D I M A C S

Center for Discrete Mathematics and Theoretical Computer Science

ANNUAL REPORT

December 1997

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

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs - Research, Bellcore, and Bell Labs

Contents

1 Overview

In this report, we summarize the accomplishments at DIMACS in the period February 1, 1997 to January 30, 1998, under the DIMACS National Science Foundation Science and Technology Center award.

Special Focus Activities

The primary scientific activity at DIMACS remains the "special focus," a focused and coordinated collection of activities that span a period of time, usually at least six months and often more than one year. The focus concentrates on an area of discrete mathematics/theoretical computer science that is ripe for rapid development, contains problems appropriate for high quality research mathematicians and computer scientists, and, in most cases, has strong potential for connections to other areas of the sciences, the mathematical sciences, and/or industry.[1] During the academic year 1996-97, we ran four special focus activities. We introduced two major new special foci, one on Networks and one on Discrete Probability, the latter joint with the Institute for Advanced Study. We also launched a modest special focus on DNA Computing, and went into the third year of the Special Year on Mathematical Support for Molecular Biology. During 1996-97, we also geared up for the 1997-98 Special Year on Massive Data Sets by running a "pre-special year" on that topic. In 1997-98, in addition to launching the Massive Data Sets special year, we are continuing the Molecular Biology, Networks, and DNA Computing special foci and, in a modest way, our special focus on Discrete Probability. We have received support from the National Security Agency to extend our Networks focus for three years. We are also planning a 1998-99 Special Year on Large Scale Discrete Optimization and are planning some 1997-98 pre-special year activities. Every special focus activity has led to dramatic scientific achievements – solutions of major open problems, the change of direction of research by a large community, the stimulus of new kinds of collaborations. The scientific achievements of our special focus activities are summarized in Section 5.

Multi-disciplinary, Multi-institutional Research Projects

In addition to special years, DIMACS has continued several major multi-disciplinary, multi-institutional research projects on "Simulations of Integrated Communications Systems" under NSF support and on "Scalable Self-Organizing Simulations" under DARPA support and laid the groundwork to continue our LINK project on software for research and education in discrete matheamtics. We have also laid the groundwork for several other major multi-institutional, multi-disciplinary projects. One is on environmental monitoring, a second on massive data sets in the biological sciences, a third involves collaborations between mathematical and biological scientists on problems of molecular biology, and a fourth is on visualization. See Section 5.8.

Continuing and Exploratory Activities

It is the plan of DIMACS to return from time to time to successful past foci. Computational Geometry is a case in point. It was the theme of our first special year. We returned to it in the summer of 1997 at our Evergreen College program (see **Undergraduate Initiatives**) and we will return to it in 1998 as part of our international partnership with the DIMATIA Center in the Czech Republic (see **Partnerships**). The 1995-96 Special Year in Logic and Algorithms is another case in point. We spent a good amount of time in 1996-97 planning to return to it with a workshop, "Logic and Algorithms a Year Later," during the summer of 1997, and are already planning to return to it as part of our DIMATIA partnership with a workshop in 1999. One way to develop topics for possible future special foci is to hold exploratory workshops. We explored one such topic

[1] We use the terms "special focus" and "special year" somewhat interchangeably. A special focus tends to be smaller in scope than a special year, but sometimes we rename something a special focus after it has been a special year and sometimes we continue a special year more than one year.

at the interface between the mathematical and the physical sciences with a workshop on disordered materials during 1996-97 and laid the groundwork for a workshop on discrete mathematical chemistry to be held in 1997-98. We also initiated some exploratory activities in the field of quantum computing by participating in the new AT&T Labs seminar on this topic and developing a summer 1997 tutorial and workshop on this topic.

Research Evaluation Report

We have embarked upon a major evaluation of activities at DIMACS. As part of this, we asked a committee consisting of Tom Leighton (MIT) and Christos Papadimitriou (UC Berkeley) to evaluate research at DIMACS. Here are a few of the highlights of their report:

- *"In our opinion, DIMACS has succeeded spectacularly in promoting high-quality research in dm/tcs, and in choosing fruitful and strategic research foci."
- *"Most celebrated breakthroughs in the past seven years in dm/tcs have had a strong DIMACS component, and were arguably catalyzed by DIMACS; the research that resulted in the PCP characterization of NP, the quantum-mechanical algorithm for factoring, and the recent polynomial-time approximation for the euclidean traveling salesman problem are three of the most salient examples."
- *"The DIMACS model of year-long focused emphasis on selected subareas often helps redirect research talent in an area of dm/tcs towards more novel and productive directions within the same area: Designers of efficient algorithms started using experimental methodology, researchers working on abstract computational models of biological systems saw the true problems facing computational biology, and researchers in logic became aware of the challenges of verification, testing, and hybrid systems."

Report on DIMACS' Special Year on Discrete and Computational Geometry

We also commissioned an evaluation of DIMACS' first special year, the 1989-90 Special Year on Discrete and Computational Geometry, as seen from the perspective of eight years. The report provides an exceptionally strong piece of evidence of the long-term impact of DIMACS' progams. It highlights the dramatic effect of DIMACS on the careers of its postdocs and students ("Talented postdocs were chosen and they profited from the environment during the special year"); the many long-term and fruitful collaborations that resulted ("the international flavor of the year stimulated cross-fertilization between communities that had not previously interacted"); the many outstanding research results that can be traced to the special year ("the impossible task of mentioning all of the accomplishments ..."); and the remarkable long-term impact on the field of research, all backed up with six pages of literature citations! The few criticisms in the report refer mostly to situations dealing with inadequate physical facilities in the first year.

Report on DIMACS' Special Year on Logic and Algorithms

A similar report on our 1995-96 Special Year on Logic and Algorithms describes the great success of the activity, with particular emphasis on the crucial role of the tutorials ("these were a boon to many researchers") and the "rich feast" of workshops; the "efficiency of the DIMACS staff"; the excellent postdocs and visitors; the many collaborations and new research results that were started ("There are many examples of new research, new and now on-going collaborations, postdocs matched with senior scientists, graduate students finding directions for their research, etc."); and the "valuable, important, and useful" interactions with industry other than our traditional participants from industrial research labs. Dealing with an activity that was more recent than our first special year, this report made more relevant comments about weaknesses and suggestions for improvement, and these have already been very useful to us in planning future special years.

Report on DIMACS' Special Year on Graph Theory and Algorithms

We also received a report on our 1991-92 Special Year on Graph Theory and Algorithms. The report speaks of the "overwhelmingly positive" response of the special year participants and "the impact the special year has had on their research and scientific and scholarly development"; the "courteous and helpful staff"; and the many scientific results that "were of special importance for the development of the field." The report concludes with a list of 70 scholarly articles that could be traced to the special year.

Regular Evaluation of our Programs

We have in place a program of regular evaluation by our workshop participants, visitors, and postdocs, run by one of our staff members. Some of the data coming out of these evaluations offers powerful evidence that we are accomplishing our purposes. For instance, a major goal of DIMACS is to facilitate collaborations. According to data obtained from our year 9 workshop participants, 93% said that they made a useful new contact at their workshop. Of those saying they had made a new contact, 73% said that they could anticipate collaborating with that person.

Undergraduate Initiatives

On the educational side, we have put a great deal of emphasis into development of ideas for new programs on undergraduate education. In the past, our major program in this direction has been our Research Experiences for Undergraduates Program, run for eight weeks each summer. This was expanded for the summer of 1997 to a program for 11 students from around the country. We launched a new module series aimed at rapid, electronic publication of modules aimed at undegraduates and teachers of undergraduates and in some cases high school teachers and students. We entered into a joint program with Evergreen State College in Washington State that introduces college teachers of Computer Science to recent research topics that are relevant to their classroom. We initiated a "Reconnect" project aimed at reconnecting faculty whose primary job is teaching undergraduates to the mathematical sciences enterprise. In this project, we will run a series of summer conferences aimed at introducing such faculty to current research topics that are relevant to the classroom and at having them develop written materials usable in the classroom; and we will run a series of two day conferences with a similar makeup aimed at two year college faculty. In the course of developing these new initiatives, we have established new partnerships with many two and four year college faculty members and their institutions and with professional organizations like the Mathematics Association of Two Year Colleges of New Jersey, which has endorsed our "reconnect" project and is helping with it.

Precollege Programs

Our highly successful precollege programs have continued, with several major new developments. One has been a concentration on ways to expand the impact of our Leadership Program beyond the DIMACS site by running off-site programs and by developing portable materials that can be used to "franchise" our programs. A second has been a new conference for high school teachers. NSF is phasing out support of Young Scholars programs. We ran the last year of our Young Scholars under NSF support and made plans for this program to continue with tuition payments by the participants and support from corporations and foundations. Meanwhile, we played a central role in the adoption of a new Curriculum Framework in Mathematics in New Jersey, cited by the governor in her 1997 "State of the State" message, and in the adoption of a discrete mathematics component by the Manhattan schools. A DIMACS volume, *Discrete Mathematics in the Schools*, was published this year, and it has been called "truly a great resource and a source of inspiration."

(Viera Proulx, Northeastern University) It was also the first DIMACS volume published in cooperation with the National Council of Teachers of Mathematics (NCTM), which co-published this with the American Mathematical Society.

Graduate Student Programs

Graduate students are and should be a central part of a science and technology center. Our $300K allocation for local graduate student support was cut from our budget starting in February, 1996, because that program did not appear to gain any leverage from DIMACS since it left choice of the participating graduate students entirely up to the individual participating departments. With the blessing of the October 1996 site visitors, we re-allocated our own funds to once again fund local graduate students, using a centralized mechanism. We have allocated $100K for this purpose for the period June 1, 1997 through August 31, 1998. Our local graduate student program consists of small DIMACS fellowships to support the research of local students, small awards for summer salary, and jobs and projects for which graduate students are paid. Initial response to this program has been highly enthusiastic. We supported seventeen students this way in the summer of 1997, and this support has already led to rather impressive scientific payback. See Section 3.2. We have also continued our programs of sponsoring DIMACS visits by non-local graduate students and continued our long-standing policy of allocating a supplement of $1,000 per workshop for support of non-local graduate students.

Postdoc Programs

One of DIMACS' goals is to involve outstanding young people in the life of the center. Our postdoc program is designed to accomplish this. We have leveraged our funds for support of postdocs by sharing postdocs with the Institute for Advanced Study and with the University of Pennsylvania (the latter a new program commencing in the Fall of 1997), obtaining grant support for postdocs, and including in our postdoc program postdocs appointed by our industrial partners. We made two significant innovations in our postdoc program during 1996-97. One is that each postdoc is assigned a faculty mentor. The second is that, because we want to train our postdocs to take an interest in educational activities and in outreach beyond the mathematical sciences, we introduced the requirement that each postdoc take part in such activities. Our postdocs meet this requirement in a variety of ways, such as helping with our K-12 programs or mentoring in our REU program. Our postdoctoral fellows from past years have beren remarkably successful in obtaining positions, both in academics and in industry. They have also rapidly risen to the top of their profession. Of the ten Sloan fellows in math and computer science announced in 1996, four were former DIMACS postdocs.

Integration of Research and Education

We have worked to integrate research and education by including educational components in our special foci, involving postdocs and graduate students in educational programs, and through our new undergraduate initiatives, the Evergreen College program, the module series, and the "Reconnect" project. We also ran the second year of our DIMACS Research and Education Institute (DREI), under a $1.9M grant from NSF. This program is aimed at integrating researchers, postdocs, students, and teachers, through a series of research workshops, a high school teacher course, and plenary activities that link these two main groups. DREI was extensively evaluated by professional evaluators Norman Webb and Kathryn Mertz, who said that "the institute met its goals and served for these separate groups in the mathematics community to engage each other and to learn from each other."

Partnerships

A key goal of DIMACS is outreach to other communities, both within and without the mathematical sciences. In the past year, we have made major progress in enhancing and expanding our many partnerships. The highly successful partnership with the Institute for Advanced Study has been expanded greatly through our joint special focus and an expanded postdoc program, and our various joint programs will be continued and expanded in the future. We inititated a joint postdoc program with the University of Pennsylvania. We initiated a partnership in computational biology with the Mt. Sinai School of Medicine, beginning with an enhancement of our REU program, and this partnership has led to a joint research proposal. We have initiated several joint activities with the Center for Molecular Biophysics and Biophysical Chemistry at Rutgers. On the international front, centers for Discrete Mathematics and Theoretical Computer Science, modeled after DIMACS, are springing up in various parts of the world. We have begun more formal contacts with these centers, in particular with the new DIMATIA Center in Prague (with NSF support). We also established a partnership with the new Erdős Institute in the Hungarian Academy of Sciences.

Minority Programs

As an outgrowth of our hosting of the second Conference of African American Researchers in the Mathematical Sciences, we saw the appearance of a DIMACS volume, *African Americans in Mathematics.*

Changes in Personnel

Stephen Mahaney, Associate Director, left DIMACS in October for a job at NSF. Julia Abrahams of ONR joins us as Associate Director in January. Robert Hochberg, one of our two education coordinators, left DIMACS in September for a position at the University of Delaware. He is continuing to work for DIMACS as a consultant, helping to organize and run our Leadership Program and our REU program and our "Reconnect" Project. Val Debellis, the other education coordinator, remains at DIMACS.

DIMACS Book Series

The DIMACS series, published with high quality hard cover volumes, reached a milestone with the publication of the 30th book in the series. There are now 39 volumes.

Overview of the Report

The rest of this report is organized as follows. In Section 2, we summarize the financial situation at DIMACS. Section 3 describes the work of the scientific personnel who were funded by DIMACS, in particular the postdoctoral and graduate student fellows. We also list all the visitors to the center in the period covered by this report. DIMACS has some 150 permanent members (see Appendix D), almost all from its five organizational members plus NEC Research, but none of them received salaries from the NSF STC award. In Section 4, we describe our scientific programs. We give information about our special focus activities – their themes, their organizers, and special year workshops. We also describe other research workshops, conferences, and scientific programs at DIMACS. In Section 5, we give highlights of scientific achievements at the center, starting with the recent special years. We also mention recent results that can be traced back to earlier special focus activities at DIMACS, and then give a selection of the exciting recent results of our postdocs, students, and visitors. We describe our multi-institutional, multi-disciplinary research projects. Finally, Section 6 describes our educational programs at the K-12 and undergraduate levels and our efforts to integrate research and education.

ABBREVIATIONS USED THROUGHOUT

ATTL = AT&T Labs - Research
BC = Bellcore
BL = Bell Labs
DM = discrete mathematics
GS = graduate student
IAS = Institute for Advanced Study
NEC = NEC Research Institute, Princeton
PD = postdoc
PM = permanent member
PU = Princeton University
RU = Rutgers University
SF = special focus
STC = science and technology center
SY = special year
TCS = theoretical computer science
V = visitor.

2 Financial Report

In this section, we summarize the center's finances, discussing its STC budget, funding from its institutional partners, the role of the New Jersey Commission on Science and Technology and the Institute for Advanced Study, and the center's external funding.

The STC Budget

The STC award from NSF for year 9 (2/1/97-1/31/98) was $1,989,483. This included the "09" base budget of $1,960,000; a $25,000 supplement for the summer 1997 REU program and $4,483 to support the DIMACS session at the AMS meeting in San Diego. In addition to the year 9 award we also had available an allowed carryover of $267,500.

Institutional Support from Partner Organizations

DIMACS receives support from its five institutional partners. Rutgers provides the director's salary, salary for part of a secretary, space, and operating and maintenance expenses with an estimated value of $333,028. Rutgers also returns to DIMACS a significant portion of the indirect costs on the main STC award and other grants, amounts between 40% and 50% of the overhead; for the one year period beginning July 1, 1996, this return totaled $334,216. Princeton provides half of a postdoc, graduate student tuitions in the years when we support students, secretarial support, and space. For the one year period beginning July 1, 1996, the amount of support provided by Princeton is estimated to be $172,540. Our industrial partners provided substantial support by appointing DIMACS postdocs, sponsoring workshops, and allowing release time for their researchers to actively participate in DIMACS administrative committees and to organize DIMACS workshops and Special Years. For the one year period beginning 7/1/96, the estimated value of their contributions are: Bell Labs $490,000; AT&T Labs - Research $765,000 and Bellcore $565,000.

All five institutions strongly encourage volunteer work by their scientists, and the number of hours spent on DIMACS committee work and organizational activities by their scientists is very large. For example, over 40 permanent members of DIMACS, representing all of the five member organizations, are actively involved in the committees that run the center and its programs, the Executive Committee, Council, Strategic Planning Committee, Budget Committee, Projects Committee, Education Committee, Postdoc Committee, Equipment Committee, Graduate Student Selection Committee, REU Committee. More are involved in an organizational role in our many special focus programs, workshops, and educational projects. Our industrial partners have estimated the value of their in-kind contributions through the participation of their employees in DIMACS organizational activities as follows: Bell Labs $200,000; AT&T Labs $250,000; Bellcore $200,000. These amounts are included in the amounts in the previous paragraph. We have not attempted to put a value on the in-kind support provided by the academic partners.

New Jersey Commission on Science and Technology

Since the start of DIMACS, with the exception of one year, the center has been receiving an annual award of $100,000 from the New Jersey Commission on Science and Technology. Recently, the Commission made the decision to phase out its support of centers over a three year period, replacing the centers program by a program of competitive awards for research projects. Most of the centers supported by the state received major cuts for the fiscal year beginning July 1, 1996 (FY97), and some were cut off entirely. However, recognizing that the State had some obligation to continue to support DIMACS, the Commission did not cut our center and we once again received $100,000 for the fiscal year. For FY98, the Commission's Budget Committee recommended that the Commission's support of centers be cut to 50% of its FY97 level. However, it was recommended that the support for DIMACS be continued in full and we did, indeed, receive the full $100,000 for the current fiscal year. While the amount of Commission support is not large, we consider it significant given the current situation.

Institute for Advanced Study Program

The Institute for Advanced Study is sharing in funding several joint projects, including our joint postdoc program and our joint special focus in discrete probability. The cost of the IAS computer science and combinatorics program, which is "created to be complementary to and supportive of the work of DIMACS," is approximately $270K per year, not including administration, secretarial support, Institute overhead, or Institute faculty salaries. IAS is also a partner with DIMACS in the DIMACS Institute for Research and Education (DREI), an externally-funded program that integrates research and education and is described in detail in Section 6.5.

External Funding

DIMACS has external funding for a variety of its special focus activities, group research activities, and educational activities. The largest sources of such funds in currently active accounts (with current and remaining years awards) are as follows:

- (1). A grant of $332,155 over three years from the National Security Agency for support of the Special Year in Network Security and Network Applications.

September 1999 to August 2000	$129,417
September 1998 to August 1999	$108,202
September 1997 to August 1998	$ 94,536

- (2). A grant of $70K over 3 years from NSF to support joint workshops and visitor exchanges between DIMACS and DIMATIA in the Czech Republic.

May 1, 1999 to June 30, 2000	$23,137
May 1, 1998 to June 30, 1999	$15,965
May 1, 1997 to June 30, 1998	$30,908

- (3). A grant of approximately $1.9M over four years from NSF for our DIMACS Research and Education Institute (DREI).

July 1998 to June 1999	$467,816
July 1997 to June 1998	$466,402
July 1996 to June 1997	$475,874

- (4). A grant of approximately $1.5M over three years from NSF for a multi-disciplinary, multi-institutional project on Simulations of Integrated Communications Systems.

September 1997 to August 1998	$466,912
September 1996 to August 1997	$460,244

  An REU supplement was received for this grant for the summer of 1997 to support one student.

- (5). A grant of $3.1M over three years from DARPA for our project on Scalable Self-Organizing Simulations.

August 1998 to July 1999	$1,163,952
August 1997 to July 1998	$1,109,721
August 1996 to July 1997	$ 857,258

- (6). A grant of $3.1M over five years from the Education and Human Resources Directorate of NSF for support of the Leadership Program in Discrete Mathematics.

April 1, 1999 to March 31, 2000	$314,056
April 1, 1998 to March 31, 1999	$803,860
April 1, 1997 to March 31, 1998	$782,009
April 1, 1996 to March 31, 1997	$711,418

- (7). A grant of $194K over two years from the Education and Human Resources Directorate of NSF for support of the Young Scholars Program in Discrete Mathematics.

March 1, 1997 to February 28, 1998	$ 86,400
March 1, 1996 to February 28, 1997	$108,000

- (8). Grants from NSF, NSA, CDC, National Institute for Allergy and Infectious Diseases (NIAID), the National Center for Human Genome Research (NCHGR), and support from Roche Molecular Systems, SmithKline Beecham, BIOSYM Tech, Merck Research, Ciba-Geigy, and Sandia, totaling $625K over three years, for support of our molecular biology activities.

- (9). Funds from ONR for approximately $122K over three years to support a senior visitor in a project on binary hierarchy (clustering).

December 1, 1997 to November 30, 1998	$30,000
December 1, 1996 to November 30, 1997	$40,000

3 Postdoctoral Fellows, Graduate Student Fellows, and Visitors

In this section, we describe the scientific work of the postdoctoral fellows and graduate student fellows who were supported by DIMACS, and we list the visitors to the Center. We include postdocs in both academic years covered by this report, 1996-97 and 1997-98, and the graduate students in the period February 1, 1997 to January 31, 1998. DIMACS has some 150 permanent members (see Appendix D), almost all from its five organizations, plus NEC Research. None of the permanent members were supported on the STC grant during the period covered by this report. However, many of them were actively involved in DIMACS as organizers of programs, serving on committees, and in other jobs. Moreover, many of them collaborated with our postdocs, students, and visitors. Some of the scientific results coming out of these collaborations are reported on in this section and others in Section 5, when we highlight the recent scientific accomplishments at the center.

3.1 Postdoctoral Fellows: September 1996 - August 1998

Amos Beimel (1996-97) (SY in Networks)
Currently at Harvard University, Division of Applied Sciences

Beimel worked on reliable communication, learning, and arithmetic branching programs.

Together with Matt Franklin (PM, ATTL), Beimel worked on reliable communication over partially authenticated networks. Reliable communication between parties in a network is a basic requirement for executing any protocol. In this work, Beimel and Franklin considered the effect on reliable communication when some pairs of parties have common authentication keys. The pairs sharing keys define a natural "authentication graph", which may be quite different from the "communication graph" of the network. They characterized when reliable communication is possible in terms of these two graphs, focusing on the very strong setting of a Byzantine adversary with unlimited computational resources. More specifically, reliable communication involves a transmitter which sends a message to a receiver. For single-pair reliable communication, we specify the transmitter and receiver, and want to succeed against any coalition of at most t faulty processors. For all-pairs reliable communication, we want to succeed for any transmitter, any receiver, and any coalition of at most t faulty processors. The network of channels defines a natural "communication graph", with an edge between two vertices for every channel between two processors. The pairs of parties sharing authentication keys define a natural "authentication graph", with an edge between two vertices for every shared key. Beimel and Franklin showed that all-pairs reliable communication is possible if and only if the communication graph is $(t+1)$-connected and the union of the two graphs is $(2t+1)$-connected.

The learnability (under various learning models) of geometric concept classes has been widely studied. Particular attention has been given to the case of discrete domains of points and concept classes which are defined as unions of boxes in this domain. One of the reasons that unions of boxes seem to be interesting concepts is that they naturally extend DNF formulae. Beimel and Eyal Kushilevitz (V) presented exact learning algorithms that learn several classes of (discrete) boxes in $\{0,\ldots,l\}^n$. In particular they learn the class of unions of $O(\log n)$ boxes in time $\mathrm{poly}(n, \log l)$, solving an open problem. And they learn the class of unions of disjoint boxes in time $\mathrm{poly}(n, t, \log l)$, where t is the number of boxes. Previously this was known only in the case where all boxes are disjoint in one of the dimensions. In particular their algorithm learns the class of decision trees over n variables that take values in $\{0,\ldots,l-1\}$, with comparison nodes in time $\mathrm{poly}(n,t,\log l)$, where t is the number of leaves, another open problem. They also learn the class of unions of $O(1)$-degenerate boxes (that is, boxes that depend only on $O(1)$ variables) in time $\mathrm{poly}(n,t,\log l)$ (generalizing the learnability of $O(1)$-DNF and of boxes in $O(1)$ dimensions).

Beimel and Ana Gal (PD) introduced the model of arithmetic branching programs, which is a generalization of modular branching programs. This model helps to understand some algebraic complexity classes and the connections between them and basic algebraic problems such as computing the rank of a matrix and checking if a matrix is singular. It is simple to show that arithmetic programs are closed under conjunction. Beimel and Gal showed that arithmetic programs are also closed under disjunction. Using these closure properties, they showed that, up to a polynomial factor in size, arithmetic programs are equivalent to dependency programs, a model introduced by Pudlak and Sgall. A consequence of this equivalence is that dependency programs are closed under conjunction, answering an open problem of Pudlak and Sgall.

Papers

A. Beimel, and E. Kushilevitz. "Learning boxes in high dimension," *Algorithmica*, 1997. Accepted for publication.

A. Beimel, and E. Kushilevitz. "Learning boxes in high dimension," In *3rd EuroCOLT*, volume 1208 of LNAI, pages 3-15, 1997. (This is an earlier version of the previous paper.)

A. Beimel, and M. Franklin. "Reliable communication over partially authenticated networks," *WDAG'97*, 1997. To appear.

A. Beimel, and A Gal. "On arithmetic branching programs," 1997. In preparation.

Talks

FOCS '96 Conference, Burlington, Vermont, October 1996. "On the applications of multiplicity automata in learning".

DIMACS Discrete Math/Theory of Computing Seminar, Rutgers University, New Jersey, November 1996. "On applications of multiplicity automata in learning."

DREI workshop, Rutgers University, New-Jersey, August 1997. "Reliable communication over partially authenticated networks".

Hui Chen (1996-97) (SY in Discrete Probability)
Currently Strategist at CS First Boston

Chen worked on broadcasting in random graphs and thresholds for ordering libraries of random clones.

A *broadcasting process* involves sending a piece of information which resides at one vertex in a graph to all the remaining vertices. At each time step, a vertex knowing the information can send it to one of its neighbors. The broadcast problem is to find the minimum number of time steps needed. The problem is NP hard in general. It is not hard to show that the edges used by the optimal broadcast algorithm form a tree. The broadcast problem is to find the tree and the minimum broadcast time. Chen considered the problem, for a random graph $G_{n,p}$, of finding the threshold for p such that with high probability (whp), a piece of information can be broadcasted in the graph in exactly $\lceil \log_2 n \rceil$ steps. (An event E_n is said to occur whp if $\mathbf{Pr}[E_n] = 1 - o(1)$ as $n \to \infty$.) The best upperbound to date is due to Frieze and Molloy who showed that if $p \geq 18 \ln n/n$ then whp there exists a rooted broadcasting tree on every vertex. Chen showed that a trivial lower bound is also sufficient for the existence of a rooted optimal broadcast tree on each vertex of a random graph $G_{n,p}$.

One of the many biological problems to which mathematicians have given attention in recent years is the Physical Mapping Problem from genetics. To attain a physical map, many copies of a genome are digested by various enzymes, creating a library of overlapping clones. One technique involves further digestion of the clones to get sets of digestion lengths. Chen has considered the technique of testing the clones for the existence of various DNA subsequences called probes. If the probes are (relatively) long, they will appear at most once in the genome, and they are called unique probes. If the probes are (relatively) short, their appearances throughout the genome will be according to a Poisson process (assuming randomness of the base pairs in the genome), and these are called Poisson probes. In either case, clone overlaps can be detected by commonality of probes, and this information will be used to reconstruct the actual overlap structure of the clone library, and the ordering of the probes on the genome. This information constitutes the physical map, which can be used by biologists for further sequencing, or as a guide to other genetic research. Chen showed that a gapless distribution of clones with distinct endpoints has a unique ordering of the atomic intervals (up to complete reversal) satisfying the so-called consecutive 1's property if and only if there are no two atomic intervals the same. The consecutive 1's property provides a simple test of whether or not a graph is an interval graph, a central concept in the history of molecular genetics and in particular in physical mapping.

Papers

Hui Chen. "Threshold of broadcast in random graphs." DIMACS Tech. Report, 1997.

Hui Chen. "Thresholds for ordering libraries of variable length clones." RUTCOR Research Report, 1997.

Scott Decatur (1996-97) (SY in Molecular Biology)

Currently at Grantham, Mayo, Van Otterloo & Co.

Decatur worked in the areas of machine learning, computational biology and information retrieval.

In machine learning, with Oded Goldreich and Dana Ron, Decatur examined the effect of computation time on the amount of training data required and the effect of various types of training data corruptions on the learning process. In a variety of PAC learning models, a tradeoff between time and information seems to exist – with unlimited time, a small amount of information suffices, but with time restrictions, more information sometimes seems to be required. In addition, it has long been known that there are concept classes that can be learned in the absence of computational restrictions, but (under standard cryptographic assumptions) cannot be learned in polynomial time *regardless of sample size*. Yet, these results do not answer the question of whether there are classes for which learning from a small set of examples is infeasible, but becomes feasible when the learner has access to (polynomially) more examples. To address this question, Decatur and his collaborators introduced a new measure of learning complexity called *computational sample complexity* which represents the number of examples sufficient for *polynomial time* learning with respect to a fixed distribution. They showed concept classes that (under similar cryptographic assumptions) possess arbitrary sized gaps between their standard (information-theoretic) sample complexity and their computational sample complexity, and demonstrated such gaps for learning from membership queries and learning from noisy examples. Decatur also considered the problem of concept learning in Valiant's PAC learning model in which the data used for learning is noisy. Specifically, he introduced a new model of noise called *constant-partition classification noise* (CPCN) which generalizes the standard model of classification noise to allow different examples to have different rates of random misclassification, e.g., the rate of false positives different from the rate of false negatives. He then showed how to learn in the presence of CPCN for any concept class learnable by statistical queries. This set of classes includes *every* concept class known to be learnable in the presence of standard classification noise. His model is the first such non-uniform generalization of the standard classification noise model that allows efficient learning of this wide range of concept classes.

In the area of computational biology, Decatur, with Richa Agarwala (PD), Serafim Batzoglou, Vlado Dancik, Martin Farach (PM, RU), Sridhar Hannenhalli (V), and Steven Skiena (V), considered the problem of determining the three-dimensional folding of a protein given its one-dimensional amino acid sequence. They used the HP model for protein folding, proposed by Dill, which models each protein as a chain of amino acid residues that are either hydrophobic or polar, and makes the assumption that hydrophobic interactions are the dominant initial driving force for protein folding. Although it is not known how to find the optimal fold in polynomial time, Hart and Istrail gave polynomial time approximation algorithms for folding proteins on the cubic lattice under the HP model. Decatur and his collaborators examined the choice of lattice by considering its algorithmic and geometric implications and argued that a triangular lattice is a more reasonable choice. They presented a set of folding rules for a triangular lattice and analyzed the approximation ratio which they achieved. In addition, they introduced a generalization of the HP model to account for residues having different levels of hydrophobicity. After developing the biological foundation for this generalization, they showed that in the new model they were able to achieve similar constant factor approximation guarantees on the triangular lattice as were achieved in the standard HP model. While the structures derived from their folding rules are probably still far from biological reality, they hope that having a set of folding rules with different properties will yield more interesting folds when combined.

In other work, it should be noted that previous algorithms for protein folding yielded results that were proven to be anywhere from 50-60% of optimal. An algorithm due to Decatur and Batzoglou (1996) has been quoted as yielding 60%. The Batzoglou-Decatur algorithm involves wrapping the residues in a tube like structure that has a solid hydrophobic core. The algorithm considers windows of 7 H residues, and determines the shape of each layer based on the number of intervening P groups. In work with REU student Peter Coles, Decatur designed an algorithm, based on the Batzoglou-Decatur algorithm, that involves several important modifications such as taking into account the location of the intervening P groups, not just the number of P-groups. Preliminary results show that the algorithm has a worst case yield of just under 61%. This is not significantly higher than 60%, but it does show that 60% is not a fundamental limiting ratio for these types of layering algorithms, a result that was previously unknown.

In the area of information retrieval, Decatur proposed a method for the incorporation of user feedback to improve the utility of web search engines such as AltaVista, Lycos, etc., in which the user may denote any link returned by the search engine as GOOD or BAD, corresponding to the user's assessment of the link's relevance to the search being conducted. This user feedback is then used to reorder all returned links from most to least relevant according to their similarity with user-labelled links. The ability to label and order by similarity often provides the user with greater power to locate desired documents than simply resubmitting a refined keyword search. Decatur constructed a prototype of such a system, called VistaLearn, which allows user feedback to order links returned by Digital's AltaVista search engine. He also described a variety of proposed functional extensions and implementation improvements to the system.

Papers

Scott Decatur, Oded Goldreich, and Dana Ron. "Computational sample complexity." In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, July 1997.

Scott Decatur. "PAC learning with constant-partition classification noise and applications to decision tree induction." In *Proceedings of the Fourteenth International Conference on Machine*

Learning, July 1997. Preliminary version appeared in *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, January 1997.

Richa Agarwala, Serafim Batzoglou, Vlado Dancik, Scott Decatur, Martin Farach, Sridhar Hannenhalli and Steven Skiena. "Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model." To appear in *Journal of Computational Biology*. Extended abstracts appeared in *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, January 1997, and *Proceedings of the First International Conference on Computational Molecular Biology*, January 1997.

Talks

6th International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL, Jan. 1997. "PAC learning with constant-partition classification noise and applications to decision tree induction."

MIT Laboratory for Computer Science Theory of Computation Seminar, April 1997. "Learning with nonuniform noise and noisy decision tree induction."

MIT class on Computational Biology, April 1997. "Protein folding in the HP model."

10th Conference on Computational Learning Theory, Nashville, TN, July 1997. "Computational sample complexity."

14th International Conference on Machine Learning, Nashville, TN, July 1997. "PAC learning with constant-partition classification noise and applications to decision tree induction."

Anna Gál (1995-97) (DIMACS - IAS Postdoc)
Currently at University of Texas at Austin

Gál worked on span programs, computing from partial solutions, and a model of arithmetic branching programs.

A large part of Gál's recent work is related to the study of span programs. The model of span programs was introduced by Karchmer and Wigderson in 1993. It is a computational model based on linear algebra with applications to lower bounds in other models, including Boolean formulae. The monotone version of this model is strongly related to the cryptographic problem of secret sharing. Proving lower bounds for monotone span programs has been a challenging open problem. Together with Amos Beimel (PD) and Mike Paterson, Gál developed a new lower bound technique that allows one to prove lower bounds for monotone span programs by considering a problem in extremal set theory. As a first application of the technique they proved a lower bound of $\Omega(n^{2.5})$ for the 6-clique function. In joint work with László Babai and Avi Wigderson (V), Gál showed that this technique can yield superpolynomial lower bounds for explicit functions. Their construction of an explicit function family that requires superpolynomial size monotone span programs is based on set systems obtained from Paley-type bipartite graphs. They proved an $n^{\Omega(\log n/\log\log n)}$ lower bound for an explicit family of monotone Boolean functions in n variables, which implies the same lower bound for the size of monotone span programs for the clique problem. Their results give the first superpolynomial lower bounds for linear secret sharing schemes. To demonstrate the power of monotone span programs they showed that this model can be much more powerful than monotone circuits. They exhibited explicit functions that have linear size monotone span programs, but require superpolynomial size monotone circuits and exponential size monotone formulae. This in particular implies that lower bound methods for the circuit model cannot be directly applied to span programs. They also showed that the perfect matching function can be computed by polynomial size (non-monotone) span programs over arbitrary fields. During the second year of her stay at DIMACS, Gál improved the above lower bounds for monotone span programs. She also gave a new characterization of span program size in terms of a combinatorial-algebraic measure which yields a new technique for proving lower bounds for monotone span programs.

Gál worked with Richard Lipton (PM, PU) and Erez Petrank (PD) on the problem of computing from partial solutions. They considered the question: Is it easier to find part of a solution than the full solution? For example, suppose we have to find only a small constant fraction of the bits in an assignment to a 3-CNF formula. Is it easier than computing the full assignment? For the three problems: 3-SAT, Graph Isomorphism, and a Shortest Lattice Vector problem, Gál, Lipton, and Petrank showed that finding only a small fraction of the solution is actually equivalent to finding the full solution.

Gál and Amos Beimel (PD) considered the model of arithmetic branching programs, which is a generalization of modular branching programs. They showed that, up to a polynomial factor in size, arithmetic programs are equivalent to dependency programs, a model introduced by Pudlák and Sgall. Using this equivalence they proved that dependency programs are closed under conjunction, answering an open problem of Pudlák and Sgall. Furthermore, they showed that span programs are at least as strong as arithmetic programs; every arithmetic program can be simulated by a span program whose size is only twice the size of the arithmetic program.

Papers

L. Babai, A. Gál, J. Kollár, L. Rónyai, T. Szabó, A. Wigderson. "Extremal bipartite graphs and superpolynomial lower bounds for monotone span programs." In *Proc. 28th ACM STOC*, 1996, pp. 603–611.

L. Babai, A. Gál, A. Wigderson. "Superpolynomial lower bounds for monotone span programs." Submitted to *Combinatorica*.

A. Gál. "A characterization of span program size and improved lower bounds for monotone span programs." In preparation.

A. Beimel and A. Gál. "On arithmetic branching programs." In preparation.

A. Gál, R. Lipton, E. Petrank. "Computing from partial solutions." In preparation.

Talks

ACM Symposium on the Theory of Computing, Philadelphia. May 1996. "Extremal bipartite graphs and superpolynomial lower bounds for monotone span programs."

Workshop on Graph Theory and Combinatorics, Balatonlelle, Hungary. July 1996. "Extremal bipartite graphs and superpolynomial lower bounds for monotone span programs."

Institute for Advanced Study, Princeton, Combinatorics and Complexity Seminar. March 31, 1997. "A characterization of span program size."

The University of Texas at Austin. April 2, 1997. "Lower bounds for monotone span programs."

Bellcore. April 29, 1997. "Fault tolerance of Boolean circuits."

The University of Toronto. May 1, 1997. "Lower bounds for monotone span programs."

Michael Goldwasser (1997-98) (SY in Massive Data Sets)

Goldwasser's research interests lie in the design and analysis of algorithms, with particular areas of interest including computational geometry, approximation algorithms, and scheduling. So far at DIMACS, he has been working on three problems. The first has been a problem from the domain of online scheduling for admission control. The second involves some issues from persistent data strucutures. The third has to do with caching with multi-size pages. The scheduling problem Goldwasser has considered is scheduling the maximum number of jobs on a single processor, without preemption, when jobs are unit-length, and have arbitrary arrival and deadline times. There is a natural greedy algorithm which was previously shown to be strongly-competitive. He has extended these results to show that the algorithm is also strongly competitive, if jobs are required to have a minimum delay of W between their arrival and deadline. The results extend the proofs to hold for all values of W.

Anders Johansson (1995-97) (DIMACS/IAS postdoc)

Johansson has worked on chromatic numbers and list chromatic numbers of graphs. A *list coloring* of a graph arises when we are given a list of allowable colors at each vertex, and we seek an ordinary graph coloring in which each vertex receives a color chosen from its list. List colorings have a variety of practical applications in channel assignment problems, traffic phasing problems, and elsewhere. The *list chromatic number* of a graph is the smallest k so that for every assignment of lists of size k to each vertex, there is always a list coloring. Johansson has obtained asymptotic upper bounds on the chromatic number or list chromatic number of a graph for several classes of graphs with no r -clique. The bound is in terms of the maximum vertex degree – a Brooks'-type bound. The methods for obtaining these results use randomized arguments that should help shed light on the behavior of "simulated annealing," a much-used heuristic for practical applications. Johansson has also worked on applying the same methods to other types of coloring and partition problems. In other work, Johansson has studied *equitable edge-colorings* of eulerian graphs, edge-colorings where the degree of each chromatic factor differs by at most one at all vertices in the graph. He has investigated the relation of this notion to a result of Lovasz and Cherkasskij concerning the possibility of linking the vertices of a set T by edge-disjoint paths, given that the vertices outside T have even degree.

Michael Krivelevich (1997-99) (Joint DIMACS/IAS Postdoc)

Krivelevich's field of interest is combinatorics and its applications to theoretical computer science. He has been working on various problems in probabilistic combinatorics, including problems on random graphs and hypergraphs (coloring, matching problems in random hypergraphs, factors in random graphs) and applications of the probabilistic method to other branches of combinatorics such as Ramsey Theory. Another direction of his research is extremal graph theory, where he has some results on coloring problems and covering problems in uniform hypergraphs. He is also interested in algorithmic problems such as approximation algorithms and algorithmic aspects of random graphs and hypergraphs.

Working with Noga Alon (V) and Benny Sudakov (V), Krivelevich has studied the *choice* (or the *list coloring*) *number* of a graph. This is the smallest k so that whenever there are sets of size k assigned to each vertex, there is always a graph coloring where the color assigned to a vertex belongs to the set assigned to the vertex. Colorings with such lists arise in a wide variety of practical problems such as channel assignments in communications. Being a natural generalization of the chromatic number, the choice number of a graph seems to be a much more complicated quantity and much less is known about it. An interesting particular question is: what is the asymptotic value of the choice number of a random graph $G(n, p)$? Alon, Krivelevich, and Sudakov have obtained some results on this problem, showing in particular that almost surely the choice number and the chromatic number of $G(n, p)$ have the same order of magnitude, namely, $np/\log(np)$.

Another interesting meta-problem is the following. It is well known that the chromatic number of a graph G with maximal degree d does not exceed $d + 1$, and this is optimal as shown by the example of a complete graph on $d + 1$ vertices. However, putting some restrictions on the graph, one may hope for a better bound. An illustrative example is the following result, obtained by Krivelevich together with Alon and Sudakov: if for a fixed $\epsilon > 0$, a graph G with maximal degree d has at most $d^{2-\epsilon}$ edges in every vertex neighborhood, then the chromatic number of G is of order at most $d/\log d$.

A fascinating conjecture is the following: Let H be a fixed graph; if G is an H-free graph of maximal degree d, then the chromatic number of G has order at most $d/\log d$. This is true for H being a cycle of any fixed length, but is open for $H = K_4$. Together with Sudakov, Krivelevich has been working on the problem of approximating the chromatic number of r-uniform hypergraphs. While the graph case (r=2) has received a considerable amount of attention and many results are known in positive (approximation algorithms with guaranteed approximation ratio) and negative (hardness of approximation) directions, the general case remains almost completely open.

Talks

Combinatorics, Complexity and Discrete Probability Seminar, Institute for Advanced Study, Princeton, NJ, September 1997. "Finding a large hidden clique in a random graph."

DIMACS Mixer at AT&T Labs, Florham Park, NJ, October 1997. "Finding a large hidden clique in a random graph."

Combinatorics Seminar, University of Pennsylvania, PA, November 1997. "Coloring random and pseudo-random graphs."

Dominic Mayers (1996-98)

Mayers works at the interface between quantum physics and information science, in particular in quantum cryptography. There are two main applications in quantum cryptography: quantum bit commitment and quantum key distribution. Quantum bit commitment is important because it was shown by Andrew Yao (PM, PU) that it is sufficiently powerful to construct most post-cold war cryptographic applications. The first quantum bit commitment protocol was proposed by Bennett and Brassard in 1984 together with a way to break it. Nevertheless, the hope remained for many years that we might be able to build an unconditionally secure quantum bit commitment protocol. In fact, in 1993 a quantum bit commitment protocol was proposed together with a security proof. This proof was generally accepted until Mayers discovered in 1995 a flaw in the protocol and in the proof. Not only did Mayers prove that this particular protocol is not unconditionally secure, but later, in part while at DIMACS, he also proved that unconditionally secure bit commitment is impossible in general. With this result, he put an end to a hope that had existed for more than 13 years. Part of this work was done while he was a postdoc at DIMACS. Mayers also obtained the first proof that quantum key distribution is unconditionally secure.

Mayers has also obtained a result in non quantum cryptography. This result is about a basic primitive called oblivious transfer (OT). In OT, one participant Alice sends a bit of her choice to another participant Bob. The bit must be received by Bob with probability 1/2. Bob must know when he receives the bit, but Alice should learn nothing about whether or not Bob has received this bit. It is well known that OT is powerful enough to construct many cryptographic applications. Very often, when OT is executed between Alice and Bob in a protocol with other participants (Charlie, Dan, etc.), it is explicitly required that OT be executed on a private channel between Alice and Bob. In other words, the other participants should learn nothing about the bit chosen by Alice and whether or not Bob has received this bit. Perhaps Mayers' most important contribution on this subject was to ask for the first time whether or not the explicit requirement for a private channel was really necessary. Such a private channel would not be necessary if every OT protocol (independently of what channel is used in the protocol) was necessarily private. Mayers proved OT is more powerful than one might first think in the sense that we don't have to worry about privacy when we use it.

Papers

D. Mayers. "Unconditionally secure quantum bit commitment is impossible." *Physical Review Letters*, Vol. 78, No. 17, April 1997, pp. 3414-3417.

Dominic Mayers and Andrew Yao. "Quantum key distribution is unconditionally secure." In preparation.

Gilles Brassard, Claude Crépeau, Dominic Mayers and Louis Salvail. "Quantum attacks against classical bit commitment schemes." In preparation.

Dominic Mayers. "Entanglement in quantum cryptography." In preparation.

Talks

Fourth workshop on Physics and Computation (also called PhysComp96), Boston University, 22-24 November 1996. In: Proceeding of Physcomp96, New England Complex Systems Institute (1996). "Unconditionally secure bit commitment is impossible."

Tutorial talk at DIMACS Workshop on Quantum Computation held at Princeton University, 1997. "Quantum cryptography."

Erez Petrank (1996-97) (SY in Networks)

Currently at IBM - Haifa Research Lab

Petrank worked on linear hashing, probabilistic checkable proofs, identity escrow, computing from partial solutions, and message authentication.

In joint work with N. Alon, M. Dietzfelbinger, P. B. Miltersen, and G. Tardos (V), Petrank studied the quality of the family of linear transformations serving as a family of hash functions. Consider the set $\mathcal{H}$ of all linear (or affine) transformations between two vector spaces over a finite field F. They studied how good $\mathcal{H}$ is as a class of hash functions; namely they considered hashing a set S of size n into a range having the same cardinality n by a randomly chosen function from $\mathcal{H}$ and looked at the expected size of the largest hash bucket. $\mathcal{H}$ is a universal class of hash functions for any finite field, but with respect to our measure different fields behave differently. If the finite field F has n elements then there is a bad set $S \subset F^2$ of size n with expected maximal bucket size $\Omega(n^{1/3})$. If n is a perfect square then there is even a bad set with largest bucket size *always* at least $\sqrt{n}$. If, however, we consider the field of two elements then Petrank and his collaborators got much better bounds. The best previously known upper bound on the expected size of the largest bucket for this class was $O(2^{\sqrt{\log n}})$. Petrank, et al. reduced this upper bound to $O(\log n \log\log n)$. Note that this is not far from the guarantee for a random function. There, the average largest bucket would be $\Theta(\log n / \log\log n)$.

Together with Joe Kilian (PM, NEC) and Gabor Tardos (V), Petrank studied how to get probabilistic checkable proofs with a zero knowledge property and constructed PCPs with strong zero-knowledge properties. First, they constructed polynomially bounded (in size) PCP's for NP which can be checked using poly-logarithmic queries, with polynomially low error, yet are statistical zero-knowledge against an adversary that makes U arbitrary queries, where U can be set to any polynomial. Second, they constructed PCPs for $\mathsf{NEXPTIME}$ that can be checked using polynomially many queries, yet are statistically zero-knowledge against any polynomially bounded adversary. These PCPs are exponential in size and have exponentially low error. Previously, it was only known how to construct zero-knowledge PCPs with a constant error probability.

In work with Joe Kilian (PM, NEC), Petrank worked on secure authorized access. The issue that arises is the tension between the need of a person to access services without revealing his identity, and the need to reveal identities of users in extreme cases, for example, when a crime is committed. They used Key escrow techniques to meet both requirements. Key escrow has proven an active and contentious field of research and discussion. Essentially all of the attention in this area has been restricted to the simple case of communication: party A sends an encrypted message $E_K(M)$ to party B; some centralized authority is given the capability to recover either K or the specific message M. Kilian and Petrank considered a new domain for the application of key escrow ideas: Escrow in identity schemes. A rapidly increasing number of interactions take place on the purely information level, and are likely to become ubiquitous. As just one example, on some toll bridges one can now electronically identify oneself to a toll booth instead of waiting in line to pay a toll. Although this application is cost effective (or will be) and more convenient, it raises critical issues of privacy; a more pervasive toll collection system might as a side effect allow the tracing of people's movements to an unprecedented degree. This work addresses the tension between party A's desire for privacy and anonymity and B's (and society's) desire to know who they are interacting with. This balance point of this tension can shift under certain often rare circumstances. For example, consider an automated parking garage. What is paramount to the garage is that the person entering it is authorized to do so; the person's precise identity is normally not a valid concern. However, suppose that on some night a person was murdered in the garage. At this point, the garage owner and society at large may have a legitimate interest in knowing who was there on that night. Or, one might wish to have a computer "chat room" in which one has conditional anonymity messages: As long as one follows the rules, one's identity is secure from even the system administrator. But if one flagrantly breaks the rules (such as arranging drug deals in the "Lion King" kiddie chat room), suitable law enforcement agencies can be appealed to in order to determine one's identity. Traditionally, access control has been all or nothing. One obtains all the information about the other person up front, with no recourse to learn more if circumstances warrant. This rigidity generally leads one to allow less privacy, since one is likely to want as much information as one can get just in case a "bad case" arises. Kilian and Petrank have given a more flexible alternative. They considered a two-tier approach to identification. On the lower tier, a person gives only as much information about themselves as is strictly necessary for ordinary circumstances. In the garage example above, this means proving authorization to use the garage. On the second tier, a person gives a more precise statement of their identity that may be needed in extraordinary circumstances. This second tier is only accessible with the help of a third party, i.e., the escrow agency, which is preferably separate from and not under the control of the party managing access (such as the garage owners). An escrowed identity system consists of an identifier, an issuer, a verifier, and an escrow agent. These parties execute various protocols, in particular issuing a certificate, checking the weak identity, and recovering the complete identity. A secure identity escrow system should work, make it hard to fake a legitimate identity, keep the escrowed identity hidden, make it hard to duck the escrow, and enhance security. Kilian and Petrank introduced the notion of an escrowed identity system and gave two concrete implementations. One is based on the RSA encryption scheme and the other on the El-Gamal signature and encryption schemes. An international patent is pending on this work (by Rutgers University and NEC).

In joint work with Anna Gál (PD) and Richard Lipton (PM, PU), Petrank studied the question: Is it easier to find part of a solution than the full solution? For example, suppose we have to find only an ϵ fraction of the bits in an assignment to a 3-CNF formula. Is it easier than computing the full assignment? For the three problems: 3-SAT, Graph Isomorphism, and a Shortest Lattice

Vector problem, they showed that finding only a small fraction of the solution is actually equivalent to finding the full solution. This can be interpreted in two ways: On the positive side, it is enough to look for an efficient algorithm that only recovers a small part of the solution, in order to completely solve any of these problems. On the negative side, any partial solution to these problems may be hard to find. For example, unless $BPP = NP$ there is no efficient algorithm which always recovers $\sqrt{n}$ bits of a satisfying assignment of a 3-CNF formula with n variables.

Message authentication is one of the most used cryptographic primitives. In this setting, two parties which share a secret key would like to make sure that the messages they exchange are not changed on the communication channel before delivery. The Cipher Block Channing (CBC) Message Authentication Code (MAC) is an authentication method which is widely used in practice. Nevertheless, until recently no solid theoretical approach was suggested for proving security of such systems. The first paper to provide rigorous proof of security for CBC MAC (conditioned on the security of the underlying cryptographic block cipher) was by Bellare, Kilian and Rogaway in 1994. They suggested a framework for proving security of CBC MAC, and also used this framework to prove that the security of CBC MAC follows from the security of the underlying block cipher when the messages have fixed length. They also suggested variants of CBC MAC that handle variable length messages, but where the length of the message has to be known in advance. In joint work with Charles Rackoff, Petrank followed the work of Bellare, et al. by studying the case in which the length of the message is not known until the message ends. This seems an important issue in particular because the transmission of fax is inherently like that: When the transmission starts, the machine does not know how many pages are going to be input. Petrank and Rackoff first presented a variant of CBC MAC, called *double MAC* (DMAC), which handles variable unknown lengths. They proved that its security is implied by the security of the underlying block cipher. Next, they argued that the basic CBC MAC is secure when applied to prefix free message space. A message space can be made prefix free by authenticating also the (usually hidden) last character which marks the end of the message. This character never appears inside the message.

Papers

N. Alon, M. Dietzfelbinger, P. B. Miltersen, E. Petrank, and G. Tardos. "Linear hashing yields small buckets." *29th ACM Symp. on Theory of Computation*, pp. 465-474, May 1997. See also DIMACS Technical Report 97-25.

A. Gál, R. J. Lipton, E. Petrank. "Computing from partial solutions." In preparation.

J. Kilian and E. Petrank. "Identity escrow." Manuscript. Appeared as DIMACS Technical Report 97-28.

J. Kilian, E. Petrank, and G. Tardos. "Probabilistic checkable proofs with zero knowledge." *29th ACM Symp. on Theory of Computation*, pp. 496-505, May 1997. See also DIMACS Technical Report 97-27.

E. Petrank and C. Rackoff. "Message authentication of unknown variable length data." Manuscript. Appeared as DIMACS Technical Report 97-26.

Talks

MIT Theory Seminar. "Linear hashing."

Princeton Theory Lunch Seminar. "Linear hashing."

STOC'97 Conference. "PCP with zero knowledge."

MIT Cryptography and Information Security Seminar. "Identity escrow."

DIMACS Research and Education Institute - DREI 97 (Cryptography and Network Security). "Identity escrow."

Patents

J. Kilian and E. Petrank, "Identity escrow." International patent pending.

Rajmohan Rajaraman (1997-98) (SY in Networks)

Rajaraman's general research interests include network algorithms, parallel and distributed computing, on-line algorithms, combinatorics, and randomness. Currently, he is focusing on certain distributed network problems that are rapidly gaining significance with the advent of high-speed networks and the growing popularity of the Internet. One such problem is that of providing efficient access to shared objects in a distributed environment. Other problems include the design of routing algorithms that provide quality of service (QoS), virtual circuit routing, and load balancing in distributed environments.

Rajaraman has considered approximation algorithms for several NP-hard problems arising in facility location. In each problem that he has studied, one is given: (i) a set of locations at which we may open a facility, (ii) a set of client locations, together with their demands, that require to be serviced by a facility, and (iii) for every pair of locations the cost of shipping a unit of demand between the two locations. In the *k-median problem*, the objective is to determine a set of at most k locations at which to open a facility so as to minimize the total shipping cost. In the *facility location problem*, we are also given a cost of opening a facility at any location, and the objective is to determine a set of locations at which to open facilities so as to minimize the total shipping and opening costs. We differentiate between the uncapacitated and the capacitated variants of the above problems. In the uncapacitated case, an unlimited amount of demand can be shipped from any facility, whereas in the capacitated case, there is an upper bound on the total demand that can be shipped from any facility. Rajaraman proved that a simple local search heuristic yields polynomial-time constant-factor approximation bounds for the metric versions of the uncapacitated k-median and facility location problems. (For the k-median problem, the algorithms require a constant-factor blowup in the parameter k.) This local search heuristic was first proposed several decades ago, and has been shown to exhibit good practical performance in empirical studies. Rajaraman also extended the above results to obtain constant-factor approximation bounds for the metric versions of the capacitated k-median and facility location problems. This work was done jointly with Madhukar Korupolu and Greg Plaxton.

In other ongoing work with Koropolu and Plaxton, Rajaraman has designed a simple algorithm to determine a placement of objects in hierarchical networks. Their placement has the property that the expected access time of a request is within a small constant factor of the optimal. Moreover, the placement is obtained by a single bottom-up pass of the hierarchical network and hence can be implemented in a distributed manner. The algorithm has practical potential in the design of World Wide Web caching schemes and in allocating files in distributed filesystems.

Papers

Madhukar Korupolu, Greg Plaxton, and Rajmohan Rajaraman. "Analysis of a local search heuristic for facility location problems." To appear in the *Proceedings of the 9th Symposium on Discrete Algorithms*, January 1998.

Madhukar Korupolu, Greg Plaxton, and Rajmohan Rajaraman. "A simple distributed algorithm for placing objects in hierarchical networks." In preparation.

Talks

AT&T, December 3 1997. "Analysis of a local search heuristic for facility location problems."

Bell Labs, November 1997. "Analysis of a local search heuristic for facility location problems."

Symposium on Discrete Algorithms, San Francisco, January 1998. "Analysis of a local search heuristic for facility location problems."

Mona Singh (1995-97) (SY in Molecular Biology)
Currently at the Whitehead Institute, MIT

Singh has worked primarily on computational problems related to protein structure prediction. Biologists are interested in predicting protein structure because the structure of a protein provides insight into its biological function. Experimental approaches to determining a protein's structure are difficult and time-consuming. As a result, much recent research has focused on applying mathematical and computational techniques to predicting protein structure from one dimensional sequence data. A first step to tackling this complex problem is a solution to the structural motif recognition problem: given a known commonly occurring three-dimensional substructure, or motif, determine whether this motif occurs in a given amino acid sequence. With Bonnie Berger (V), Singh developed an iterative algorithm that uses randomness and statistical techniques to improve existing methods for recognizing protein structural motifs. Their algorithm is particularly effective in situations where large numbers of sufficiently diverse examples of the motif are not known. These are precisely the situations that pose significant difficulties for previously known methods. They have implemented their algorithm, and show that it is able to substantially improve performance on the coiled coil motif. In collaboration with biologists at the Whitehead Institute and Genentech Corporation, Berger and Singh and collaborators Peter Kim, James Berger, and Andrea Cochran applied their program to the histidine kinase protein family. Most members of the protein family are sensors that allow microorganisms to adjust to varying conditions in their environments. Their program identifies a common but previously undescribed coiled-coil region in the histidine kinase family. This work also suggests mechanisms by which the predicted coiled coil may mediate kinase regulation and signal transduction. There are some interesting medical applications, as some dangerous bacteria, such as those causing Whooping cough and Salmonella, need histidine kinase proteins to be fully virulent.

In the summer of 1997, Singh supervised Alantha Newman, a student in the REU program. With Martin Farach (PM, RU) and Scott Decatur (PD), Singh and Newman worked on a project studying simple models of protein folding. One way to simplify the protein folding problem is to model a protein as a sequence of hydrophobic (H) and polar (P) elements, with potential conformations limited to self-avoiding walks on lattices. They looked at a variety of related problems arising from a recent computationally intensive study done by biophysicists at NEC that suggests that structural regularities found in proteins (e.g., secondary structures or motifs) are the result of good "designability."

Papers

Bonnie Berger and Mona Singh. "An iterative method for improved protein structural motif recognition." *Journal of Computational Biology* (1997) 4(3):261–273. An earlier version is in the *1st Annual Conference on Computational Molecular Biology*, January 1997.

Mona Singh, Bonnie Berger, Peter S. Kim, James Berger, and Andrea Cochran. "Computational learning reveals coiled coil-like motifs in histidine kinase linker domains." Submitted.

Mona Singh, Peter Kim and Bonnie Berger. "Sequence analysis of retrovirus envelope proteins." In preparation.

Talks

University of Houston, October 1996. "An iterative method for improved protein structural motif recognition."

1st Annual Conference on Computational Molecular Biology, January 1997. "An iterative method for improved protein structural motif recognition."

Brown University, March 1997. "An iterative method for improved protein structural motif recognition."

Cornell University, April 1997. "An iterative method for improved protein structural motif recognition."

Yale University, April 1997. "An iterative method for improved protein structural motif recognition."

Whitehead Institute for Biomedical Research, August 1997. "An iterative method for improved protein structural motif recognition."

Elizabeth Sweedyk (1997-98) (SY in Molecular Biology, Joint DIMACS-U Penn Postdoc)

Sweedyk's research focuses on algorithms for various computational problems, particularly problems related to biology and databases. One problem she is currently working on is how to determine maximal conserved regions in the genomes of two different species. She recently developed a constant factor approximation algorithm for this problem. She is also working on clustering algorithms and is currently planning experiments to study various heuristics on large databases.

Lubos Thoma (1996-97)

Currently at Institute for Advanced Study

Thoma's primary research interests are in the areas of extremal combinatorics, probabilistic methods, 0-1 laws for random structures, and structural complexity.

A family of sets is called a Δ-*system* (respectively, a *weak* Δ-*system*) if any two sets have the same intersection (respectively, the same cardinality of intersection). Denote by $f(k,r)$ (respectively, $g(k,r)$) the largest size of a k-uniform family of sets which does not contain a Δ-system (respectively, a weak Δ-system) consisting of r sets. Erdős and Rado proved in 1960 that $(r-1)^k < f(k,r) < k!(r-1)^k$ and conjectured that for each r, there exists a constant C_r so that $f(k,r) < (C_r)^k$. This general conjecture is likely very hard. Denote by $F(n,r)$ (respectively, $G(n,r)$) the maximum number of subsets of an n-element set which do not contain a Δ-system (respectively, a weak Δ-system) consisting of r sets. Erdős and Szemerédi proved in 1978 that $G(n,3) > n^{\log n/4 \log\log n}$ and conjectured that there exists an ϵ so that $G(n,3) < (2-\epsilon)^n$, which was later proved. In collaboration with V. Rödl, Thoma gave a construction showing $G(n,r) \geq 2^{\frac{1}{3}\cdot n^{1/5} \log^{4/5}(r-1)}$ for n large enough. This improves the result of Erdős and Szemerédi.

In joint work with Alan Frieze (V) and Michel Karonski, Thoma studied matchings and hamiltonian cycles in sums of random trees. A *random tree* on the set $[n] = \{1,\ldots,n\}$ is a tree on this set chosen uniformly at random from the family of all trees on the set $[n]$. Let k be a positive integer. For trees $T_1,\ldots,T_k$, all of them on the set $[n]$, we define their sum as the graph on the vertex set $[n]$ and edge set being the union of edge sets of the trees $T_1,\ldots,T_k$, where the parallel edges coalesce. We denote by $\mathbf{ST}_n(k)$ the sum of k random trees on $[n]$. Let f be a mapping from $[n] \to [n]$. Let $D(f)$ be its associated functional digraph, i.e., the graph with vertex set $[n]$ and edges $(i, f(i))$, $i \in [n]$. Similarly to the sum of trees, we define the sum of mappings. We use $\mathbf{SM}_n(k)$ for the sum of k random mappings. Frieze showed in 1996 that $\mathbf{SM}_n(2)$ almost surely has a perfect matching and Cooper and Frieze showed in 1996 that $\mathbf{SM}_n(4)$ almost surely has a hamiltonian cycle. The problem of whether or not $\mathbf{SM}_n(3)$ almost surely has a hamiltonian cycle is one of the most important open questions in the theory of random graphs. Frieze, Karonski, and Thoma proved that

$$\lim_{\substack{n\to\infty \\ n \text{ even}}} Prob(\mathbf{ST}_n(1) \text{ has a perfect matching }) = 0$$

and

$$\lim_{\substack{n\to\infty \\ n \text{ even}}} Prob(\mathbf{ST}_n(2) \text{ has a perfect matching }) = 1$$

and that

$$\lim_{n\to\infty} \text{Prob}(\mathbf{ST}_n(5) \text{ has a hamiltonian cycle }) = 1.$$

In joint work with V. Rödl, Thoma studied the number of dependent arcs in acyclic orientations of graphs. An arc is called *dependent* if the reversal of its orientation creates an oriented cycle. For a graph G, let $d_{min}(G)$ be the minimum number of dependent arcs in an acyclic orientation of G. Similarly, we can define $d_{max}(G)$. To consider the minimum percentage of dependent arcs in an acyclic orientation of a graph, we define

$$r_{\chi,g} = \sup\{\frac{d_{min}(G)}{e(G)} \ : \ \chi(G) = \chi, \text{girth}(G) = g\} \ .$$

Rödl and Thoma determined the exact value of $r_{\chi,g}$. This result extends work by D.C Fisher, K. Fraughnaugh, L. Langley, and D.B. West.

Joel Spencer (V) and Thoma have developed a procedure characterizing limit values of probabilities for a random structure having a property from a given language L. The method is based on completions of almost sure theories. Let L be a (first or second order) language. For each $n \in \mathbf{N}$ we are given a probability space (Ω_n, μ_n), where Ω_n is a finite set of models of L and μ_n is a probabilistic measure on Ω_n. (In the particular application by Spencer and Thoma, we set (Ω_n, μ_n) to be the random graph $G(n,p)$.) We assume all subsets of Ω_n are μ_n-measurable. For $\psi \in L$ we write $\mu_n(\psi)$ for $\mu_n(\{M \in \Omega_n : M \models \psi\})$, the probability ψ holds in a "random" model. In 1991, T. Luczak and J. Spencer classified probabilities p with $p = \Theta(\log n/n)$ at which the $0-1$ law does not hold, i.e., thresholds in the neighborhood of the point where the random graph $G(n,p)$ becomes connected. Namely, the 0-1 law does not hold for the probabilities of the form $p_{k,l} = p_{k,l}(n) := (\ln n + l \ln\ln n + c + o(1))/(kn)$ where $l \geq k-1 \geq 0$. Spencer and Thoma have obtained a full characterization of limit probabilities for the random graph $G(n, p_{k,l})$ and first order properties of graphs. Due to the generality of the results, one can expect the procedure developed by Spencer and Thoma to play a role in characterizing limiting probabilities for other random structures.

Thoma worked with T. Łuczak (V) on the 0-1 law for some second order monadic theories. Let u_n denote a random mapping on the set $\{1,\ldots,n\}$. Łuczak and Thoma proved that the limit $\lim_{n\to\infty} Prob(u_n \models \psi)$ exists for every sentence ψ of the second order monadic language of mappings. Note that this was also proved independently using different techniques by A. Woods. Let $k \geq 3$ be an integer. The random k-regular graph $RR_k(n)$ is a probability space of all k-regular graphs on $\{1,\ldots,n\}$ with the uniform distribution. Łuczak and Thoma showed that there is a sentence φ in the second order monadic theory of graphs such that $\limsup_{n\to\infty} Prob(RR_k(n) \models \varphi) = 1$ while $\liminf_{n\to\infty} Prob(RR_k(n) \models \varphi) = 0$. That is, they showed nonconvergence for random regular graphs and the second order monadic theory of graphs. In 1994, S. Shelah (PM, RU) and J. Spencer (V) showed that for the probability p satisfying $pn \to c, 0 < c < \infty$, the limit $\lim_{n\to\infty} P(G(n,p) \models \psi)$ exists provided $c < 1$, and does not exist provided $c > 1$. Here, ψ is any sentence of the second order monadic language of graphs. Łuczak and Thoma proved a strengthening of the results of Shelah and Spencer.

Papers

V. Rödl and L. Thoma. "On the size of set systems on $[n]$ not containing weak (r, Δ)-Systems." DIMACS Report 96-57, to appear in *Journal Combinatorial Theory A*.

A. Frieze, M. Karonski, and L. Thoma. "On perfect matchings and hamiltonian cycles in sums of random trees." DIMACS Report 96-58, submitted to *SIAM J. Discr. Math.*

J. Spencer and Lubos Thoma, "On the limit values of probabilities for the first order properties of graphs." DIMACS Report 97-35, submitted to Proceedings of the DIMATIA-DIMACS Conference, *The Future of Discrete Mathematics.*

V. Rödl and Lubos Thoma. "On the number of dependent arcs in acyclic orientations of a graph." In preparation.

T. Łuczak and Lubos Thoma. "The nonconvergence for second order monadic properties of random regular graphs." In preparation.

T. Łuczak and Lubos Thoma. "The convergence law in the neighborhood of the double jump." In preparation.

Talks

Banach Center Minisemester on Selected Topics in Discrete Mathematics, Stefan Banach International Mathematical Center, Warsaw, Poland, September 1996. "On asymptotic packings and the random greedy algorithm."

Bell Labs (DIMACS Mixer), Murray Hill NJ, October 1996. "On the complexity of cover graph recognition for finite lattices."

Department of Mathematics, Emory University, Atlanta GA, October 1996. "On the size of set systems not containing weak Δ-systems."

Penn Combinatorics Seminar, Department of Mathematics, University of Pennsylvania, Philadelphia PA, February 1997. "Acyclic orientations of graphs, finite posets, and lattices."

Combinatorics, Complexity, and Discrete Probability Seminar, School of Mathematics, Institute for Advanced Study, Princeton NJ, March 1997. "On set systems not containing (weak) Δ-systems."

Seminar on Discrete Probability, DIMACS, Rutgers University, Piscataway NJ, May 1997. "Limit probabilities of the first order properties of graphs."

Special Year on Logic and Algorithms - One Year Later Workshop, DIMACS, Rutgers University, Piscataway NJ, July 1997 . "On the limit probabilities of the first order properties of graphs."

Géza Tóth (1997-99) (DIMACS-IAS Postdoc)

Tóth's research has focused on Euclidean Ramsey theory, geometric graphs, unit distance graphs and art gallery theorems.

Tóth has collaborated with Pavel Valtr (PD) to find an improvement in the famous Erdős-Szekeres Theorem. In 1933, Esther Klein raised the following question. Is it true that for every n there is a least number $g(n)$ such that among any $g(n)$ points in general position in the plane there are always n in convex position? This question was answered in the affirmative in a classical paper of Erdős and Szekeres. In fact, they showed that

$$2^{n-2} + 1 \leq g(n) \leq \binom{2n-4}{n-2} + 1.$$

This theorem turned out to be the starting pont of many new developments in graph theory and combinatorial geometry, and it was also a rediscovery of Ramsey's theorem which has many analogues, extensions, and applications in several fields of mathematics. The lower bound of the Erdős-Szekeres theorem, $2^{n-2}+1$, is sharp for $n = 2, 3, 4, 5$ and has been conjectured to be sharp for all n. However, the upper bound, $\binom{2n-4}{n-2}+1 \approx c\frac{4^n}{\sqrt{n}}$, was not improved for 60 years. Recently, Chung and Graham managed to improve it by 1. Shortly after, Kleitman and Pachter showed that $g(n) \leq \binom{2n-4}{n-2} + 7 - 2n$. Tóth and Valtr obtained a further improvement, roughly by a factor of 2 They showed that any set of $\binom{2n-5}{n-2}+2$ points in general position in the plane contains n points in convex position.

Let $\mathcal{F}$ denote a family of pairwise disjoint convex sets in the plane. $\mathcal{F}$ is said to be in *convex position*, if none of its members is contained in the convex hull of the union of the others. For any fixed $k \geq 3$, J. Pach and Tóth have estimated $P_k(n)$, the maximum size of a family $\mathcal{F}$ with the property that any k members of $\mathcal{F}$ are in convex position, but no n are. In particular, for $k = 3$, they have improved the triply exponential upper bound of T. Bisztriczky and G. Fejes Tóth by showing that $P_3(n) < 16^n$. Tóth has also, for any fixed $k \geq 5$, given a linear upper bound on $P_k(n)$. This improves the quadratic bounds of Bisztriczky and Fejes Tóth.

Pach and Tóth have developed Erdős-Szekeres-type theorems for segments and non-crossing convex sets. They proved that there is a function $N(n)$ with the following property. If $\mathcal{F}$ is a family of at least $N(n)$ plane convex sets with non-empty interiors, such that any *two* members of $\mathcal{F}$ have at most two boundary points in common and any *three* are in convex position, then $\mathcal{F}$ has n members in convex position. This result generalizes a theorem of T. Bisztriczky and G. Fejes Tóth. The statement does not remain true if two members of $\mathcal{F}$ may share four boundary points. This follows from the fact that there exist infinitely many straight-line segments such that any three are in convex position, but no four are. However, there is a function $M(n)$ such that every family of at least $M(n)$ segments, any *four* of which are in convex position, has n members in convex position.

A three-dimensional *grid drawing* of a graph G is a placement of the vertices at distinct integer points so that the straight-line segments representing the edges of G are pairwise non-crossing. J. Pach, T. Thiele, and Tóth have shown that for any fixed $r \geq 2$, every r-colorable graph of n vertices has a three-dimensional grid drawing that fits into a box of volume $O(n^2)$. The order of magnitude of this bound cannot be improved. The results are applicable to the variety of problems arising in the important area of graph drawing.

A *geometric graph* is a graph drawn in the plane so that the vertices are represented by points in general position, the edges are represented by straight line segments connecting the corresponding points. Improving a result of Pach and Törőcsik, Tóth and Pavel Valtr (PD) showed that a geometric graph on n vertices with no $k+1$ pairwise disjoint edges has at most $k^3(n+1)$ edges. On the other hand, they constructed geometric graphs with n vertices and approximately $\frac{3}{2}(k-1)n$ edges, containing no $k+1$ pairwise disjoint edges.

Papers

G. Tóth, "Finding convex sets in convex position." In preparation.

J. Pach, T. Thiele, and G. Tóth. "Three-dimensional grid drawings of graphs." In: *Proceedings of Conference on Discrete and Computational Geometry: Ten Years Later (Mount Holyoke*, 1996; B. Chazelle et al., eds.), *Contemporary Mathematics*, AMS, Providence (accepted). DIMACS TR 97-73.

J. Pach and G. Tóth. "A generalization of the Erdős-Szekeres theorem to disjoint convex sets." *Discrete and Computational Geometry* (accepted). DIMACS TR 97-66.

G. Tóth and P. Valtr. "Note on the Erdős-Szekeres theorem." *Discrete and Computational Geometry* (accepted). DIMACS TR 97-31.

G. Tóth and P. Valtr. "Geometric graphs with few disjoint edges." Manuscript.

J. Pach and G. Tóth. "Erdős-Szekeres-type theorems for segments and non-crossing convex sets." Manuscript.

Talks

Geometry Seminar, Courant Institute, Oct. 1997. "A generalization of the Erdős-Szekeres theorem to convex sets."

Seminar on Combinatorial Computing, City University of New York, Oct. 1997. "A small improvement of the Erdős-Szekeres theorem."

924th AMS Meeting, Montreal, Canada, Sept. 1997. "A generalization of the Erdős-Szekeres theorem to convex sets."

Pavel Valtr (1996-97)

Currently at Charles University, Prague

Valtr worked on extremal graph theory, geometric graphs, and robot motion planning.

The famous theorem of Turán shows that a graph with low independence number contains many edges. Intuitively, a graph with low independence number also contains relatively small subgraphs with many edges. This intuition was quantitatively expressed in a conjecture of Erdős who considered subgraphs on $\approx \sqrt{n}$ vertices in a graph on n vertices with independence number smaller than $\approx \sqrt{n}$. His conjecture giving an asymptotically best bound was recently proved by Alon. Alon extended the conjecture to a case when "$\approx \sqrt{n}$" is replaced by an arbitrary number between $\log n$ and n, and proved the extended conjecture in some special cases. Valtr proved the conjecture in general. Moreover, his result considerably extends Alon's conjecture in a natural way. It asymptotically answers the following question for all n, m, k: given a graph on n vertices with independence number m, how many edges must be contained in at least one subgraph on k vertices?

In robot motion planning, the environment in which a robot can move often corresponds to a subset of R^d, called *free space*. Typically, it is easy to move a robot along a line segment contained in the free space. This leads to a question in which cases one can find a constant number of points (*milestones*) in the free space, such that every point in the free space can be moved along a segment to a milestone. With pre-computed paths joining the milestones, this would give a fast and simple way to find a path of the robot between any two points in the free space. In two dimensions, Matoušek and Kalai showed how to find the set of milestones for any simply connected free space such that every point in the free space is connected, say, to at least $1/100$ of the points in the free space by a line segment. An analogous result was conjectured in more dimensions. Valtr disproved this conjecture by constructing a "bad" free space in three dimensions.

A *geometric graph* is a graph drawn in the plane so that the vertex set consists of points in the plane and the edge set consists of line segments joining the vertices. Two of the basic questions about geometric graphs ask how many edges (asymptotically) can be contained in a geometric graph (a) with no k pairwise disjoint edges, (b) with no k pairwise intersecting edges, where k is a constant. With condition (a), the question was partially answered by Pach and Törőcsik who in 1993 gave an upper bound linear in n, the number of vertices. Valtr strengthened this result by showing a linear bound also when condition (a) is essentially weakened. As a consequence, he obtained an upper bound $O(n \log n)$ on the number of edges in geometric graphs satisfying condition (b). This improved the previous upper bound $O(n \log^{2k-6} n)$ proved jointly by Agarwal, Aronov, Pach, Pollack, and Sharir. The results are related to the many practical questions arising in the field of graph drawing.

As mentioned above, Pach and Törőcsik in 1993 bounded the number of edges in a geometric graph on n vertices with no k pairwise disjoint edges by a function linear in n. Their bound is $(k-1)^4 n$. Géza Tóth (PD) and Valtr lowered this bound to $\frac{4}{21}k^3 n$. They also improved the previous almost trivial lower bound $(k+0.5)n$ by a constant factor

In 1996, Károlyi, Pach, and Tóth wrote a paper considering a new type of problem for geometric graphs. They started to investigate classical Ramsey-type questions naturally modified for geometric graphs. These results have been extended by the above authors together with Valtr. The work makes use of results and techniques in three different fields (extremal graph theory, combinatorial geometry, and Ramsey theory).

Motivated by a problem concerning convex polytopes, Kalai raised the problem if, for any $\varepsilon > 0$, it is possible to find a finite collection of disjoint ellipsoids in R^d of diameter at most 1 such that these elipsoids cover a disk of radius 1 (say) after they are enlarged by factor $1+\varepsilon$. Several researchers asked if this is possible in the special case of disks in the plane. Jiři Matoušek (V) and Valtr showed that this is not possible.

Papers

P. Valtr. "On the density of subgraphs in a graph with bounded independence number." To appear in *J. Comb. Theory (Ser. B)*.

P. Valtr. "On galleries with no bad points." In preparation.

Gy. Károlyi, J. Pach, G. Tóth, and P. Valtr. "Ramsey-type results for geometric graphs." II. *Proc. of the 13th ACM Symp. on Computational Geometry*, Nice, June 4-6, 1997. Also to appear in *Discrete and Computational Geometry.*

P. Valtr. "On geometric graphs with no k pairwise parallel edges." To appear in *Discrete and Computational Geometry.*

P. Valtr. "Graph drawings with no k pairwise crossing edges." To appear in *Graph Drawings '97, Rome*, September 1997.

G. Tóth and P. Valtr. "Note on the Erdős-Szekeres theorem." To appear in *Discrete and Computational Geometry.*

G. Tóth and P. Valtr. "Geometric graphs with few disjoint edges." In preparation.

P. Valtr. "Distributions producing random Voronoi diagrams of large complexity." In preparation.

J. Matoušek and P. Valtr. "Packing disks of different radii." In preparation.

Talks

Rider University, Lawrenceville, NJ., AMS meeting, October 6, 1996. "A positive fraction Erdős-Szekeres theorem."

Future of Discrete Mathematics, Stiŕin. May 20, 1997. "Several recent results on geometric graphs."

DIMACS Mixer Series, DIMACS. September 13, 1996. "Several problems related to the Erdős-Szekeres theorem."

Courant Institute, Discrete Lunch Bunch. October 3, 1996. "Ramsey-type results for planar point configurations".

DIMACS Seminar, DIMACS. October 24, 1996. "On the probability that n random points are in convex position".

Courant Institute, Geometry Seminar, October 29, 1996. "A positive fraction Erdős-Szekeres theorem."

R. Pollack's lecture, Courant Institute. April 16 and 23, 1997. "Generalized DS-sequences and geometric graphs".

University of Alaska, seminar talk, Fairbanks. July 17, 1997. "Several recent results on geometric graphs."

Shibu Yooseph (1997-98) (SY in Molecular Biology - DIMACS-U. Penn Postdoc)

Yooseph's research interest is in the design and analysis of algorithms. He is currently working on problems in computational biology and also in computer networks. His research in this area has focused on the problem of determining evolutionary relationships between objects from a given set. These objects can be genes, proteins, biological species, etc. The knowledge of evolutionary relationships is useful in the prediction of protein structure, drug design, sequence alignment, etc. His previous results in this area have included new algorithms for constructing evolutionary trees (under various criteria like compatibility and parsimony) in the presence of special kind of data (*polymorphic character data*) and a new model (*local consensus model*) for extracting information from a given set of evolutionary trees. His current research has focused on performance analysis of various tree construction methods. This kind of analysis is done using computer simulations. The analysis, in addition to providing a better understanding of currently used methods, has also resulted in new methods for obtaining better estimates of evolutionary trees.

Evolutionary tree reconstruction is a challenging problem with important applications in biology and linguistics. With S. Nettles, K. Rice, and T. Warnow (V), Yooseph has explored hybrid tree construction methods. He and his collaborators have explored the problem of reconstructing very large evolutionary trees from biomolecular sequences. Using both analytical and experimental techniques, they showed that each of the existing popular methods has significant problems in obtaining accurate reconstructions of some trees. As suggested by their experimental results, they then proposed a new method that combines the existing methods into hybrid methods. They studied one such hybrid experimentally, and showed that in significant parts of the parameter space, it achieves results that are equal to or better than the best of its components. Their experimental performance study examined more than a thousand sets of sequences simulated on more than a hundred different model trees.

In biology, one of the most promising approaches to tree reconstruction is to find the "maximum parsimony" tree, while in Linguistics, the use of the "maximum compatibility" method has been very useful. However, these problems are NP-hard, and current approaches to solving these problems amount to heuristic searches through the space of possible tree topologies (a search which can, on large trees, take months to complete). Maria Bonet (PD), Mike Steel (V), Tandy Warnow (V), and Yooseph developed a new technique, *Optimal Tree Refinement*, for reconstructing very large trees. Their technique is motivated by recent experimental studies which have shown that certain polynomial time methods reliably return *contractions* of the true tree. They have studied the use of this technique in solving parsimony and compatibility, and obtained both hardness results and polynomial time algorithms.

Daniel Huson (V), S. Nettles, L. Parida, Tandy Warnow (V), and Yooseph have developed a divide-and-conquer approach to reconstructing large evolutionary trees. The approach in essence *bootstraps* the performance of any given phylogenetic method. They have found a particular variation of the technique based upon the popular distance method *Neighbor-Joining*, and showed experimentally the improvement in performance they obtained over naive Neighbor-Joining. They also have have noted, through their experimentation with different versions of the general method, suggestions for which techniques might work to reduce different types of errors.

Yooseph's research on computer networks has focused on the rate allocation problem in ATM networks. With Martin Farach (PM, RU), S. Muthukrishnan (PM, BL), and Lisa Zhang (PM, BL), Yooseph has studied the problem of allocating available bandwidth in ATM networks to different sessions with specific demands. The challenges in this problem include the design of fast algorithms which can be implemented in a distributed environment. These algorithms typically meet session demands either completely or partially so that the final allocation satisfies a certain criterion (for instance, the *max-min* fairness criterion). Yooseph and his collaborators have been studying the convergence of existing rate allocation algorithms and have also been looking into the design of new algorithms.

Papers

S. Nettles, K. Rice, T. Warnow, and S. Yooseph. "Hybrid tree construction methods." Submitted to the *Second Annual International Conference on Computational Molecular Biology*, 1998.

M. Bonet, M. Steel, T. Warnow, and S. Yooseph. "Better methods for solving parsimony and compatibility." Submitted to the *Second Annual International Conference on Computational Molecular Biology*, 1998.

D. Huson, S. Nettles, L. Parida, T. Warnow, and S. Yooseph. "A divide-and-conquer approach to tree reconstruction." Submitted to *Algorithms and Experiments*, 1998.

3.2 Graduate Student Fellows: February 1997 - January 1998

3.2.1 DIMACS-Supported Graduate Students at Rutgers University, February 1997 - January 1998

Sarmad Abbasi

Abbasi worked under the mentorship of Endre Szemeredi (PM, RU) on a problem in combinatorial pattern matching. Let Σ be a finite alphabet with C letters. For any two strings $x = x_0 \cdots x_{n-1}$ and $y = y_0 \cdots y_{n-1}$ of length n, we let $S(x,y)$ denote the size of the longest common consecutive substring between x and y; that is, $S(x,y)$ is the largest k such that, $x_i \cdots x_{i+k-1} = y_j \cdots y_{j+k-1}$ for some $i, j \in \{0, \ldots, n-1\}$. (Here the indices are modulo n.) Abbasi showed that for x and y chosen uniformly among all possible strings of length n, $S(x,y)$ is highly concentrated around $2\log_C n$. More precisely, for any $a \geq 1$ $\Pr[|S(x,y) - 2\log_C n| > a] \leq \frac{18C^{-a}}{n^2}$. This enabled him to show that the expected value of $S(x,y)$ is within a constant additive term of $2\log_C n$.

Abbasi also worked on embedding problems in graph theory, in particular the so called "Bollobas-Komlos" Conjecture. Bollobas and Komlos have conjectured the following result. For any fixed k and Δ and $\beta > 0$, there is an α and a threshold n_0 such that the following statement holds for any $n > n_0$. Let G be any graph on n vertices with minimum degree $\frac{k-1}{k}n + \beta n$. Then G contains as subgraphs all n-vertex k chromatic, maximum degree Δ graphs whose bandwidth is at most αn. He proved this conjecture for $k = 2$. Furthermore, he showed that the conjecture is true for all k provided the notion of bandwidth is changed to a closely related notion called k-partite bandwidth. Let H be a k colorable graph with color classes $V_1, \ldots, V_k$. We say that H has *k-partite bandwidth* b if one can order each color class

$$V_i = v_i^1, v_i^2, \ldots, v_i^{k_i}$$

such that, if vertex v_i^s is adjacent to vertex v_j^t then $|s - t| \leq b$. Alternatively, there are functions $f_i : V_i \to |V_i|$ such that if $v \in V_i$ and $w \in V_j$ are two adjacent vertices then $|f_i(v) - f_j(w)| \leq b$.

Papers

S. Abbasi. “Longest common consecutive substring in two random strings.” DIMACS Technical Report. 1997.

Bernardo Abrego

Abrego worked on combinatorial geometry together with Silvia Fernandez, and under the mentorship of József Beck (PM, RU). He obtained several results, all of them regarding the maximum number of equilateral triangles that can be determined from a set of n points in some Euclidean space, and in some cases for sets of n points satisfying a certain geometric constraint. Highlights of the previous include the upper bounds $O(n^{2.1})$ and $O(n^{13/7})$ for the number of triangles, and unit triangles in 3 space; the precise asymptotic expression for the number of unit triangles in d-space under the restriction of points in general position; and the reduction of optimal configurations on the plane to subsets of the equilateral triangle lattice. Abrego also worked with Beck and Fernandez on geometric discrepancy problems.

Paper

B. Abrego and S. Fernandez. “On the number of equilateral triangles on euclidean spaces.” In preparation.

Talk

Combinatorial Geometry session of the 3rd Joint Meeting AMS-SMM. Oaxaca, Mexico, December 3-7. “On the number of equilateral triangles on euclidean spaces.”

Subramanian Balakrishna

Balakrishna is working on problems involving protein contact maps, under the direction of Joel Lebowitz (PM, RU). In collaboration with Michele Vendruscolo, Lebowitz, and Eytan Domany, Balakrishna is studying the statistical properties of physical contact maps. A *contact map* is a coarse-grained representation of a protein's structure, amenable to numerical studies of folding. Balakrishna and his collaborators have shown that the number of contact maps corresponding to the possible configurations of a polypeptide chain of N amino acids, represented by N step Self-Avoiding Walks, scales exponentially with N. This is done in general dimension D by establishing lower and upper bounds on the number of maps. Balakrishna, et al. also have exact numerical enumeration results on the square and triangular lattices in D=2. Based on the enumeration data, they have analyzed interesting properties of these maps pertinent to the mechanism of protein folding.

Vicky Choi

Choi worked on combinatorial geometry under the direction of Rafael Ostrovsky (PM, BC). Let $g(n)$ be the smallest integer such that any subset of $g(n)$ points in the plane in general position must always contain the vertices of a convex n-gon. In 1935, Erdős and Szekeres showed that $g(n) \leq \binom{(2n-4)}{(n-2)} + 1$ and conjectured that $g(n) = 2^{n-2} + 1$. In 1960, they showed the lower bound $2^{n-2} + 1 \leq g(n)$. However, the bounds have stood unchanged until recently, when Chung and Graham improved on the upper bound, by 1. Although the improvement in the paper was tiny, it led to the hope that the methods might suggest substantial reductions in the upper bound. In fact, Geza Tóth (PD) and Pavel Valtr (PD) improved the upper bound. Choi worked on the problem of finding additional improvements in the upper bound, in collaboration with Bill Steiger (PM, RU) and Valtr.

Rick Desper

Phylogeny reconstruction is a central problem in the field of computational biology. Desper studied a widely-used model, the Cavender-Farris tree model of evolution. With Martin Farach (PM, RU), Sampath Kannan (V), and Andris Ambainis (V), Desper considered the reconstruction problem from the standpoint of a metric introduced on the space of Cavender-Farris trees. The metric, variational distance, is the L_1 distance between the two probability distributions. Letting n denote the number of living species, the number of leaves of the tree, and k the number of samples, which biologically is the number of sites of DNA in n aligned sequences of DNA, Desper and his collaborators showed that no algorithm can approximate a tree with an expected variational distance better than $\sqrt{n/k}$. They showed an algorithm which achieved this variational distance, up to a constant based upon minimum edge probabilities.

Paper

R. Desper, M. Farach, S. Kannan, and A. Ambainis. "Nearly tight bounds on the learnability of evolution." In preparation.

Talk

38th Symposium on Foundations of Computer Science, Miami, Florida. 1997. "Nearly tight bounds on the learnability of evolution."

Paul Dreyer

Given an $n \times n$ square grid, there are the outlines of n^2 1×1 squares, $(n-1)^2$ 2×2 squares, and so on. What is the fewest number of edges that can be removed from the grid such that there are no complete outlines of squares remaining? Using techniques from tiling problems and other graph theoretic methods, Dreyer and Therese Biedl (GS, RU) solved the problem for all n and proved similar results for rectangular grids for which the side lengths differ by no more than six. They also introduced another version of the problem where they asked for the fewest number of edges to "break" all of the rectangles in a grid.

Paper

P. Dreyer and T. Biedl. "Rectangle breaking in grids." *Congressus Numerantium*, 1997.

Talk

The 28th Southeastern Conference on Combinatorics, Graph Theory and Computing. Boca Raton, Florida. 1997. "Rectangle breaking in grids."

Amr Elmasry

Elmasry worked on problems of computational geometry under the direction of Diane Souvaine (PM, RU). A lot of work has been done on calculating shortest paths in polygonal regions. In this classic shortest path work, the shortest path between two points is the shortest path achievable by a piece of string, i.e., it can wrap around corners. For work in wireless communications, there CAN be some wrapping around vertices, but the "shortest path" achievable by a signal tends to travel in straight lines, and then be reflected off a wall or an obstacle where the angle of incidence equals the angle of reflection. Elmasry worked to develop a package that calculates multiple reflections. Elmasry implemented some ray tracing subroutines in C language to help in building the required system. He also suggested how multireflection methods could enhance the traditional ray tracing strategy.

Linchun Gao

Under the direction of Alex Kogan (PM, RU), Gao worked on computer implementation of the Logical Analysis of Data (LAD) methodology aimed at detecting patterns from large data sets. The computer implementation of LAD consists of a number of combinatorial, enumeration, and combinatorial optimization algorithms based on specially designed data structures. The main components of the present implementation of LAD are a binarization module, a pattern generation module, and a theory construction module. These modules require the input of a significant number of control parameters guiding the selection of the processing subroutines as well as the execution of them. The values of the control parameters may have a strong influence on the running time and accuracy of the results. Gao worked on parameter experimentations. There are nine parameters in the binarization step. The two most sensitive parameters are the confidence interval around each cut point and the minimal separability of each pair of points (optimization). She used three data sets to do the experiments. They are a data set for the Chinese economy, an oil data set and a depress data set. With the increment of confidence interval, the number of final cut points increased dramatically for the China and oil data sets, but this was not true for the depress data set. Also, the min error and bin time increased along the increment of confidence interval for the China and oil data sets, but not for the depress data set. As to the minimal separability of each pair of points, for all three data sets, the number of final cut points and bin time increased dramatically when this parameter increased, but the min error decreased meanwhile. Gao did further work on the application aimed at finding some general principles in Chinese economical development by processing the data they currently have and helping the decision maker to find strategies to promote further development of the Chinese economy. The data set is composed of observations collected from 29 provinces ranging from the years 1985 to 1995. She helped to develop a code which can generate a "training set" and a "test set", can run LAD automatically, get patterns, and check conjectures.

Daniel Kudenko

The goal of Kudenko's research was to improve the performance of standard machine learning algorithms on sequence categorization tasks, where a good featurization is not known from the beginning. DNA sequence categorization is a good example for such a domain. Working under the direction of Haym Hirsh (PM, RU), Kudenko developed a greedy set covering algorithm that generates boolean features which test for the existence and frequencies of certain sets of *n*grams that are shared by many positive training examples. The implementation of these features is based on a variation of the suffix tree data structure. The performance of these features is compared to straightforward feature representations, e.g., where the feature set represents the existence of single *n*grams that occur somewhere in the training data. Kudenko continued to work to improve the algorithm and to test it on biological datasets and other sequential data. The major improvement that he achieved was the change of the choice of the seed (i.e., starting element) for a set cover. In a previous version of the feature generation algorithm, it had been picked randomly. It turned out that a greedy choice of that seed (i.e., picking the seed that yields the biggest cluster), improved performance significantly. Kudenko tested the feature generation algorithm on two types of domains: DNA sequences and texts. Using the generated features alone didn't always improve the performance as he wished; therefore he tried to append the generated features to straightforward representations. There was an improvement in error rates almost everywhere, sometimes quite dramatic.

Jie Lai

Lai worked on WiPPET - Wireless Propagation & Protocols Evaluation Testbed - together with Narayan Mandayam, Andy T. Ogielski (PM, RU), Roy Yates, Owen Kelly (PD), Jignesh Panchal (GS, RU) and David Pandian (GS, RU). They are developing a comprehensive suite of reusable software modules (classes) for modeling and simulations of wireless communications networks. The principal classes provide modeling of geography, mobility, radio propagation and interference, radio resource management protocols, networking protocols and packet traffic sources. Subclassing allows one to model various specific systems with a minimum of programming. Their goal is to achieve high performance parallel simulations, both for a large number of entities and for depth of modeling detail. Fast parallel simulations will allow exhaustive studies of performance and quality of service at the system level, while modeling interference at the packet level. Existing discrete event simulators do not accommodate systems of such complexity. This work is a part of a multi-institutional project "Scalable Self-Organizing Simulations (S3)" funded by DARPA/ITO. S3 aims to achieve a generational transition in performance and manageability of simulations of telecommunication systems.

Paper

Jignesh Panchal, Owen Kelly, Jie Lai, Narayan Mandayam, Andrew T. Ogielski and Roy Yates. "Wippet, a virtual testbed for parallel simulations of wireless networks." Submitted to *PADS '98* (Banff, Alberta, Canada, May 26-29, 1998).

Vincenzo Liberatore

Multipaging is the paging problem when more than one page can be requested at each step. In the uniform cost model, a paging algorithm is charged for the number of pages loaded from disk to fast memory. Uniform multipaging is similar to the on-line paging problem, except that more than one page can be requested in one step. Liberatore, working under the direction of Martin Farach (PM, RU), has established a general reduction from uniform multipaging to paging. As a consequence, he obtained the first strongly competitive randomized algorithm for uniform multipaging.

Paper

Vincenzo Liberatore. "Uniform multipaging reduces to paging." DIMACS TR 97-69.

Talk

DIMACS Mixer, AT&T Labs, Florham Park, NJ, October 1997. "Uniform multipaging."

Daniel Lloyd

The goal of Lloyd's research was to find an approximation algorithm which calculates the best fit ultrametric based on a dissimilarity matrix. This algorithm would have applications in DNA analysis: inferring phylogenies, and clustering in general. Several methods were explored including general heuristical approaches, and finally relaxations based on semidefinite programs. Current work in semidefinite programming by several other researchers has shed new light on possible approaches to this problem. Lloyd continues his research with Martin Farach (PM, RU) on this problem, focusing on using relaxations based on semidefinite programs.

Sachin Lodha

Lodha worked on a cryptography and security project under the guidance of Rafail Ostrovsky (PM, BC) and William Aiello (PM, BC). The topic of the research was Certificate Authority (CA) and related issues in public key cryptography, and involved a mixture of theoretical work using combinatorics and number theory and some programming.

Jignesh Panchal

Panchal worked, under the direction of Andy Ogielski (PM, RU), on the DIMACS DARPA project S3 ("Scalable Self-Organizing Simulation"). The main goals of this project are development and implemention of network modeling techniques and parallel simulation frameworks to facilitate construction of large models and exploit scalability to reduce the simulation time. He used a parallel simulation engine, GTW (Georgia Time Warp), and TeD, the metalanguage developed by Georgia Tech, as a tool. He implemented WiPPET (Wireless Propagation & Protocols Evaluation Testbed). It is capable of simulating all the functionality of a Wireless voice network. He continues to work on simulation of higher protocol layers, realistic packet traffic, etc.

Papers

Jignesh Panchal, Owen Kelly, Jie Lai, Narayan Mandayam, Andrew T. Ogielski and Roy Yates. "Wippet, a virtual testbed for parallel simulations of wireless networks." Submitted to *PADS '98* (Banff, Alberta, Canada, May 26-29, 1998).

Owen Kelly, Jie Lai, Narayan Mandayam. Andrew T. Ogielski, Jignesh Panchal and Roy Yates. "Developers report." Wireless Information Network Laboratory, Rutgers. WINLAB TR KLM97, 1997. In preparation.

James Cowie, Richard Fujimoto, Owen Kelly, Jie Lai, Andy Ogielski, Jignesh Panchal and Kalyan Perumalla. "Scalable virtual testbeds for end-to-end modeling of telecommunication networks." *IEEE Communications Magazine*, 1998. In preparation.

Talk

WINLAB Research Review, October 21, 1997. "Virtual testbed for parallel simulations of wireless networks."

Gregory Perkins

A goal in the now huge field of wireless communications is a sound and fast method for signal strength prediction in urban and in building situations. A proposed and somewhat proven method is to use ray tracing to simulate the propagation of the radio signal. Perkins built his own modeling method using ray-tracing which is a merging of several ideas from the field and some of his own. Currently, this system is complete and is being tested for its accuracy and speed. Perkins continues to work on a variety of problems. Ray tracing has fast become the accepted method for radio wave propagation prediction when one is trying to predict signal strengths at various locations in an urban or in a building situation. The accuracy of the method has been well documented by several research groups which shows that ray-tracing prediction results lie roughly around 5-10dB of the mean signal strength when compared to measurements performed on sight by hand. This is good but is still a fair bit off desired accuracy levels, especially since some predicted values are very wrong. A second problem deals with an upper bound for the number of transmitters needed to cover an area. Given a layout of an urban area or a floor of a building, determine an upper bound for the number of transmitters necessary to provide complete coverage, where coverage is defined as every point in the layout attains a mean signal strength of at least beta from at least one of the transmitters. Another problem Perkins is working on is returning the number of intervening obstacles. Given a layout of obstacles (all convex), two locations (xt, yt) and (xr, yr), return the number of intervening obstacles to a line of sight path. Preprocessing on the layout based upon the location (xt, yt) is allowed. Another problem is characterizing similar reflection regions. Given a layout of convex obstacles and a transmitter location, a *reflection region* is defined to be the set of points that all have the same set of reflection paths. A *reflection path* is a path from the transmitter to the point in question that reflects off of k walls along the way, where $k \geq 0$. A *path* is made up of the set of walls it reflects off of. Questions arise in the wireless field as to how many of these regions can exist for an arbitrary layout and, more importantly, how many big/small regions does one expect to arise in the average case and the worst case.

Lorant Porkolab

Porkolob obtained with Leo Khachiyan (PM, RU) a new complexity bound for the problem of maximizing a given linear objective function over the set of integral points contained in a given convex semi-algebraic set (which also includes integer semidefinite programming as a special case).

Papers

L. Porkolab and L. Khachiyan. "Computing integral points in convex semi-algebraic sets." 1997. In preparation.

L. Khachiyan, L. Porkolab. "Computing integral points in convex semi-algebraic sets." *38th Annual Symposium on Foundations of Computer Science*, pp. 162-171.

L. Porkolab. "On the complexity of real and integer semidefinite programming." Technical Report DCS-TR-334, Department of Computer Science, Rutgers University, 1997.

Talk

38th Annual Symposium on Foundations of Computer Science, Miami Beach, October 1997. "Computing integral points in convex semi-algebraic sets."

Jason Schattman

Schattman worked, under the guidance of Alex Kogan (PM, RU), on the Logical Analysis of Data (LAD), a relatively new methodology aimed at detecting structural information about data sets. Its goal is to detect logical patterns which distinguish observations in one class from all other observations. In the LAD framework, we are given a data set in which some of the data are classified as positive and the rest negative. LAD is a multimoduled algorithm that seeks to ferret out the variables in the data set that are most relevant to determining the data's classifications. LAD operates in three stages. It first binarizes all numerical variables so that all data are converted to 0 or 1. Secondly, LAD processes the binary data to find a set of boolean conjunctions that have certain properties with respect to the data. These conjunctions are called patterns. A *pattern* must output a 1 for at least one (and ideally many) positive data points, and a 0 for most (ideally all) negative data points. Finally, LAD selects a subset of the patterns that collectively outputs a 1 on a sufficiently large number of positive data points. This subset, called a *theory*, is used to classify future data points whose classification is unknown. The success or value of a theory lies in the accuracy of its classification predictions. An implementation of LAD in C++ has been succesfully tested on numerous data sets. Schattman implemented in C++ an alternative algorithm proposed by Kogan for the pattern generation module of LAD. The current LAD implementation computes a complete enumeration of all patterns up to a fixed degree. Generating patterns of high degree in this way is very expensive computationally, for the number of possible patterns is exponential in the number of binary variables. Schattman's goal was to implement a pattern finding algorithm that would find high-degree, high-quality patterns in a computationally efficient way. The algorithm Schattman implemented builds patterns one conjunct at a time, adding conjuncts until the conjunction covers sufficiently few negative points. While the pattern is being built, the next conjunct is chosen from the set of candidate conjuncts using one of several heuristics. This approach to pattern generation made possible the identification of very high degree patterns, which could not have been found using current LAD in reasonable time. On most data sets, the algorithm required no more than a few seconds to find enough patterns to cover the entire positive data set. Most importantly, Kogan and Schattman found that many of the patterns found also had very high coverage of the positive data points, which translated into high prediction accuracy on future unclassified data. Although many individual long patterns could be found with high positive coverage, the theories that were formed from the entire set of such patterns often did not predict as accurately as the theories found by the current LAD. Kogan and Schattman conjectured that requiring complete coverage of the positive data set contributed to overfitting of the data. Therefore they proposed that the new algorithm serve not as a theory former but as a source of long patterns for the main LAD program. The new algorithm is being embedded into the current LAD source code so that the user can input an option to augment his run with long pattern generation. New versions of the algorithm may include randomization to avoid local optima, and differing heuristics for computing the next conjunct of a pattern.

Lijie Shi

The Logical Analysis of Data (LAD) is a new methodology aimed at detecting structural information about datasets. It was first proposed by Peter Hammer (PM, RU) and developed by the RUTCOR LAD research group. The main characteristic of LAD is the detection of logical patterns which distinguish observations in one class from all the other observations. How to find a good pattern was the main task for Shi in her research under the guidance of Hammer and Alex Kogan (PM, RU). For a given set of binary data, a *pattern* is a conjunction of literals which covers at least one positive point and covers no negative points. One measure of a pattern is the coverage, i.e., the number of the positive points it covers. Because the problem of finding a pattern that maximizes the coverage is NP-hard, one could not expect to find a polynomial time algorithm. Shi, Hammer, and Kogan (PM, RU) proposed a greedy algorithm for it. The idea is to try to find a literal by conjuncting it to the partial-pattern such that the resulting new partial-pattern will have the greatest number of positive points and simultaneously decrease the number of negative points by as many as possible. Then, one needs to make the pattern prime. Shi implemented this algorithm using C++ and determined that the pattern generated by it was always close to optimal.

Tonguc Unluyurt

Suppose we are given a boolean function. Often, this function is related to a system. We would like to figure out the value of the function at a particular state which is unknown to us. But we can learn the value of argument i by paying a cost of c_i. Also we know that the argument will be 1 with probability p_i. Unluyurt worked on the problem of finding a decision tree which will have a minimum expected cost. (It is assumed that the components fail or work independently of each other.) This problem has many applications where the boolean function describes some system. Unluyurt worked on the problem and its applications under the mentorship of Endre Boros (PM, RU). He gave a polynomial algorithm to find a minimum expected cost decision tree for a double regular boolean function. This result generalizes the results in the literature for k-out-of-n systems. Unluyurt also worked on diagnosing read-once systems which consist of identical components. He gave a polynomial algorithm to find a minimum expected cost decision tree for networks, assuming that each component is identical.

Paper

Tonguc Unluyurt. "Optimal diagnosis of double regular functions." In preparation.

Talk

Fifth International Conference on Artificial Intelligence and Mathematics, Fort Lauderdale, Jan. 1998. "Optimal diagnosis of double regular functions."

Yang Yu

For x, y elements of some (finite) poset P, write $p(x < y)$ for the probability that x precedes y in a random (uniform) linear extension of P. For $u, v \in [0,1]$ define $f(u,v) = \inf\{p(x < z) : p(x < y) \geq u,\ \ p(y < z) \geq v\}$, where the infimum is over all choices of P and distinct $x, y, z \in P$. Addressing an issue raised by Fishburn, Yang and Jeff Kahn (PM, RU) gave the first nontrivial lower bounds on the function f. This is part of a more general geometric result, the exact determination of the function $g(u,v) = \inf\{\Pr(X_1 < X_3) : \Pr(X_1 < X_2) \geq u,\ \ \Pr(X_2 < X_3) \geq v\}$, where the infimum is over $X = (X_1 \ldots X_n)$ chosen uniformly from some compact convex subset of a Euclidean space. These results are mainly based on the Brunn-Minkowski Theorem and a theorem of Keith Ball, which allowed for the reduction to a 2-dimensional version of the problem.

Papers

Yang Yu. "On proportional transitivity of ordered sets." Submitted to *Discrete & Computational Geometry.*

Yang Yu and Jeff Kahn. "Log-concave functions and poset probabilities." Submitted to *Combinatoria.B*

3.2.2 DIMACS-Supported Graduate Students at Princeton University, February 1997 - January 1998

Steven Kleinstein

When a foreign substance is introduced into our bodies, our immune system acts to neutralize that substance. This is a complex process involving the collective and co-ordinated response of approximately 10^{12} cells. In the effort to fit detailed experimental findings into a comprehensive model of the immune system, computer simulations are just beginning to play a role. Kleinstein together with J.P. Singh (PM, PU) and Philip Seiden, has been working, along with a group of scientists in the molecular biology department at Princeton, on a model of the immune system called IMMSIM. The IMMSIM model is actually not a single model, but a collection of models that cooperate with each other to produce a simulation of the immune system. Each of these models is called a module and they have developed a flexible software framework to connect these modules together. Currently, they simulate the humoral response only, although an effort is under way to add in the cellular response. As it now stands, IMMSIM contains a model of the bone marrow, thymus and a lymph node. While many models in theoretical immunology are differential equation-based, dealing with entire populations of cells and molecules, IMMSIM represents all of the cells and molecules in the system explicitly. This is much more expensive computationally, but there are also numerous benefits. For example, it is much easier to model and analyze the causality between events. Explicit representations are also more realistic biologically as there are genetic recombination and mutation events that operate in the immune system to produce an enormous number of different types of cells. Many other methods seek to group these diverse cells into a small number of equivalence classes based on just one or a few characteristics. Entities are the 'atomic objects' that populate the IMMSIM modules and constitute the flows between them. Interactions between and internal dynamics within entities combine to create the dynamics of the simulation. There are currently seven types of entities in the IMMSIM model: Antibodies, Antigens, Antigen Presenting Cells, B-Cells, Immune Complexes, Plasma Cells and T-Cells. Kleinstein's current research focuses on the biological phenomena of repertoire shift and affinity maturation. These are the processes by which an initial immune response, which is typically made up of low-affinity cells, comes to be dominated by high-affinity ones. This is accomplished through mutation and selection processes. Using IMMSIM, he and his collaborators have put upper and lower bounds on the magnitudes of these processes as well as obtained a feel for where the values of many other unknown parameters might lie. Their hope is that these 'in machina' experiments will lead to 'in vivo' and 'in vitro' experiments that will confirm or deny these findings.

Paper

S. Kleinstein, J. Pal Singh, Philip E. Seiden. "Putting limits on selection and mutation in a simulation of the humoral immune response." Submitted to the *Pacific Symposium on BioComputing.*

Anil Ruia

Ruia worked, under the direction of Sanjeev Arora (PM, PU), on approximation algorithms on NP-hard problems, especially geometric problems. He studied the polynomial time approximation scheme for TSP and other approximation problems as developed in recent work of Arora. He tried to improve the running time of the algorithm to make it useful for practical purposes. He also tried to adapt the algorithm or to use ideas from it to get a polynomial time approximation scheme for the Triangulation and Steiner Triangulation problems.

Konstantinos Tsioutsiouliklis

Tsioutsiouliklis worked on network flow problems under the guidance of Robert Tarjan (PM, PU). In particular, he considered 'all pairs minimum cuts in networks'. Based upon recent work of Chekuri et al, "Experimental Study of Minimum Cut Algorithms", as well as on previous work of Gomory-Hu and Gusfield who present different algorithms for the all pairs network flow problem, he has developed implementations of these algorithms. The goal of his work is an experimental study of the behavior of these algorithms on several network cases, as well as the development of several heuristics to improve their behavior.

3.3 Center Visitors: February 1997- January 1998

- **Andris Ambainis**; Institute of Mathematics and Computer Science, University of Latvia; Visit Related to: Molecular Biology; Dates: 4/2/97 - 4/5/97
- **Imre Barany**; Hungarian Academy of Science; Visit Related to: General; Dates: 9/29/97 - 10/13/97
- **Amir Ben-dor**; Technion; Visit Related to: Molecular Biology; Dates: 4/21/97 - 5/3/97
- **Volodia Blinovsky**; Inst. for Info. Transmission Problems, Russian Academy of Sciences; Visit Related to: General; Dates: 4/20/97 - 4/28/97
- **Graham Brightwell**; London School of Economics; Visit Related to: Discrete Probability; Dates: 9/2/97 - 9/24/97
- **Ondrej Cepek**; Charles University; Visit Related to: General; Dates: 1/27/97 - 2/6/97
- **Mary Cryan**; University of Warwick; Visit Related to: Molecular Biology; Dates: 3/31/97 - 5/15/97
- **Bhaskar DasGupta**; Rutgers,Camden-Department of Computer Science; Visit Related to: General; Dates: 10/1/96 - 6/30/97
- **Guoli Ding**; Louisiana State; Visit Related to: General; Dates: 9/5/97 - 12/1/97

- **Eytan Domany**; Weizmann Institute;
  Visit Related to: General; Dates: 5/4/97 - 5/18/97
- **Peter L. Erdős**; Math. Inst. of the Hungarian Academy of Science;
  Visit Related to: Molecular Biology; Dates: 1/20/97 - 1/24/97
- **Weiwu Fang**; Chinese Academy of Science;
  Visit Related to: General and Molecular Biology; Dates: 11/15/96 - 1/4/98
- **Stephan Foldes**; National Autonomous University of Mexico;
  Visit Related to: General; Dates: 11/3/97 - 12/20/97
- **Bruno Gaujal**; INRIA - Intitut National de Recherche Informatique et en Automatique;
  Visit Related to: Networks; Dates: 7/5/97 - 8/31/97
- **Vladimir Gurvich**; Technion;
  Visit Related to: General; Dates: 8/1/97 - 8/31/97
- **Leonid Gurvits**; NEC Research Institute;
  Visit Related to: General; Dates: 10/24/96 - 6/30/97
- **Olle Haggstrom**; Chalmers University;
  Visit Related to: Discrete Probability; Dates: 3/22/97 - 4/2/97 and 5/28/97 - 6/7/97
- **Armin Haken**; Czech Academy of Science;
  Visit Related to: General; Dates: 8/1/96 - 7/31/97
- **Fang Hao**; Georgia Tech;
  Visit Related to: General; Dates: 6/17/97 - 9/19/97
- **Micha Hofri**; Rice University;
  Visit Related to: General; Dates: 10/6/97 - 11/4/97
- **Daniel Huson**; University of Pennsylvania;
  Visit Related to: Molecular Biology; Dates: 9/1/97 - 8/31/98
- **Klaus Jansen**; University of Trier ;
  Visit Related to: Networks; Dates: 4/21/97 - 5/3/97
- **Mark Jerrum**; University of Edinburgh, Scotland;
  Visit Related to: Discrete Probability; Dates: 3/21/97 - 6/7/97
- **Gyula Katona**; Math Institute - Hungarian Academy of Sciences;
  Visit Related to: General; Dates: 7/14/97 - 7/25/97
- **Oleg Kedrov**; Inst. of Seismology, Joint Institute of Earth Physics, Russian Academy of Sciences;
  Visit Related to: Massive Data Sets; Dates: 3/6/97 - 4/6/97
- **Jeong Han Kim**; Microsoft;
  Visit Related to: Discrete Probability; Dates: 4/21/97 - 4/25/97
- **Eyal Kushilevitz**; Technion;
  Visit Related to: General; Dates: 1/26/97 - 2/14/97 and 10/5/97 - 11/5/97

- **Yanpei Liu**; Northern Jiaotong University, Beijing;
  Visit Related to: General; Dates: 6/15/97 - 6/28/97
- **Nathan Lulov**; Harvard University;
  Visit Related to: Discrete Probability; Dates: 3/24/97 - 5/16/97
- **Russell Lyons**; Indiana University (Bloomington);
  Visit Related to: Discrete Probability; Dates: 5/30/97 - 6/7/97
- **Angus MacIntyre**; Oxford;
  Visit Related to: General; Dates: 5/4/97 - 6/1/97
- **Eddy Mayoraz**; Dalle Molle Inst. of Perceptive Artificial Intelligence;
  Visit Related to: General; Dates: 6/21/97 - 7/21/97
- **Boris Mirkin**; Rutgers University;
  Visit Related to: General; Dates: 2/1/97 - 1/31/98
- **Vadim Mottl'** ; Tula State University;
  Visit Related to: Massive Data Sets; Dates: 3/6/97 - 4/6/97
- **Ilya Muchnik**; Rutgers University;
  Visit Related to: Massive Data Sets; Dates 2/1/97 - 1/31/98
- **Gianpaolo Oriolo**; University of Rome;
  Visit Related to: General; Dates: 5/12/97 - 7/12/97
- **Yuval Peres**; Math Institute, Hebrew University;
  Visit Related to: Discrete Probability; Dates: 3/22/97 - 3/29/97
- **Boris Pittel**; Ohio State University;
  Visit Related to: Discrete Probability; Dates: 3/21/97 - 6/7/97
- **Neil Robertson**; Ohio State University;
  Visit Related to: General; Dates: 7/7/97 - 9/10/97
- **Christian Rodriguez**; University of Buenos Aires;
  Visit Related to: Molecular Biology; Dates: 12/19/97 - 2/14/98
- **Jose Rolim**; University of Geneva;
  Visit Related to: Networks; Dates: 4/28/97 - 5/9/97
- **Adi Rosen**; University of Toronto;
  Visit Related to: General; Dates: 9/29/97 - 10/29/97
- **Sunil Shende**; University of Nebraska;
  Visit Related to: General; Dates: 9/23/97 - 5/30/98
- **Amin Shokhrollahi**; University of California, Berkeley;
  Visit Related to: General; Dates: 9/20/97 - 9/26/97
- **Bruno Simeone**; La Sapienza University;
  Visit Related to: General; Dates: 4/1/97 - 4/15/97
- **Ron Solomon**; Ohio State;
  Visit Related to: General; Dates: 3/3/97 - 3/30/97

- **Joel Spencer**; New York University, Courant Institute;
  Visit Related to: Discrete Probability; Dates: 1/11/97 - 4/11/97
- **Aravind Srinivasan**; National University of Singapore;
  Visit Related to: Networks; Dates: 11/24/97 - 12/13/97
- **Jeff Steif**; Chalmers University of Technology;
  Visit Related to: Discrete Probability; Dates: 1/2/97 - 1/15/97 and 3/28/97 - 4/16/97
- **Gernot Stroth**; Halle University;
  Visit Related to: General; Dates: 3/2/97 - 3/9/97
- **Tamas Szonyi**; Attila Jozsef University;
  Visit Related to: General; Dates: 4/18/97 - 5/3/97
- **Andras Telcs**; International Business School, Budapest;
  Visit Related to: Discrete Probability; Dates: 3/20/97 - 4/3/97
- **Prasad Tetali**; Georgia Institute of Technology;
  Visit Related to: General; Dates: 2/1/97 - 3/3/97
- **Robin Thomas**; Georgia Tech;
  Visit Related to: General; Dates: 7/7/97 - 9/10/97
- **Jacob van den Berg**; CWI;
  Visit Related to: Discrete Probability; Dates: 1/10/97 - 7/1/97
- **Boban Velickovic**; University of Paris;
  Visit Related to: General; Dates: 9/25/97 - 10/25/97
- **Nikolai Verechtchaguine**; Maimonid State Academy;
  Visit Related to: General; Dates: 7/23/97 - 8/22/97
- **Eric Vigoda**; University of California, Berkeley;
  Visit Related to: Discrete Probability; Dates: 5/20/97 - 8/1/97
- **Van Ha Vu**; Yale University;
  Visit Related to: Discrete Probability; Dates: 5/30/97 - 6/14/97
- **Dominic Welsh**; Oxford University;
  Visit Related to: Discrete Probability; Dates: 3/16/97 - 3/28/97
- **Jia-Feng Weng**; Melbourne University;
  Visit Related to: Networks; Dates: 4/27/97 - 5/11/97
- **David Wilson**; University of California, Berkeley;
  Visit Related to: Discrete Probability; Dates: 1/10/97 - 6/10/97
- **Tomoyuki Yamakami**; Princeton University;
  Visit Related to: General; Dates: 9/1/97 - 8/31/98
- **Michele Zito**; University of Liverpool;
  Visit Related to: General; Dates: 11/3/97 - 12/2/97

4 Special Year, Focus and Workshop Programs

4.1 Special Year on Networks, 1996-00

Overview:

The Special Year on Networks began in July 1996 and will continue through August 2000 with the help of a three year grant received from the National Security Agency. The program is stressing postdoctoral training and workshops that will reach a broad segment of the DM and TCS research community as well as engage practitioners.

Organizers:

- Stuart Haber, Bellcore / Surety Technologies
- David Johnson, AT&T Labs - Research
- Mihalis Yannakakis, Bell Labs

Scientific Advisory Board: Baruch Awerbuch, Johns Hopkins; Br. Badrinath, Rutgers; Endre Boros, Rutgers; Ding-Zhu Du, Minnesota; David Goodman, Rutgers; Jack Holtzman, Rutgers; Frank Hsu, Fordham; Srinivasan Keshav, Cornell; Arjen Lenstra, Citibank; Udi Manber, Arizona; Ray Miller, Maryland; Debasis Mitra, Bell Labs; Clyde Monma, Bellcore; Panos Pardalos, Florida; K.K. Ramakrishnan, AT&T Labs - Research; Greg Shannon, Spanning Tree Technologies; Scott Shenker, Xerox; Subhash Suri, Washington U. in St. Louis; Eva Tardos, Cornell; Peter Winkler, Bell Labs; and Yacov Yacobi, Microsoft

Sponsors:

The National Science Foundation
The New Jersey Commission on Science and Technology
The National Security Agency
AT&T Labs - Research
Bell Labs
Bellcore
Rutgers University
Princeton University

Theme of program:

This special year is intended to bring together researchers and practitioners in a wide variety of fields connected to the common theme of "Networks" and ranging from questions of infrastructure to ones of applications. Special emphasis is placed on Internet-related topics, as this is a young and growing area where theoretical computer science and discrete mathematics have a chance to make a real impact.

The special year's topics can be grouped into four overlapping areas, which we list below along with examples of planned or possible workshop topics.

- **Network Security:** protocols for electronic commerce, public key infrastructure, network security and threat models.
- **Network Applications:** information retrieval, databases and the Web, multimedia transmission, and coding distributed and collaborative computing.

- **Network Control:** routing, congestion and admission control, network protocols, mobile computing.
- **Network Design:** topology and connectivity, switch design, facilities location.

The goal is to engage theoreticians with these problems and foster communication between them and real-world practitioners. Practitioners have much to tell theorists about what the key issues are, and theorists may themselves be able to provide crucial insights and solutions to practitioners.

Workshops, Miniworkshops and Special Events 1997-98:

- **Workshop on Architecture and Algorithmic Aspects of Communication Networks**
  - Dates: January 26-28, 1997
  - Location: Rutgers University
  - Attendance: 70
  - Organizers: Eva Tardos, Cornell University; Serge Plotkin, Stanford University; Israel Cidon, Technion and Sun Microsystems

  The size of the future networks, combined with the need for bandwidth reservation and QoS guarantees, poses numerous challenging problems. These problems are exacerbated by the convergence of the data, telephone and video services that have to be supported in the same framework. In parallel, the recent advances in theoretical computer science and discrete mathematics bring us closer to being able to address many of these issues. Unfortunately, these advances have not had a significant impact on practical networks, mostly because of a lack of continuous dialog between the research communities.

  The aim of the workshop was to bring together and allow the exchange of research ideas and interests between experts in networking, theory, ATM and Iternet. Both theoretical and applied issues were considered. Primary areas of focus of the workshop included:

  - Switching: Packet scheduling and QoS guarantees, switching architecture.
  - Routing: Link state algorithms, dynamic online routing of virtual circuits, generation and maintenance of routing tables, rerouting due to bandwidth fragmentation, service restoration after link failure.
  - Network topology design.
  - Multicast: Membership management, multicast routing, QoS in multicast, policing.
  - Admission Control.
  - Fairness.

  Most of the workshop was devoted to talks by distinguished invited speakers but there were also a few contributed talks and an hour long panel discussion.

- **Workshop on Economics, Game Theory, and the Internet**
  - Dates: April 18-19 1997
  - Location: Rutgers University
  - Attendance: 80
  - Organizers: Eric Friedman, Rutgers; Scott Shenker, Xerox PARC

The Internet poses many interesting challenges for theoretical economics and game theory. The highly asynchronous interactions and large degree of anonymity inherent in the Internet require that many traditional game theoretic notions of equilibrium, learning, and mechanism design need to be re-examined in this context. New cost sharing issues (such as when receivers share multicast transmissions, or when flows share the use of a link) arise in network settings. Incentive issues must be considered when designing basic network algorithms and protocols. In addition, regulation of the Internet and other telecommunication networks raises other theoretical issues. This workshop focused on identifying new areas of theoretical interest that arise out of the institutional and technical nature of the Internet and other communication networks.

Each day of the two day workshop began with two 90 minute plenary talks followed by more standard 30 minute research reports in the afternoons. Plenary speakers included Herve Moulin (Duke University) who discussed the applications of fair division, which has been much studied in economics, to problems in the Internet; Bill Sharkey (Federal Communications Commission) who described the Telecommunications Act of 1996 and its implications; Hal Varian (University of California - Berkeley) on intellectual property rights in a digital environment; and Scott Shenker (Xerox PARC) who gave a tutorial on the structure of the Internet and a review of important open problems.

The speakers were quite diverse, coming from Computer Science, Economics, Engineering, and Physics; they covered topics such as economic analyses of the Internet, game theoretic learning on the Internet, science policy related to the Internet, and general analyses of networks.

- **Workshop on Network Design: Connectivity and Facilities Location**
  - Dates: April 28-30, 1997
  - Location: Nassau Inn, Princeton, NJ
  - Attendance: 70
  - Organizers: Ding-Zhu Du, University of Minnesota; Panos Pardalos, University of Florida

  Connectivity and facilities location are two important topics in network design with applications in data communication, transportation, production planning, and VLSI designs. There are two issues concerning these two topics: design and optimization. They involve combinatorial design and combinatorial optimization. Finding the solution of design problems and the optimal or approximate solution of the related optimization problem are challenging tasks because no polynomial time algorithms are known. Such problems include some variations of Steiner tree problems (such as multiple connected Steiner network, independent flow problem, and subset-interconnection designs), topology network design, nonlinear assignment problems (such as quadratic assignment problems), problems in facilities location and allocation and network problems appearing in VLSI design. The workshop focused on combinatorial, algorithmic, and applicational aspects of these problems.

  A refereed volume with papers presented at the Workshop will be published early next year by the American Mathematical Society in the DIMACS series.

- **Workshop on Network Switching**
  - Dates: July 7-9, 1997
  - Location: Princeton University
  - Attendance: 17
  - Organizers: Frank Hwang, AT&T; Ding-Zhu Du, University of Minnesota

Recent development in Internet has required new technology in network switching. This includes routing algorithm designs and network designs to meet various communication requirements. This workshop focused on all aspects of network switching, including the following topics:

- Network Environment: circuit-switching, packet switching, ATM switching, multirate switching, wireless.
- Routing: point-to-point, permutation, broadcast, conferencing.
- Network Topology: multistage, rings, cubes, grids, stars, general connectors, fault-tolerant networks.
- Switching Component: concentrator, superconcentrator, partial concentrator, expander, multiplier, copy network, n-generalizers, optical network.
- Nonblockingness: strictly nonblocking, wide-sense nonblocking, rearrangeable.
- Optimization: hardware minimization, optimal network design.

A refereed volume with papers presented at the Workshop will be published the next year by the American Mathematical Society in the DIMACS series and selected papers will also be published in a special issue of *Networks*.

- **Design and Formal Verification of Security Protocols**
  - Dates: September 3-5, 1997
  - Location: Rutgers University
  - Attendance: 110
  - Organizers: Hilarie Orman, DARPA; Catherine Meadows, Naval Research Laboratory

Research has been intensifying in the application of formal methods to cryptographic protocol verification. The goal of this workshop was to facilitate this process by bringing together those who are involved in the design and standardization of cryptographic protocols, and those who are developing and using formal methods techniques for the verification of such protocols.

The workshop was organized as a series of related talks by individuals and panels of experts, scheduled with ample time for questions. The goal was to encourage audience participation in order to foster understanding of the technology and how it could be applied, and this supported the overall goal of getting researchers and protocol designers to exchange ideas on how to make substantial progress in this field. We alternated papers with panels soliciting new paths for research with emphasis on proposals addressing new protocols with respect to their formal and informal analysis.

Other topics of interest included:

- Progress in belief logics
- Use of theorem provers and model checkers in verifying crypto protocols
- Interaction between protocols and cryptographic modes of operation
- Methods for unifying documentation and formal, verifiable specification
- Methods for incorporating formal methods into crypto protocol design
- Verification of cryptographic API systems
- Formal definition of correctness of a cryptographic protocol

- Arithmetic capability required for proofs of security for number theoretic systems
- Formal definitions of cryptographic protocol requirements
- Design methodologies
- Emerging needs and new uses for cryptographic protocols
- Multiparty protocols, in particular design and verification methods

The workshop papers were published on-line the day before the workshop started, and a CD-ROM with the papers was produced as well.

- **End-to-End Network Modelling and Simulations**
  - Dates: October 23-24, 1997
  - Locations: Rutgers University
  - Attendance: 100
  - Organizers: Sandeep Bhatt, Bellcore; Andrew Ogielski, Rutgers; Ashok Erramilli, QMetrix

  This workshop focused on current developments in modeling, simulation, and algorithmic techniques which take into account the observed and measured behavior of networks. In order to obtain a reasonably complete picture of the state of the art, and to stimulate further research, the organizers invited practitioners of industrial and defense network modeling together with researchers involved in the analysis and simulations of networks and protocols. They were particularly interested in identifying areas requiring further research, and in creating an informal forum for practitioners, researchers, and students to hear each other and to exchange ideas.

  The primary topics covered in the workshop were:

  - Network modeling and simulation practices
  - Modeling and simulation of wireless networks
  - Models of traffic sources and aggregated traffic
  - Modeling and simulation of network protocols
  - Parallel simulation techniques for large telecommunication systems
  - Algorithmic analyses of network and protocol models

  The demographic mix of the participants (20% were from corporations unaffiliated with DIMACS, 40% were from corporations affiliated with DIMACS (AT&T, Bellcore, Lucent) and 40% were from academia and national laboratories) indicates that we have succeeded in creating an atmosphere conducive to university-industry collaborations, and in reaching out to the networking and communications industries.

  All workshop viewgraphs will be published on the DIMACS web pages together with extra papers and links provided by the speakers.

- **Networks in Distributed Computing**
  - Dates: October 27-29, 1997
  - Location: Rutgers University
  - Attendance: 50
  - Organizers: Marios Mavronicolas, University of Cyprus; Michael Merritt, AT&T Labs - Research; Nir Shavit, Tel Aviv University and MIT

The primary goal of the Workshop was to bring together researchers from both industry and academia interested in issues related to networks as they appear in the context of distributed computing. We invested special effort to expose the theoretical community carrying out basic research in the field of distributed computing to fundamental problems and challenges related to recent technological advances in the networking industry. We invited both theoreticians and practitioners, in a hope that the latter would carry through their practical motivation, experience, and insights to the Workshop

Topics of discussion included: ATM networking technology, issues of routing and flow control in modern communication networks, service pricing, security, optical networking, management of Internet, mobile computing, and formal verification of communication protocols.

Events Planned for the Rest of 1997-98

- **Workshop on Ethical and Social Dimensions of Information Technology**
  - Dates: February 20-21, 1998
  - Location: Princeton University
  - Organizers: Helen Nissenbaum and Bernard Chazelle, Princeton University; Joan Feigenbaum, AT&T Labs - Research
- **Workshop on Multichannel Optical Networks: Theory and Practice**
  - Dates: March 15-17, 1998
  - Location: Rutgers University
  - Organizers: Peng-Jun Wan, Illinois Institute of Technology; Ding-Zhu Du, University of Minnesota; Panos Pardalos, University of Florida

4.2 Special Focus on Discrete Probability, 1996-98

Overview:

This program began during the 1996 - 1997 academic year and continued into 1997 - 1998 at a much lower level. It is jointly organized and sponsored with the Institute for Advanced Study, which is cohosting visitors and workshops with DIMACS.

Organizers:

- Enrico Bombieri, Institute for Advanced Study
- Jeff Kahn, Rutgers University
- Peter Winkler, Bell Labs

Sponsors:

The National Science Foundation
The National Security Agency
AT&T Labs - Research
Bell Labs
Bellcore
The Institute for Advanced Study
Rutgers University

Princeton University

Theme of program:

Probabilistic methods and ideas have by now become central to large parts of discrete mathematics and theoretical computer science. At the same time, recent years have seen a number of major developments in parts of probability and statistical mechanics dealing with topics of an essentially discrete nature (e.g., percolation, random walks and Markov chains, particle systems, branching processes).

A major goal of this program is to foster interaction between the DM/TCS community and the probability and statistical community. This is already happening to some extent, with both the application of increasingly sophisticated tools from probability to DM and TCS, and the solution of problems from probability by combinatorial means. Nonetheless, such interactions, and even awareness of what these communities have to offer each other, are still far more exception than rule. So the program is intended to give a substantial push to a broader and more rapid development of these connections. The program is also reaching out from the mathematical sciences to the physical sciences, especially Statistical Mechanics.

Workshops 1997 - 1998:

- **Workshop on Statistical Physics Methods in Discrete Probability, Combinatorics and Theoretical Computer Science**
  - Dates: March 23-27, 1997
  - Location: Institute for Advanced Study and Rutgers University
  - Attendance: 90
  - Organizers: Jennifer T. Chayes, UCLA and Dana Randall, Georgia Tech

  This workshop focused on new developments and directions at the interface between statistical physics and the fields of discrete probability, combinatorics, information theory and theoretical computer science. The workshop brought together 90 researchers in various fields with the intention of developing a common language and recognition parallels among the fundamental problems they are addressing. Speakers focused on topics which included: Critical percolation - finite-size scaling and conformal invariance; Potts models, random cluster models and the Tutte polynomial; dynamics of the giant component in the random graph model; hard core models and independent sets; noisy channels and network reliability; Markov sampling methods; and phase transitions in random structures and algorithms.

  This workshop received additional funding from the National Science Foundation and the National Security Agency.

- **Workshop on Probabilistic Analysis of Algorithms**
  - Dates: May 11-14, 1997
  - Location: Princeton University
  - Attendance: 95
  - Organizers: Alan Frieze, CMU and Michael Molloy, Toronto

  This workshop addressed problems arising in the mathematical analysis of algorithms with emphasis on probabilistic techniques. The theory of algorithms has undergone an extraordinarily vigorous development over the last 20 years or so, and probability theory has emerged as one of its most vital partners. By its nature probabilistic analysis cuts across the boundaries of Computer Science, Discrete Mathematics, Operations Research and Probability Theory.

This workshop focused on the analysis of algorithms in the average case, assuming a probabilistic distribution on input instances. The goal is to obtain good estimates of various performance measures of algorithms such as solution quality and running time.

The main objective was to explain why many simple algorithms, often used in practice, perform much better than as predicted by worst-case analysis. Pathological cases are generally rare and simple algorithms can often be shown to perform well in the average case. This is especially true in the case of NP-complete problems where it seems that the probabilistic approach may be the best way and indeed may be the only way to understand the success of heuristic combinatorial algorithms.

The workshop focused on five major sub-areas: Algorithmic Theory of Random Graphs and Related Combinatorial Problems; Analysis of Euclidean Functionals; Linear Programming and Integer Programming Problems; Analysis of Fundamental Algorithms; and Average Case Complexity.

- **Workshop on Microsurveys in Discrete Probability**
  - Dates: June 2-6, 1997
  - Location: Institute for Advanced Study, Princeton, New Jersey
  - Attendance: 76
  - Organizers: David Aldous, Berkeley, and Jim Propp, MIT

The goal of this workshop was to focus on "microsurveys" of recently active areas within discrete probability. The five day program included three sessions for open problems and scheduled topics which included: A simple family of tree-valued Markov processes; connections between hyperplane arrangements and card-shuffling; exact sampling from Markov chains; random walks in scenery: every vertex has a color; the emergence of the giant component in the random graph process; the scale-invariant Poisson process; mixing times for Markov chains; and quadratic dynamical systems.

Workshops Planned for the Rest of 1997-98

- **Workshop on Randomization Methods in Algorithm Design**
  - Dates: December 12 - 14, 1997
  - Location: Princeton University
  - Organizers: Panos Pardalos, University of Florida; Sanguthevar Rajasekaran, University of Florida; Jose Rolim, University of Geneva

4.3 Special Year on Mathematical Support for Molecular Biology, 1994-98

Overview: This program began in August 1994 as a special year intended to last 25 months. With significant additional external funding of some $525K plus additional funds from Rutgers, the center was able to plan an extensive series of tutorials and workshops, to support 4 postdoctoral fellows for two year appointments, host several long term visitors, and host a distinguished lecturer series.

The program is now extending into a fourth year of activity, aided by an additional grant of $100K from the National Center for Human Genome Research. The current year has two new postdoctoral fellows in a joint postdoc program with the University of Pennsylvania, several workshops, and an active seminar series. The first three years built up several collaborative relationships with other centers and departments in the biological sciences, and this year is building on those relationships.

Organizers:

- Fred Roberts (Chair), Rutgers
- Joachim Messing (Chair), Waksman Institute - Rutgers
- Lawrence Shepp (Co-Chair), Rutgers
- Michael Waterman (Co-Chair), University of Southern California
- Martin Farach (Assistant Chair), Rutgers

Steering Committee: David Axelrod, Rutgers; Michael Bulmer, Rutgers; Charles Cantor, Boston University; Daniel Gusfield, UC - Davis; Leroy Hood, Center for Molecular Biotechnology, Washington; Richard Karp, Washington; Eric Lander, MIT - Whitehead Institute; Eugene Lawler, UC - Berkeley; Joshua Lederberg, Rockefeller; Joseph Naus, Rutgers: Michiel Noordewier, Rutgers: Wilma Olson, Rutgers: David B. Searls, University of Pennsylvania; William Sofer, Rutgers; Tandy Warnow, University of Pennsylvania; Peter Winkler, Bell Labs; and Andrew Yao, Princeton.

Collaborating Centers and Institutes:

- Ecology & Evolutionary Biology Department, Princeton University
- Computer Science Department and Genome Center, University of Pennsylvania
- Molecular Biology Department, Princeton University
- Center for Molecular Biophysics and Biophysical Chemistry, Rutgers University
- Center for Molecular Biotechnology, University of Washington
- Program in Mathematics and Molecular Biology, UC - Berkeley
- Center for Theoretical and Applied Genetics, Rutgers University
- Mt. Sinai School of Medicine
- Waksman Institute for Molecular Genetics, Rutgers University

Sponsors:

AT&T Labs - Research
Bell Labs
Bellcore
BIOSYM Tech
Centers for Disease Control
Ciba-Geigy Corp.
Department of Energy
International Union of Crystallography
Merck Research

National Center for Human Genome Research
National Institute of Allergy and Infectious Diseases
National Science Foundation
National Security Agency
New Jersey Commission on Science and Technology
Princeton University
Program in Mathematics and Molecular Biology, UC-Berkeley
Promega Corporation
Roche Molecular Systems
Rutgers University
Sandia National Laboratories
The Sloan Foundation
Smith-Kline Beecham
University of Pennsylvania

Theme of program:

Discrete structures have been used to formulate some of the most fundamental concepts of molecular biology. It is now well-known that information storage within a cell is by means of long nucleic molecules which can be thought of as long strings of smaller units called nucleotides. Some of the most important problems of modern molecular biology involve combinatorial and algorithmic questions about these strings. Models of proteins, DNA, and RNA molecules as sequences form a basic tool of molecular biology.

Throughout its history, molecular biology has interacted with methods of discrete mathematics and theoretical computer science. The first nucleic acid sequence was determined in 1965 by R.W. Holley and his co-workers at Cornell University (it contained 77 bases). The method used was the fragmentation stratagem, which involved the combinatorial problem of reconstructing an unknown sequence from fragments obtained by certain enzyme decompositions. Graph-theoretical methods, using eulerian chains, were developed by Hutchinson to give algorithms for implementing the fragmentation stratagem. Of course, these methods are not used any more, and indeed were used only for a short time before other, more efficient methods were adopted. However, they played an important role in the development of molecular biology. In this special year, we are investigating similar uses of DM/TCS in molecular biology.

The special year has a number of specific goals. These are to expose a large community of discrete mathematicians/theoretical computer scientists to problems of molecular biology that are fundamentally DM/CS problems; provide an opportunity for intensive collaboration to discrete mathematicians/theoretical computer scientists who have already discovered these problems; introduce some outstanding young people to the field in a very concentrated way; forge lasting partnerships between mathematical scientists and biological scientists; involve biological scientists in a primary role in defining the agenda of computational biology; and introduce molecular biologists to a community of mathematical scientists who are interested in helping them with their problems.

Workshops in 1997 - 1998:

- **Workshop on DNA Topology II**
  - Dates: April 3-4, 1997
  - Location: Rutgers University
  - Attendance: 35

- Organizers: Wilma Olson, Rutgers; Bernard Coleman, Rutgers; Tamar Schlick, New York University; Irwin Tobias, Rutgers

The workshop attracted a diverse audience of biologists-biochemists and mathematicians-biophysicists by presenting a 50/50 mix of biologically and mathematically oriented topics. Sponsorship was provided by DIMACS; The Program on Mathematics and Molecular Biology (PMMB) at University of California, Berkeley; and The Center for Molecular Biophysics and Biophysical Chemistry (MBBC) at Rutgers University.

The focus of the workshop was on the energetics and dynamics of supercoiled and knotted DNA, and the effects of protein activity binding and primary base sequence on DNA structure and topology. Topics of discussion included the following: use of elastic rod theory to model the structure and dynamics of DNA, computer simulations, statistical mechanics of DNA topology (i.e., numerical methods for visualizing and manipulating DNA topology), effects and/or analysis of drug, carcinogen, and protein-binding activity on DNA structure and topology, DNA biophysical properties (i.e. stretching, catenation).

A set of papers presented by the participants is currently under review for publication in "Biophysical Journal".

- **Joint DIMACS/Universidad de Buenos Aires Tutorial on Computational Biology**
  - Dates: July 21-26, 1997
  - Location: University of Buenos Aires
  - Attendance: 73
  - Organizers: Dannie Durand, University of Pennsylvania; Martin Farach, Rutgers University; R. Ravi, Carnegie Mellon University; Mona Singh, Princeton University

Computational Molecular Biology is the field of computer science that has emerged to solve algorithmic problems in determining sequences and analyzing them. Specific research efforts in this area include sequencing and mapping, pairwise and multiple sequence comparison, protein structure determination and evolutionary tree reconstruction. This tutorial was developed to give a basic introduction to the major algorithmic research areas in computational biology. It consisted of 18 hours of instruction presented as six three-hour segments.

Of the 73 participants in the course, 50% were biologists, 40% were computer scientists and 10% were from other scientific disciplines. The primary goal of biology students who attended the course was to better understand the methods on which the sequence analysis tools that they use in their research are based. Computer science students attended the course primarily to better understand algorithmic methods used in computational biology.

As a result of this presentation in Argentina, two computer science students will visit DIMACS in January of 1998 to collaborate with Martin Farach. Ten of the best students applied for these two visiting positions.

- **mmCIF Software Developers' Workshop**
  - Dates: October 17-19, 1997
  - Location: Rutgers University
  - Attendance: 50
  - Organizers: Helen Berman, Rutgers University; Paula Fitzgerald, Merck Laboratories; Brian McMahon, IUCr; John Westbrook, Rutgers University

CIF is the Crystallographic Information File format. It was designed as a dictionary of data items sufficient for archiving the small molecule crystallographic experiment and its results. mmCIF, the macromolecular Crystallographic Information File format, is an extension of CIF to support the macromolecular crystallographic experiment, to improve the publication process for experimental papers, and to facilitate the development of tools so that computer programs could easily interface with the CIF.

With a mix of tutorial, hands-on and discussion sessions, the workshop addressed such topics as: adapting existing applications to mmCIF, creating new applications to exploit the capabilities of DDL2 and mmCIF, use of existing mmCIF software tools and packages, and detailed presentations on the content of the DDL2 and mmCIF dictionaries. Discussion was focused on the requirements for new mmCIF software and dictionary content, and on the coordination of these ongoing development activities.

This workshop was jointly sponsored by the National Science Foundation, the Rutgers University Bioinformatics Initiative, DIMACS, and the International Union of Crystallography.

Events Planned for the Rest of 1997-98

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

Events Under Development

- **Workshop on Biological Databases, tentative Spring 1998**
- **Miniworkshop on Reconstructing Very Large Trees in Biology, tentative May 1998**
- **Workshop on Structural Informatics**

4.4 Special Focus on DNA Computing, 1997-98

Overview:

The Special Focus on DNA Computing began in February 1997 and is continuing through 1997-98. It builds on two earlier DIMACS workshops in the area and on collaborations between computer scientists and biologists stimulated by the Special Year on Mathematical Support for Molecular Biology.

Organizers:

- Richard Lipton, Princeton University
- Laura Landweber, Princeton University

Sponsors:

The National Science Foundation
The New Jersey Commission on Science and Technology
AT&T Labs - Research
Bell Labs

Bellcore
Rutgers University
Princeton University

Theme of Program:

DNA computing presents great challenges both to computer sciences and molecular biology in such areas as using parallelism, designing new algorithms, and dealing with errors in molecular processes. The Focus on DNA Computing attempts to introduce researchers in discrete mathematics and theoretical computer science to parts of molecular biology relevant to computation. Molecular Biology is an area that is increasingly dependent on DM and TCS competencies and a field that has elements of being fundamentally an "information science," and so it is natural to explore connections between it and computing. DNA Computing has attracted interest from researchers from areas such as molecular biology, genetics, supercomputing, and parallel computing, to collaborate with scientists working in DIMACS core areas, on new models of computation.

Workshops, Miniworkshops and Special Events, 1997-98

- **Third DIMACS Workshop on DNA Based Computers**
  - Dates: June 23-25, 1997
  - Location: University of Pennsylvania
  - Attendance: 123
  - Organizers: Harvey Rubin, University of Pennsylvania (chair); Len Adleman, University of Southern California; Carter Bancroft, Mt. Sinai; Anne Condon, Wisconsin; Richard Lipton, Princeton; David Wood, Delaware

  In order to set the stage for the multidisciplinary nature of the conference, two tutorial sessions were held. David Wood reviewed some basic concepts in computation for the biologists and biochemists and Laura Landweber reviewed ideas and concepts in molecular biology for the computer scientists and mathematicians. There were sessions entitled "Making it Work in the Lab," "Proposed Applications," "Proposed Algorithmic Approaches," and "Models of DNA Computers." As for "making it work in the lab," some researchers stressed the need for a standard set of benchmarks for key primitive operations. Under "proposed applications," an invited paper suggested using DNA computation to operate on unknown pieces of DNA. This would have applications in DNA sequencing, DNA fingerprinting, and detection of mutations in populations. Among the proposed algorithmic approaches discussed were connections between protein folding problems and NP complete problems and grappling with the ability of DNA to implement nondeterminism as a native mode of computation. As for models of DNA computers, there was some argument for the development of local parallelism in biomolecular computation as opposed to the more generally understood distributed parallelism. Throughout the workshop, every opportunity was taken to encourage the interaction between the theoreticians and the lab people.

  This workshop was sponsored by the National Science Foundation, NEC Research Institute, Promega Corporation, The Sloan Foundation, The University of Pennsylvania, and DIMACS.

Events Planned for the Rest of 1997-98

- **Workshop on Nucleic Acid Selection and Computing**
  - Dates: March 15-17, 1998

  - Location: Princeton University
  - Organizers: Laura Landweber and Richard Lipton, Princeton University; Robert Dorit, Yale University; Andrew Ellington, Indiana University

Events Under Development

- **Fourth Annual DIMACS Workshop on DNA Based Computers, tentative June 1998**

4.5 Special Year in Massive Data Sets, 1997-98

Overview:
The DIMACS special year on massive data sets began with prespecial year activities in 1996-97 and is continuing in 1997-98 as a full-fledged special year.

Organizers:

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

Sponsors:
The National Science Foundation
The New Jersey Commission on Science and Technology
AT&T Labs - Research
Bell Labs
Bellcore
Rutgers University
Princeton University

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

Workshops, Miniworkshops and Special Events, 1997-98

- **Session on Mathematical, Statistical, and Algorithmic Problems of Very Large Data Sets**
  - Dates: January 8, 1997
  - Location: American Mathematical Society Meeting, San Diego, CA
  - Organizers: Joan Feigenbaum, AT&T Labs - Research; Fred Roberts, Rutgers University

  In an effort to acquaint the mathematical sciences community with the problems and challenges in this area, the Division of Mathematical Sciences and DIMACS organized a session on massive data sets at the San Diego meeting of the American Mathematical Society. This session described some of the challenging mathematical, statistical, and algorithmic problems inherent in organizing and using enormous amounts of data. The emphasis was on basic issues that transcend particular applications. Speakers explained why existing mathematical, statistical, and algorithmic methods break down on the enormous data sets that scientists and technologists now encounter regularly and attempted to delineate the boundaries at which these breakdowns occur. The session included a discussion of specific DIMACS and DMS programs that focus on huge data sets.

  The speakers were: Laney Kulsrud, Institute for Defense Analyses, Center for Communications Research, Princeton; David Karger, Massachusetts Institute of Technology; Sarbani Basu, Aarhus University; and Ken Church, AT&T Labs - Research.

- **Miniworkshop on Exploring Large Data Sets Using Classification, Consensus, and Pattern Recognition Techniques**
  - Dates: May 29-30, 1997
  - Location: Rutgers University
  - Attendance: 58
  - Organizers: F. R. (Buck) McMorris, University of Louisville, and Ilya Muchnik, Rutgers University

  This two day workshop featured the presentation of 21 papers from three separate research areas: academic research in methods, academic research in applications, and industrial research.

  The papers on methods covered the latest results in DM/TCS approaches (convex Boolean functions, hierarchical bases, support vectors, enhanced search, etc.) and the applied papers related to challenging high-tech issues (QSAR, computational finance, very large networks, etc.).

  From the discussions at the workshop it appears that there is not much experience in systematic C2P-supported handling of massive data sets. The large data sets currently maintained by even an advanced company such as AT&T amount to a dozen Gb with just elementary statistical tools implemented at present. The C2P theoretical support under development appears oriented mostly to thousands, not millions of entities. To achieve linear or sub-linear complexity of the algorithms, the theories and methods seem to call for a two-level organization. On the first level, clusters of relatively small samples should be found and analyzed. Next the cluster characteristics should then be extended onto the entire data set.

  Additional topics discussed concerned the preprocessing stage of the data set and analyses such as recognition of false data entries, filling in missed data entries, choice of appropriate discretization grid, and consistency among different data sources.

- **Workshop on Massive Data Sets in Telecommunications**
  - Date: October 13-15, 1997
  - Location: Rutgers University
  - Attendance: 70
  - Organizer: Joan Feigenbaum, AT&T Labs - Research

  Due to the enormous amounts of data involved in telecommunications systems, the telecommunications industry poses some of the most challenging massive data sets problems for researchers in the mathematical sciences. Many classical algorithmic and mathematical problems are relevant, including clustering, compression, error-detection and -correction, nearest-neighbor searching, and shortest-path and connectivity computations.

  This workshop attracted many participants from Bell Labs, AT&T Labs and Bellcore as well as a contingent from academia. Dealing specifically with the problems of the telecommunications industry, it was focused towards such industry specific problems as network operations, management, billing, fraud detection, quality control, marketing, information retrieval, and law enforcement.

Events Planned for the Rest of 1997-98

- **Workshop on Mathematical Methods for Massive Data**
  - Dates: April 27-28, 1998
  - Location: Princeton University
  - Organizers: Helene Kulsrud, Institute for Defense Analyses; Robert Grossman, University of Illinois
- **Workshop on Astrophysics and Algorithms: Massive Astronomical Data Sets**
  - Dates: May 6-8, 1998
  - Location: Princeton University
  - Organizers: Michael Vogeley, Princeton University; David Spergel, Princeton University; Don Sun, Bell Labs; Tony Tyson, Bell Labs
- **Workshop on External Memory Algorithms and Visualization**
  - Dates: May 20-22, 1998
  - Location: Rutgers University
  - Organizers: James Abello, AT&T Labs; A.Birilis, AT&T Labs; Steve Mahaney, NSF; Yossi Matias, Bell Labs; T.Munzner, Stanford; W. Sweldens, Bell Labs; Jeff Vitter, Duke University; Kiem-Phong Vo, AT&T Labs.
- **9th Annual Conference on Combinatorial Pattern Matching**
  - Dates: July 20-22, 1998
  - Location: Rutgers University
  - Organizers: Martin Farach, Rutgers University; Udi Manber, University of Arizona

Events Under Development

- **Workshop on Dimensionality, tentative March 1998**
- **Workshop on Biological Databases, tentative Spring 1998**
- **Implementation Challenge, tentative April 1998**
- **Workshop on Biological Applications, tentative June 1998**
- **Workshop on Analysis of Algorithms, tentative July 1998**

4.6 Special Year on Large Scale Discrete Optimization, 1998-99

Overview:

The Special Year on Large Scale Discrete Optimization will begin in September 1998. Some pre-special year activities are being held during 1997-98. The year builds on the 1992-93 Special Year on Combinatorial Optimization.

Organizers:

- Endre Boros, Rutgers University (local chair)
- William Pulleyblank, IBM
- Michael Trick, Carnegie Mellon University (general chair)
- Bob Vanderbei, Princeton University
- Vijay Vazirani, Georgia Tech

Sponsors:

The National Science Foundation
The New Jersey Commission on Science and Technology
AT&T Labs - Research
Bell Labs
Bellcore
Carnegie Mellon University
Georgia Institute of Technology
Rutgers University
Princeton University

Theme of Program:

This special year is motivated by developments in the past ten years. There have been three simultaneous advances: at the algorithmic level, with new and interesting algorithms for solving large scale discrete problems both exactly and approximately; at the implementation level, providing codes and systems for solving such problems; and at the practical level, with interest in the fast solution of real world problems. Examples of such applications include airline crew scheduling, chemical process design, and telecommunication network design. Our goal in this year is to unite groups working at the three above levels. The confluence of exciting recent work on approximation and other algorithms and greatly improved software/hardware for optimization makes large scale discrete optimization a practical, useful approach to solving problems of practical interest.

The special year organizers hope to provide forums to increase research contacts between groups that do not frequently come in contact - for example practitioners and academics, operations researchers and computer science researchers, theoretically oriented researchers (in areas such as interior point IP methods, combinatorial geometry approaches, approximation methods, and group formulations) and computationally oriented researchers, including those interested in AI techniques such as constraint programming.

Workshops Planned for the Rest of 1997-98

- **Joint DIMACS-CMU-Georgia Tech. Workshop on Large Scale Discrete Optimization**
- Dates: May 27-29, 1998
- Location: Rutgers University
- Organizers: R. Ravi, Carnegie Mellon University, and Vijay Vazirani, Georgia Tech.

Events Under Development for 1998-99

- **Large Scale Discrete Optimization in Engineering (including structure design and chemical process design)**
- **Large Scale Discrete Optimization in Production and Scheduling**
- **Large Scale Discrete Optimization in Transportation**
- **Large Scale Discrete Optimization in Robotics and Vision**
- **Algorithms for Large Scale Discrete Optimization**
- **Semidefinite Programming and Large Scale Discrete Optimization**
- **Constraint Programming and Large Scale Discrete Optimization**
- **Large Scale Discrete Optimization in Data Mining**
- **Algorithm Implementation Challenge on Large Scale Discrete Optimization**

4.7 Other Research Workshops, Conferences, and Institutes

Overview:

In addition to workshops that are part of the concentrated activity of special years, DIMACS has other workshops and conferences that are often used to visit topics of special current interest, that might develop into special focus topics, that "revisit" former special focus topics, or that involve new collaborations with other units.

Workshops, Miniworkshops and Special Events, 1997-98:

- **Workshop on Combinatorial Optimization and Disordered Materials: Recent Progress and Algorithmic Challenges**
  - Dates: May 15-17, 1997
  - Location: Rutgers University
  - Attendance: 85
  - Organizers: Deborah Franzblau, City University of New York; Phillip Duxbury, Michigan State University; Joel Lebowitz and Paul Leath, Rutgers University

This three day workshop was intended to encourage a closer dialog between computer science, mathematics and theory by using graph theory methods in the analysis of disordered systems. Topics discussed included: the use of algorithms from graph theory in the application of matching methods to rigidity percolation and the structure of glasses; the use of min-cut/max-flow and min-cost algorithms in the analysis of the ground states of random magnets and flux line arrays; and the use of the greedy algorithm to grow trees which led to new insights into percolation and the spin glass ground state.

A strength of this workshop was the fact that many of the leading people working in statistical mechanics and the mathematics of disordered systems provided the physical and mathematical background against which the new techniques should be tested. It was also useful to have selected talks on methods for "hard" combinatorial problems in disordered systems.

- **DIMATIA Inaugural Conference, The Future of Discrete Mathematics**
  - Dates: May 19-24, 1997
  - Location: Stirin Castle, Czech Republic
  - Attendance: 53
  - Organizers: Jarik Nesetril, Jan Kratochvil, DIMATIA; Fred Roberts, Rutgers University; Ronald Graham, AT&T Labs - Research

This six day conference was the first of several events planned to encourage the long-term collaboration of members of DIMACS and its partner center in the Czech Republic, DIMATIA, the Center for Discrete Mathematics and Theoretical Informatics. This conference in particular was intended to bring DIMATIA to the international attention of the DM/TCS community by coinciding with the formal opening of the center.

The program featured leaders in the discrete mathematical community discussing the major themes and trends that they see for the field, in particular in its connections with theoretical computer science and its growing role in various applications which range from optimization to molecular biology, and the directions they recommend be explored.

The conference can boast of having drawn an international mix of speakers with researchers from the USA, Czech Republic, Spain, Hungary, Russia, France, Austria, Canada, Poland, England and Germany. Participants included outstanding young researchers including two holders of the prestigious European award for young mathematicians.

The main lecturers discussed combinatorial number theory, relating to the work of the late Paul Erdős, and also with an emphasis on discrepancy; the solution of the even cycle problem; the common features between Ramsey theory and sequences of random variables; meaningless statements in combinatorial optimization; geometric discrepancy; efficient representability of 3-dimensional polytopes; new probabilistic methods in combinatorics; random trees and random graphs; algorithmic applications of the classification of simple groups; approximation algorithms; the algebraic method in graph theory; color critical graphs and hypergraphs; random walks, mixing, and sampling; minimax theorems related to connectivity augmentation; relations between combinatorics and mathematical logic, in particular the use of mathematical logic to sove the complexity of the cutting plane method and also in research on proof complexity; Ramsey theory and linear algebra; graph homomorphisms and branching random walks; extremal problems and the Szemeredi regularity lemma; graph theoretic aspects of the protein folding problem; and cryptography.

The program and abstracts were published in the KAM/DIMATIA series and proceedings are presently being prepared and will be published by AMS in the DIMACS series.

- **Workshop on Logic and Algorithms - One Year Later**
  - Dates: July 23-25, 1997
  - Location: Rutgers University
  - Attendance: 80
  - Organizers: Eric Allender, Rutgers University; Robert Kurshan, Bell Labs; Moshe Vardi, Rice University

  The 1995-1996 Special Year on Logic and Algorithms (SYLA) was aimed at bridging a dichotomy in computer science between two major branches of research, one in algorithms and complexity, and the other one in models and semantics. SYLA focused on three bridge areas: computer-aided verification, finite-model theory, and proof complexity.

  The goal of the "One Year Later" workshop was to follow up on SYLA. The program of the workshop consisted of 11 hour-long survey talks as well as 18 shorter 1/2-hour-long talks. The speakers reported on research results that arose during the Special Year or following it. Many of these results were obtained in collaborative work that started during SYLA. During the Special Year, we conducted many specialized workshops. In this workshop, we focused on bringing the three topics together, in an attempt to provide participants with a broader view of the field.

- **The DIMACS Research and Education Institute 1997 (DREI): Cryptography and Network Security**
  - Dates: July 27 - August 15, 1997
  - Location: Rutgers University
  - Attendance: 30 teachers, 170 researchers
  - Organizers: Stephen Greenfield, Rutgers University; Joan Feigenbaum, AT&T Labs - Research; Steven Rudich, Carnegie Mellon University

  DREI is a three week institute which attempts to encourage high school and college teachers to interact with researchers with the goal that the teachers, having more insight into the current advances in research, will include cutting edge research in their curriculum and thus encourage their students to pursue careers in the sciences, and that the researchers will gain new insights into pedagogy and interest in mathematical sciences eduation at all levels. Details about DREI can be found in Section 6.5. This 3-week summer institute featured three research workshops:

  - **Underlying Mathematics**
  - Dates: July 27 - August 2, 1997
  - Organizer: Eric Bach, University of Wisconsin

  The research program included sessions on number theory and number-theoretic algorithms, probability and statistics, information theory, cryptosystems, and cryptanalysis.

  - **Computer Science**
  - Dates: August 3-9, 1997
  - Organizer: Steven Rudich, Carnegie Mellon University

The research program included sessions on definitions and proofs of security, complexity theory, protocols, and formal models.

- **Network Security**
- Dates: August 10-15, 1997
- Organizer: Ross Anderson, Cambridge University

The research program included sessions on applications and implementations, threat models, break-ins, and infrastructure.

- **Workshop and Tutorial on Quantum Computing**
  - Dates: August 11-15, 1997
  - Location: Princeton University
  - Attendance: 60
  - Organizers: Peter Shor, AT&T Labs - Research; Andrew Yao, Princeton University

  A three-day tutorial on quantum computing was held August 11-13 and directed towards computer scientists; a two-day scientific workshop on recent results in quantum computing followed the tutorials.

  The tutorial topics included: quantum algorithms, quantum error-correcting codes and fault-tolerant computing, quantum information theory, and potential physical implementations of quantum computers. The tutorial was directed to computer scientists and mathematicians with relatively little knowledge of quantum computing, with the intention of giving them enough background to understand current research in quantum computing. Judging from the comments received by the organizers, the tutorial was very successful.

  The workshop talks were research talks submitted by participants, mainly on their current subjects of research. The topics ranged from quantum algorithms and quantum communication complexity through quantum coding theory to realizations of quantum computers. These talks included a number of exciting new results.

Events Planned for the Rest of 1997-98

- **Workshop on Discrete Mathematical Chemistry**
  - Dates: March 23-25, 1998
  - Location: Rutgers University
  - Organizers: Pierre Hansen, University of Montreal; Patrick Fowler, University of Exeter: Maolin Zheng, Chemical Abstracts
- **From Erdős to Algorithms: Applications of the "Probabilistic Method" - A Joint DIMACS/DIMATIA Workshop**
  - Dates: June 10-12, 1998
  - Location: Rutgers University
  - Organizers: Joel Spencer, New York University; Jarik Nesetril, Charles University and DIMATIA; Ronald Graham, AT&T Labs - Research; Jiri Matousek and Pavel Valtr, Charles University and DIMATIA

- **Control: Non-Linear, Hybrid, and Learning Approaches**
  - Dates: June 16-20, 1998
  - Location: Rutgers University
  - Organizer: Eduardo Sontag, Rutgers University
- **DIMACS Research and Education Institute 1998 (DREI): Topics in Graph Theory**
  - Dates: July 20 - August 7, 1998
  - Location: Rutgers University
  - Organizers: Linda Lesniak, Drew University; Glenn Hurlbert, Arizona State University

5 Research Highlights

5.1 Special Year in Networks

In addition to the research work arising from its STC-sponsored special year, DIMACS has two multi-disciplinary, multi-institutional research projects in network modeling and parallel simulations, each funded separately, but closely related to our Special Year in Networks. These projects relate to some very important problems in networking research, aiming to provide the methodology and software tools for creating software descriptions of very large networks – such as the global Internet – which could then be simulated on multiprocessor machines or on multiple networked machines, to speed up the simulations. Parallel discrete event simulations of telecommunications systems involve very difficult problems, and so far have not been successfully implemented in commercial systems despite a great desire for such tools from industry and the military. The scientific work on these issues is described in Sections 5.8.1 and 5.8.2.

A Local Search Heuristic for Facility Location Problems Madhukar Korupolu, Greg Plaxton, and Rajmohan Rajaraman (PD) have considered approximation algorithms for several NP-hard problems arising in facility location. In each problem, we are given: (i) a set of locations at which we may open a facility, (ii) a set of client locations, together with their demands, that require to be serviced by a facility, and (iii) for every pair of locations the cost of shipping a unit of demand between the two locations. In the *k-median problem*, the objective is to determine a set of at most k locations at which to open a facility so as to minimize the total shipping cost. In the *facility location problem*, we are also given a cost of opening a facility at any location, and the objective is to determine a set of locations at which to open facilities so as to minimize the total shipping and opening costs. We differentiate between the uncapacitated and the capacitated variants of the above problems. In the uncapacitated case, an unlimited amount of demand can be shipped from any facility, whereas in the capacitated case, there is an upper bound on the total demand that can be shipped from any facility. Koropolu, Plaxton, and Rajaraman proved that a simple local search heuristic yields polynomial-time constant-factor approximation bounds for the metric versions of the uncapacitated k-median and facility location problems. (For the k-median problem, their algorithms require a constant-factor blowup in the parameter k.) This local search heuristic was first proposed several decades ago, and has been shown to exhibit good practical performance in empirical studies. Koropolu, Plaxton and Rajaraman also extended the above results to obtain constant-factor approximation bounds for the metric versions of the capacitated k-median and facility location problems.

A Simple Distributed Algorithm for Placing Objects in Hierarchical Networks Madhukar Korupolu, Greg Plaxton, and Rajmohan Rajaraman (PD) have designed a simple algorithm to determine a placement of objects in hierarchical networks. Their placement has the property that the expected access time of a request is within a small constant factor of the optimal. Moreover, the placement is obtained by a single bottom-up pass of the hierarchical network and hence can be implemented in a distributed manner. The algorithm has practical potential in the design of World Wide Web caching schemes and in allocating files in distributed filesystems.

Nearly Linear Time Approximation Schemes for Euclidean TSP, Steiner Tree and other Geometric Problems Much effort has been invested in designing approximation algorithms for NP-hard geometric problems such as Euclidean Traveling Salesman and Minimum Steiner Tree. In 1996, Sanjeev Arora (PM, PU) presented new algorithms for many of these problems that are approximation schemes: for any fixed $c > 0$, the algorithm computes a solution of cost at most $1 + c$ times the optimal. The running time was $n^{O(1/c)}$ in the plane and slightly superpolynomial in higher dimensions. (Mitchell independently discovered a similar scheme for the planar case.) Arora has now found a new approximation scheme that is simpler and much more efficient. For any fixed $c > 0$, it finds a $(1+c)$-approximate solution in $n * \text{polylog}(n)$ time, which is nearly linear in n. The scheme also works in nearly linear time on instances in d dimensions (for fixed d). His algorithm generalizes to many other geometric problems, including the Minimum Steiner Tree problem, and the Euclidean Matching problem.

A New Model of Generalized Steiner Networks J.F. Weng (V) studied 2-dimensional networks with a Steiner topology. Suppose T is such a network. If any edge e joining two Steiner points s and s' satisfies the condition that the other edges of s, s' are, in pairs, parallelly lying on two sides of e, then T is called a *generalized Steiner network*. As an example, Weng showed that the drive shafts for an underground mine form a network of this type since the shafts have to have a maximum gradient. He gave an algorithm that can compute generalized Steiner networks with given Steiner topologies in linear time.

Interconnecting Highways As a result of the workshop on Network Design, Ding-zhu Du (V), Frank Hwang (PM), and Guoliang Xue worked on the problem of constructing roads of minimum total length to interconnect n highways under the constraint that the roads can intersect each highway only at one point in a designated interval which is a line-segment. They found a set of optimality conditions for the problem and showed how to construct a solution to meet this set of optimality conditions.

Compliance Checking in the PolicyMaker Trust Management System Emerging electronic commerce services that use public-key cryptography on a mass-market scale require sophisticated mechanisms for managing trust. For example, any service that receives a signed request for action is forced to answer the central question "Is the key used to sign this request authorized to take this action?" In some services, this question reduces to "Does this key belong to this person?" In others, the authorization question is more complicated, and resolving it requires techniques for formulating security policies and security credentials, determining whether particular sets of credentials satisfy the relevant policies, and deferring trust to third parties. Blaze, Feigenbaum, and Lacy identified this "trust management problem" as a distinct and important component of network services and described a general tool for addressing it, the "PolicyMaker trust management system." At the heart of a trust management system is an algorithm for "compliance checking." The inputs to the compliance checker are a request, a policy, and a set of credentials. The compliance checker returns yes or no, depending on whether the credentials constitute a "proof" that the request complies with the policy. Thus a central challenge in trust management is to find an appropriate notion of proof and an efficient algorithm for checking proofs of compliance. Stimulated by discussions at the DIMACS workshop on Trust Management in Networks, Matt Blaze (PM, ATTL), Joan Feigenbaum (PM, ATTL), and Martin Strauss (PM, ATTL) have developed the notion of proof that is used in the current version of the PolicyMaker trust management system. They showed that this notion of proof leads to a compliance-checking problem that is undecidable in its most general form and is NP-hard even if restricted in several natural ways. They identified a special case of the problem that is solvable in polynomial time and is widely applicable. The algorithm that they gave for this special case has been implemented and is used in the current version of the PolicyMaker system.

Reliable Communication over Partially Authenticated Networks Reliable communication between parties in a network is a basic requirement for executing any protocol. Amos Beimel (PD) and Matt Franklin (PM, ATTL) considered the effect on reliable communication when some pairs of parties have common authentication keys. The pairs sharing keys define a natural "authentication graph", which may be quite different from the "communication graph" of the network. They characterized when reliable communication is possible in terms of these two graphs, focusing on the very strong setting of a Byzantine adversary with unlimited computational resources. More specifically, reliable communication involves a transmitter which sends a message to a receiver. There are two versions of this problem: "single-pair" and "all-pairs". For single-pair reliable communication, one specifies the transmitter and receiver, and wants to succeed against any coalition of at most t faulty processors. For all-pairs reliable communication, one wants to succeed for any transmitter, any receiver, and any coalition of at most t faulty processors. The network of channels defines a natural "communication graph", with an edge between two vertices for every channel between two processors. The pairs of parties sharing authentication keys define a natural "authentication graph", with an edge between two vertices for every shared key. Success for the all-pairs reliable communication problem depends in a natural way on the connectivity of these graphs. Beimel and Franklin showed that all-pairs reliable communication is possible if and only if the communication graph is $(t+1)$-connected and the union of the two graphs is $(2t+1)$-connected. When there is a specific sender a and a specific receiver b, the situation is not so straightforward. In this case the characterization involves a family of inductively defined graphs, each one containing all edges of the communication graph and some of the edges of the authentication graph.

Escrowed Identity Joe Kilian (PM, NEC), and Erez Petrank (PD) worked on secure authorized access. The issue that arises is the tension between the need of a person to access services without revealing his identity, and the need to reveal identities of users in extreme cases, for example, when a crime is committed. Kilian and Petrank used Key escrow techniques to meet both requirements. Key escrow has proven an active and contentious field of research and discussion. Essentially all of the attention in this area has been restricted to the simple case of communication: party A sends an encrypted message $E_K(M)$ to party B; some centralized authority is given the capability to recover either K or the specific message M. Kilian and Petrank considered a new domain for the application of key escrow ideas: Escrow in identity schemes. One distinctive feature of this application is that Escrowed identity may actually enhance privacy. By default, many identity schemes often require a person to give their entire identity "up front." A protocol in which this information is only released under special circumstances may prove an acceptable, and more private, substitute. Another distinctive feature is that Escrowed identity schemes work to the advantage of at least one of the parties invoking them. In traditional key escrow systems, both parties only lose by following the escrow system, and have everything to gain by bypassing it (which is generally quite easy to do). The motivation for traditional identification schemes in access control, communication and commerce need not be discussed. A rapidly increasing number of interactions take place on the purely information level, and are likely to become ubiquitous. As just one example, on some toll bridges one can now electronically identify oneself to a toll booth instead of waiting in line to pay a toll. Although this application is cost effective (or will be) and more convenient, it raises critical issues of privacy; a more pervasive toll collection system might as a side effect allow the tracing of people's movements to an unprecedented degree. Kilian and Petrank addressed the tension between party A's desire for privacy and anonymity and B's (and society's) desire to know who they are interacting with. This balance point of this tension can shift under certain often rare circumstances. For example, consider an automated parking garage. What is paramount to the garage is that the person entering it is authorized to do so; the person's precise identity is normally not a valid concern. However, suppose that on some night a person was murdered in the garage. At this point, the garage owner and society at large may have a legitimate interest in knowing who was there on that night. Or, one might wish to have a computer "chat room" in which one has conditional anonymity messages: As long as one follows the rules, one's identity is secure from even the system administrator. But if one flagrantly breaks the rules (such as arranging drug deals in the "Lion King" kiddie chat room), suitable law enforcement agencies can be appealed to in order to determine one's identity. Traditionally, access control has been all or nothing. One obtains all the information about the other person up front, with no recourse to learn more if circumstances warrant. This rigidity generally leads one to allow less privacy, since one is likely to want as much information as one can get just in case a "bad case" arises. Kilian and Petrank have given a more flexible alternative. They considered a two-tier approach to identification. On the lower tier, a person gives only as much information about themselves as is strictly necessary for ordinary circumstances. In the garage example above, this means proving authorization to use the garage. On the second tier, a person gives a more precise statement of their identity that may be needed in extraordinary circumstances. This second tier is only accessible with the help of a third party, i.e., the escrow agency, which is preferably separate from and not under the control of the party managing access (such as the garage owners). An escrowed identity system consists of the following parties: An Identifier (the party who identifies themselves to the verifier, the user of the garage in the above example); An Issuer (the party which issues certificates that allow the identifier to identify themselves in an escrowed manner); A Verifier (typically the access provider who verifies the first-tier identity process as well as the escrow proof for the second-tier identity (the garage keeper in the above example); and Escrow Agent(s) (parties who use information forwarded by the verifier to make a second-tier identification of the identifier. These parties execute the following protocols: Issuing a certificate; checking the weak identity; recovering the complete identity. Kilian and Petrank have developed the properties of identity escrow systems. A typical requirement is that faking a legitimate identity is hard: A computationally bounded user that has not been issued a certificate by the issuer will fail to convince the verifier of having a certificate with probability almost 1. A second requirement is that ducking the escrow is hard: The probability that a computationally bounded user passes the check for weak identity but his identity cannot be recovered from the transcript by the escrow agents is almost 0. Kilian and Petrank have developed two concrete implementations of escrow identity systems. One is based on the RSA encryption scheme and the other on the El-Gamal signature and encryption schemes. An international patent is pending on this work (by Rutgers University and NEC).

Message Authentication Message authentication is one of the most used cryptographic primitives. In this setting, two parties which share a secret key would like to make sure that the messages they exchange are not changed on the communication channel before delivery. The Cipher Block Channing (CBC) Message Authentication Code (MAC) is an authentication method which is widely used in practice. Nevertheless, until recently no solid theoretical approach was suggested for proving security of such systems. The first paper to provide rigorous proof of security for CBC MAC (conditioned on the security of the underlying cryptographic block cipher) was by Bellare, Kilian and Rogaway in 1994. They suggested a framework for proving security of CBC MAC, and also used this framework to prove that the security of CBC MAC follows from the security of the underlying block cipher when the messages have fixed length. They also suggested variants of CBC MAC that handle variable length messages, but where the length of the message has to be known in advance. Erez Petrank (PD) and Charles Rackoff followed the work of Bellare, Kilian and Rogaway by studying the case in which the length of the message is not known until the message ends. This seems an important issue in particular because the transmission of fax is inherently like that: When the transmission starts, the machine does not know how many pages are going to be input. Petrank and Rackoff first presented a variant of CBC MAC, called *double MAC* (DMAC), which handles variable unknown lengths. They proved that its security is implied by the security of the underlying block cipher. Next, they argued that the basic CBC MAC is secure when applied to prefix free message space. A message space can be made prefix free by authenticating also the (usually hidden) last character which marks the end of the message. This character never appears inside the message.

5.2 Special Focus in Discrete Probability

Solution of a Question of Shannon. The *Shannon capacity* of an undirected graph G is the limit, as n tends to infinity, of the n-th root of the independence number of the nth normal power of G. Motivated by a question in information theory, Shannon conjectured in 1956 that the capacity of a disjoint union of two graphs is always equal to the sum of their capacities (he showed that it is always at least that big, and that equality holds when at least one of the graphs is perfect). Noga Alon (V) showed that this is false in a very strong sense: the capacity of a union can be bigger than any fixed power of the sum of the capacities. This has the following somewhat counter-intuitive information theoretic interpretation. By using two channels alternately, it is sometimes possible to obtain capacity that exceeds considerably the sum of the capacities of each of them.

Statistical Mechanics, the Bethe Lattice, and Random 3-Colorings of a Binary Tree Motivated by problems from statistical mechanics where we try to model the behavior of physical phenomena (like magnetism) using the random behavior of small particles, Graham Brightwell (V) and Peter Winkler (PM, BL) have done extensive work on the Bethe lattice (Cayley tree). They have obtained two theorems classifying systems which exhibit phase transitions, and another classification involving ergodicity of graph homomorphisms, plus results on memory in branching random walks. The first classification theorem specifies the constraint graphs for which there are node activities with multiple Gibbs distributions. An illustration from the paper by Brightwell and Winkler appears on the cover of the 1997-98 catalog of the Graduate School of Rutgers - New Brunswick. The illustration shows a random 3-coloring of a binary tree subject to the condition that no red point may be adjacent to a green one. Brightwell and Winkler have used these random colorings as an abstract mathematical model of the phenomena studied in statistical mechanics and have shown that such colorings can exhibit phase transitions akin to the phase transitions exhibited by physical systems. (The illustration shows a "phase" in which green is favored over red, even though the local conditions are the same for those two colors.) The picture was chosen for the cover of the catalog in part because it symbolized the collaborations between mathematicians and computer scientists and researchers in other fields of science, such as physics, chemistry, biology, or the social sciences, collaborations in which graduate students participated through DIMACS.

A Startling Example in Statistical Mechanics Graham Brightwell (V), Olle Haggstrom (V), and Peter Winkler (PM, BL) have used Brightwell-Winkler theory (above), together with an idea of Schonmann and Tanaka, to devise a startling example in statistical mechanics. Random green-yellow-red colorings, with no red node adjacent to a green one and equal conditional probability x for red and green, constitute the Widom-Rowlinson model of statistical mechanics. On a cubic lattice it is known that when x is high one gets multiple phases, i.e., spontaneous symmetry-breaking, and when x is low one doesn't. Physicists believe that there is just one critical value of x where the change takes place, but this remains open (as it does for various other models as well). Applying the calculations on the tree version of this model to the k-branching tree and modifying the tree led Brightwell, Haggstrom, and Winkler to an example where there were three critical points. They proved that there is a unique critical point on the k-branching tree, namely $x = ((k+1)/k)^k/(k-1)$; but if one takes k=40 and adds 7 dangling edges to each node of the tree, bizarre things happen: symmetry is broken for x in the open interval (.217935, .401318) and again after $x = 4.55188$, but elsewhere there is only one (symmetric) phase. There are other known examples of triple critical points but they involve external fields; we believe this is the first case of multiple critical points in a symmetric system.

A Tutte Polynomial for Colored Graphs Bela Bollobas (V) and Oliver Riordan defined a polynomial W on graphs with colors on the edges, by generalizing the spanning tree expansion of the Tutte polynomial as far as possible. In particular, they gave necessary and sufficient conditions on the edge weights for this expansion not to depend on the order used and gave a contraction-deletion formula for W analogous to that for the Tutte polynomial. They showed that generalizations of the Tutte polynomial obtained from its rank generating function formulation, or from a random cluster model, can be obtained from W. Finally, they found the most general conditions under which W gives rise to a link invariant.

Dependent Random Graphs and Spatial Epidemics Jacob van den Berg (V), G.R. Grimmett, and R.B. Schinazi have extended certain exponential decay results of subcritical percolation to a class of locally dependent random graphs, introduced by Kuulasmaa as models for d-dimensional spatial epidemics. In these models, infected individuals eventually die (are removed) and are not replaced. They combined these results with certain continuity and re-scaling arguments in order to improve our knowledge of the phase diagram of a modified epidemic model in which new susceptibles are born at some positive rate. In particular, they showed that, throughout an intermediate phase where the infection rate lies between two certain critical values, no coexistence is possible for sufficiently small positive values of the recovery rate. This result provides a converse to results of Durrett–Neuhauser and Andjel–Schinazi. They also showed that such an intermediate phase indeed exists for every $d \geq 1$ (i.e., that the two critical values mentioned above are distinct). An important technique used is the general version of the BK inequality for disjoint occurrence, proved in 1994 by David Reimer (GS, RU).

The Existence of Finitary Codings for a Class of Random Fields. Jeff Steif (V) and Jacob van den Berg (V) studied the existence of finitary codings from an i.i.d. process to Markov random fields. They showed that the presence of a phase transition is an obstruction for the existence of the above coding; this yields a large class of Bernoulli shifts for which no such coding exists. Conversely, they showed that for the stationary distribution of a monotone exponentially ergodic probabilistic cellular automata such a coding exists. These constructions were inspired by the Propp-Wilson algorithm for exact simulation. In particular, combining the above with a result of Martinelli and Olivieri, this yielded the fact that for the plus state for a d-dimensional ferromagnetic Ising model, $d \geq 2$, there is (not) such a map when the interaction parameter is below (above) its critical value.

Threshold of Broadcast in Random Graphs A broadcasting process involves sending a piece of information which resides at one vertex in a graph to all the remaining vertices. At each time step, a vertex knowing the information can send it to one of its neighbors. The broadcast problem is to find the minimum number of time steps needed. The problem is NP hard in general. It is not hard to show that the edges used by the optimal broadcast algorithm form a tree. The broadcast problem is to find the tree and the minimum broadcast time. Hui Chen (PD) considered the problem, for a random graph $G_{n,p}$, of finding the threshold for p such that with high probability (whp), a piece of information can be broadcasted in the graph in exactly $\lceil \log_2 n \rceil$ steps. (An event E_n is said to occur whp if $\mathbf{Pr}[E_n] = 1 - o(1)$ as $n \to \infty$.) The best upperbound to date is due to Frieze and Molloy who showed that if $p \geq 18 \ln n/n$ then whp, there exists a rooted broadcasting tree on every vertex. Chen showed that a trivial lower bound is also sufficient for the existence of a rooted optimal broadcast tree on each vertex of a random graph $G_{n,p}$.

Mixing Times of Lozenge Tiling and Card Shuffling Markov Chains David Wilson (V) introduced a simple technique that can be applied to a number of classes of Markov chains to obtain upper bounds and lower bounds on their mixing times, the number of steps the Markov chains take to approach their equilibrium distribution. One application is to a class of Markov chains introduced by Luby, Randall, and Sinclair to generate random tilings of regions by lozenges. The mixing time bounds derived by Wilson improve substantially upon the previous results. In another application he resolved a few questions raised by Diaconis and Saloff-Coste, by lower bounding the mixing time of various card shuffling Markov chains. His lower bounds are within a constant factor of their upper bounds. This resolved a question that had been open for about a decade. When he applied this technique to modify an analysis of Bubley and Dyer, he obtained an $O(n^3 \log n)$ upper bound on the mixing time of the Karzanov-Khachiyan Markov chain for linear extensions.

Exact Sampling Algorithms for Storage and Network Models Robert Lund and David Wilson (V) developed algorithms for simulating random samples from the invariant measure of storage and network models without relying upon the analytic form of the invariant measure. Since the invariant measure of many storage and network models is intractable, the proposed algorithms can be used to study the invariant measure of such systems through simulation techniques. Backwards coupling methods were used along with the sample path monotonicities inherent in these models to develop an efficient algorithm for simulating from the limiting distribution. Previously, some sort of truncation was required.

Sampling Points in a Convex Body Russ Bubley, Martin Dyer and Mark Jerrum (V) have described a new method for proving the polynomial-time convergence of an algorithm for sampling (almost) uniformly at random from a convex body in high dimension. Previous approaches had been based on estimating conductance via isoperimetric inequalities. Bubley, Dyer, and Jerrum showed that a more elementary coupling argument can be used to give a similar result.

Set Partitions with Unequal Block Sizes A. Knopfmacher, A. M. Odlyzko (PM, ATTL), B. Pittel (V), L. B. Richmond, D. Stark, G. Szekeres, and N.C. Wormald have determined the asymptotic behavior of the number of set partitions of an n-element set into blocks of distinct sizes. The behavior is more complicated than is typical for the set partition problems. Although there is a simple generating function, the usual analytic methods for estimating coefficients fail in the direct approach.

Confirming two Conjectures about Integer Partitions Boris Pittel (V) has settled two old conjectures about integer partitions. For a given integer n, let Λ_n denote the set of all integer partitions $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0$, $(m \geq 1)$, of n. For the dominance order "$\preceq$"on Λ_n, Pittel showed that if two partitions λ, μ are both chosen from Λ_n uniformly at random, and independently of each other, then $\Pr(\lambda \preceq \mu) \to 0$ as $n \to \infty$. This statement answers affirmatively a question posed by Macdonald in 1979. The proof is based on the limit joint distribution of the largest parts counts found by Fristedt. A slight modification of the argument confirms a conjecture made by Wilf in 1982, namely that, for n even, the probability of a random partition being graphical is zero in the limit.

Matchings and Hamiltonian Cycles in Sums of Random Trees A *random tree* on the set $[n] = \{1, \ldots, n\}$ is a tree on this set chosen uniformly at random from the family of all trees on the set $[n]$. Let k be a positive integer. For trees $T_1, \ldots, T_k$, all of them on the set $[n]$, we define their sum as the graph on the vertex set $[n]$ and edge set being the union of edge sets of the trees $T_1, \ldots, T_k$, where the parallel edges coalesce. We denote by $\mathbf{ST}_n(k)$ the sum of k random trees on $[n]$. Let f be a mapping from $[n] \to [n]$. Let $D(f)$ be its associated functional digraph, i.e., the graph with vertex set $[n]$ and edges $(i, f(i))$, $i \in [n]$. Similarly to the sum of trees we define the sum of mappings. We use $\mathbf{SM}_n(k)$ for the sum of k random mappings. Alan Frieze showed in 1996 that

$\mathbf{SM}_n(2)$ almost surely has a perfect matching and Cooper and Frieze showed in 1996 that $\mathbf{SM}_n(4)$ almost surely has a hamiltonian cycle. The problem of whether or not $\mathbf{SM}_n(3)$ almost surely has a hamiltonian cycle is one of the most important open questions in the theory of random graphs. Alan Frieze (V), Michel Karonski, and Lubos Thoma (PD) have proved that

$$\lim_{\substack{n\to\infty \\ n \text{ even}}} Prob(\mathbf{ST}_n(1) \text{ has a perfect matching }) = 0$$

and

$$\lim_{\substack{n\to\infty \\ n \text{ even}}} Prob(\mathbf{ST}_n(2) \text{ has a perfect matching }) = 1$$

and that

$$\lim_{n\to\infty} \text{Prob}(\mathbf{ST}_n(5) \text{ has a hamiltonian cycle }) = 1.$$

Probability that a Random Structure has a Given Property Joel Spencer (V) and Lubos Thoma (PD) have developed a procedure characterizing limit values of probabilities for a random structure having a property from a given language L. The method is based on completions of almost sure theories. Let L be a (first or second order) language. For each $n \in \mathbf{N}$ we are given a probability space (Ω_n, μ_n), where Ω_n is a finite set of models of L and μ_n is a probabilistic measure on Ω_n. (In the particular application by Spencer and Thoma, we set (Ω_n, μ_n) to be the random graph $G(n,p)$.) We assume all subsets of Ω_n are μ_n-measurable. For $\psi \in L$ we write $\mu_n(\psi)$ for $\mu_n(\{M \in \Omega_n : M \models \psi\})$, the probability ψ holds in a "random" model. In 1991, T. Luczak and J. Spencer classified probabilities p with $p = \Theta(\log n/n)$ at which the $0-1$ law does not hold, i.e., thresholds in the neighborhood of the point where the random graph $G(n,p)$ becomes connected. Namely, the 0-1 law does not hold for the probabilities of the form $p_{k,l} = p_{k,l}(n) := (\ln n + l \ln\ln n + c + o(1))/(kn)$ where $l \geq k-1 \geq 0$. Spencer and Thoma have obtained a full characterization of limit probabilities for the random graph $G(n, p_{k,l})$ and first order properties of graphs. Due to the generality of the results, one can expect the procedure developed by Spencer and Thoma to play a role in characterizing limiting probabilities for other random structures.

DIMACS special foci cross-fertilize each other. There was considerable cross-fertilization between the special focus on discrete probability and that on molecular biology. The following results are examples of the impact of that cross-fertilization, and were already mentioned in Section 5.2.

Thresholds for Ordering Libraries of Variable Length Clones One of the many biological problems to which mathematicians have given attention in recent years is the Physical Mapping Problem from genetics. To attain a physical map, many copies of a genome are digested by various enzymes, creating a library of overlapping clones. One technique involves further digestion of the clones to get sets of digestion lengths. Hui Chen (PD) has considered the technique of testing the clones for the existence of various DNA subsequences called probes. If the probes are (relatively) long, they will appear at most once in the genome, and they are called unique probes. If the probes are (relatively) short, their appearances throughout the genome will be according to a Poisson process (assuming randomness of the base pairs in the genome), and these are called Poisson probes. In either case, clone overlaps can be detected by commonality of probes, and this information can be used to reconstruct the actual overlap structure of the clone library, and the ordering of the probes on the genome. This information constitutes the physical map, which can be used by biologists for further sequencing, or as a guide to other genetic research. Chen showed that a gapless distribution of clones, with distinct endpoints, has a unique ordering of the atomic intervals (up to complete reversal) satisfying the so-called consecutive 1's property if and only if there are no two atomic intervals the same. The consecutive 1's property provides a simple test of whether or not a graph is an interval graph, a central concept in the history of molecular genetics and in particular in physical mapping.

Sequence Reconstruction Problems of sequence reconstruction led to some interesting results by Richard Arratia (V), Bela Bollobas (V), Don Coppersmith, and Greg Sorkin. The original question they considered arises in SBH (Sequencing by Hybridization), which is the attempt to reconstruct a long sequence from the knowlege of which q-tuples it contains, without being given the knowledge of the order in which those q-tuples occur. For random sequences, through deBruijn graphs, this reduces to a question of enumerating Euler circuits. Arratia, et al. showed that the interlacement graph, which gives only the relation between pairs of repeated pairs of (q-1)-tuples, determines the number of Euler circuits. The tool for this leads to a new graph polynomial, which may be of independent interest.

5.3 Special Year in Mathematical Support for Molecular Biology

Structural Motif Recognition Biologists are interested in predicting protein structure because the structure of a protein provides insight into its biological function. Experimental approaches to determining a protein's structure are difficult and time-consuming. As a result, much recent research has focused on applying mathematical and computational techniques to predicting protein structure from one dimensional sequence data. A first step to tackling this complex problem is a solution to the *structural motif recognition problem*: given a known commonly occurring three-dimensional substructure, or motif, determine whether this motif occurs in a given amino acid sequence. Bonnie Berger (V) and Mona Singh (PD) developed an iterative algorithm that uses randomness and statistical techniques to improve existing methods for recognizing protein structural motifs. The algorithm is particularly effective in situations where large numbers of sufficiently diverse examples of the motif are not known. These are precisely the situations that pose significant difficulties for previously known methods. They have implemented their algorithm, and showed that it is able to substantially improve performance on the coiled coil motif. In collaboration with biologists at the Whitehead Institute and Genentech Corporation, Berger, Singh, with Peter Kim, James Berger, and Andrea Cochran, have applied their program to the histidine kinase protein family. Most members of the protein family are sensors that allow microorganisms to adjust to varying conditions in their environments. Their program identifies a common but previously undescribed coiled-coil region in the histidine kinase family. Their work also suggests mechanisms by which the predicted coiled coil may mediate kinase regulation and signal transduction. There are some interesting medical applications, as some dangerous bacteria, such as those causing Whooping cough and Salmonella, need histidine kinase proteins to be fully virulent.

The HP Model for Protein Folding Richa Agarwala (PD), Serafim Batzoglou, Vlado Dancik, Scott Decatur (PD), Martin Farach (PM, RU), Sridhar Hannenhalli (V) and Steven Skiena (V) worked on the problem of determining the three-dimensional folding of a protein given its one-dimensional amino acid sequence. They used the HP model for protein folding proposed by Dill in 1985, which models each protein as a chain of amino acid residues that are either hydrophobic or polar, and makes the assumption that hydrophobic interactions are the dominant initial driving force for protein folding. Although it is not known how to find the optimal fold in polynomial time, Hart and Istrail in 1995 gave polynomial time approximation algorithms for folding proteins on the cubic lattice under the HP model. Agarwala, et al. examined the choice of lattice by considering its algorithmic and geometric implications and argued that a triangular lattice is a more reasonable choice. They presented a set of folding rules for a triangular lattice and analyzed the approximation ratio which they achieved. In addition, they introduced a generalization of the HP model to account for residues having different levels of hydrophobicity. After developing the biological foundation for this generalization, they showed that in the new model they were able to achieve similar constant factor approximation guarantees on the triangular lattice as were achieved in the standard HP model. While the structures derived from their folding rules are probably still far from biological reality, it is hoped that having a set of folding rules with different properties will yield more interesting folds when combined.

Ultraselfish Genes in Mice The ability of the t-haplotype, an ultraselfish gene in mice, to persist in natural populations has been a subject of debate in the theoretical biology literature for forty years. *Ultraselfish genes*, which have been observed in several species, are genes that are both aggressive and deleterious. They are transmitted to offspring more often than their genetic counterparts in violation of Mendel's laws, yet they are harmful to the organism that carries them in that they cause sterility and often carry recessive lethal mutations. Ultraselfish genes have also been observed in fruitflies, mosquitoes, wheat, tomatoes and fungi. Because such genes challenge the basic notions of Mendelian genetics and Darwinian evolution, understanding their behavior in populations is an important research problem in evolutionary biology. Dannie Durand (V) has been working on problems in this area in collaboration with Lee Silver in the Molecular Biology department at Princeton since shortly after the Special Year in Mathematical Support for Molecular Biology started in 1994-95. Their first project, addressing the ability of the t-haplotype to persist at low frequencies in natural populations, was reported in *Genetics* in April 1997. Their results overturned the longstanding hypothesis that this low level persistence is due to a balance between loss of the gene due to stochastic effects in small populations and reintroduction of the gene through migration. Using Monte Carlo simulation, they showed that although the proposed model does predict the gene frequencies seen in nature, the system is not stable and therefore not a biologically plausible explanation for the observed phenomena. Their simulations suggested that, in the context of the proposed model, stability can only be obtained if the aggressive transmission of ultraselfish genes is reduced through mechanisms such as reduced fitness or female choice. These new hypotheses must now be tested through laboratory and field work. Currently Durand and Silver are working on a study of the role of stochastic variation in the evolution of the t-haplotype. Once an ultraselfish gene arises, what properties maximize its chance for survival? The stochastic fluctuations in gene frequencies in finite populations, called genetic drift, have been shown to be a major mechanism in the evolution of normal genes but, until now, there has been no systematic study of genetic drift in populations of ultraselfish genes. Durand and Warren Ewens developed a Markov model of lethal t-haplotypes in finite populations. To their knowledge, this is the first analytical model of genetic drift of ultraselfish genes. All previous studies have used simulations or infinite population models. They also used Monte Carlo simulations to predict the variation in expected gene frequency and the time to the loss of both the sterile and the lethal variants of the t-haplotype as a function of population size and aggressiveness. The results, which are surprisingly different from the normal case, explain why the particular t-haplotype variants seen in nature have a competitive advantage. They suggest answers to two outstanding questions in t-haplotype evolution: Why are t-haplotypes in nature transmitted to 90% - and not 100% - of offspring? Why did the lethal variant only appear recently in the history of t-haplotype evolution?

Constructing Optical Maps by Physical Mapping Optical Mapping is an emerging new technology for sequencing DNA. S. Muthukrishnan (PM, BL and former PD) and L. Parida have formulated combinatorial problems relevant for realizing this technology and provided efficient theoretical and practical algorithms for these problems.

Group Testing with Sequences Many tests can be carried out on a set of samples, rather than on a single sample. For example, if one wants to detect the presence of an antibody in a set of samples, one can simply mix a fraction of each sample and test for the antibody in the newly formed sample. If the test is sensitive enough, one will detect if the antibody is present in any of the constituent samples. *Group testing* is the problem of efficiently finding out information about specific samples from tests on groups of samples. Group testing issues arise in genomics from problems like clone-probe incidence detection. Motivated by the economics of maintaining a clone library by robotic means, Martin Farach (PM, RU), Sampath Kannan (V), Manny Knill (V), and S. Muthukrishnan (PM, BL) designed and posed many novel group testing problems. They solved many of these problems by using classical coding theory results.

Longest Common Consecutive Substring in two Random Strings Let Σ be a finite alphabet with C letters. For any two strings $x = x_0 \cdots x_{n-1}$ and $y = y_0 \cdots y_{n-1}$ of length n, let $S(x, y)$ denote the size of the longest common consecutive substring between x and y; that is, $S(x, y)$ is the largest k such that,

$$x_i \cdots x_{i+k-1} = y_j \cdots y_{j+k-1}$$

for some $i, j \in \{0, \ldots, n-1\}$ (Here the indices are modulo n.) Sarmad Abbasi (GS, RU) showed that for x and y chosen uniformly among all possible strings of length n, $S(x, y)$ is highly concentrated around $2 \log_C n$. More precisely, for any $a \geq 1$

$$\Pr[|S(x, y) - 2 \log_C n| > a] \leq \frac{18C^{-a}}{n^2}.$$

This enabled him to show that the expected value of $S(x, y)$ is within a constant additive term of $2 \log_C n$.

Constructing evolutionary trees is a fundamental problem of biology. We have obtained a variety of results about this problem, as described in the following examples.

Nearly Tight Bounds on the Learnability of Evolution Evolution is often modeled as a stochastic process which modifies DNA. One of the most popular and successful such processes are the Cavender-Farris (CF) trees, which are represented as edge weighted trees. The *Phylogeny Construction Problem* is: given k samples drawn from a CF tree, output a CF tree which is close to the original. Each CF tree naturally defines a random variable, and the gold-standard for reconstructing such trees is the Maximum Likelihood Estimator of this variable. This approach is notoriously computationally expensive. Andris Ambainis (V), M. Farer, Martin Farach (PM, RU), and Sampath Kannan (V) showed that a very simple algorithm, which is a variant on one of the most popular algorithms used by practitioners, converges on the true tree at a rate which differs from the optimum by a constant. They did this by analyzing upper and lower bounds for the convergence rate of learning very simple CF trees, and then showing that the learnability of each CF tree is sandwiched between two such simpler trees. The results relied on the fact that, if the right metric is used, the likelihood space of CF trees is smooth.

The Maximum Agreement Subtree The *Maximum Agreement Subtree (MAST)* of a set of leaf-labelled trees (such as phylogenies) is the largest tree which is a subtree of all the input trees. For two trees, Martin Farach (PM, RU), Teresa Przytycka (V) and Mikkel Thorup (V) gave efficient algorithms for finding the MAST. In the case of k trees, the *kMAST problem*, it is NP-hard to find the MAST, even if $k = 3$, assuming the degree of every tree is bounded. However, Farach, Przytycka, and Thorup were able to show that the problem is polynomial as long as one of the input trees has bounded degree. This is almost always the case since phylogenies are usually binary. Their algorithm is fast in theory and in practice and has been used to enumerate all possible solutions.

The Linear-Cost Subtree-Transfer Distance between Phylogenetic Trees Different phylogenetic trees for the same group of species are often produced either by procedures that use diverse optimality criteria or from different genes in the study of molecular evolution. Comparing these trees to find their similarities and dissimilarities (i.e., distance) is thus an important issue in computational molecular biology. Several distance metrics including the nearest neighbor interchange (nni) distance and the subtree-transfer distance have been proposed and extensively studied in the literature. Bhaskar DasGupta (V), Xin He, Tao Jiang (V), Ming Li (V) and John Tromp have considered a natural extension of the subtree-transfer distance, called the linear-cost subtree-transfer distance, and studied the complexity and efficient approximation algorithms for this distance as well as its relationship to the nni distance. The linear-cost subtree-transfer model seems more logical than the (unit-cost) subtree-transfer model in some applications. They showed that the linear-cost subtree-transfer distance is in fact identical to the nni distance on unweighted phylogenies. They found an algorithm to compute an optimal linear-cost subtree-transfer sequence between unweighted phylogenies in $O(n^2 \log n + n2^{O(d)})$ time, where d denotes the linear-cost subtree-transfer distance. Such an algorithm is usual when d is small. They showed that computing the linear-cost subtree-transfer distance between two weighted phylogenetic trees is NP-hard, provided one allows multiple leaves of a tree to share the same label (i.e., the trees are not necessarily uniquely labeled). Finally they showed that there is an efficient approximation algorithm for computing the linear-cost subtree-transfer distance between weighted phylogenies with performance ratio 2.

The Number of Nucleotide Sites Needed to Accurately Reconstruct Large Evolutionary Trees. Biologists seek to reconstruct evolutionary trees for increasing number of species, n, from aligned genetic sequences. How fast must the sequence length N grow, as a function of n, in order to accurately recover the underlying tree with probability $1 - \epsilon$, if the sequences evolve according to simple stochastic models of nucleotide substitution? Mike Steel (V), Laszlo Szekely (V), and Peter Erdős (V) showed that for a certain model, a reconstruction method exists for which the sequence length N can grow surprisingly slowly with n (sublinearly for a wide range of parameters, and even as a power of $\log n$ in a narrow range, which roughly meets the lower bound from information theory). By contrast a more traditional technique (maximum compatibility) provably requires N to grow faster than linearly in n. Their approach is based on a new, and computationally efficient approach for reconstructing phylogenetic trees from aligned DNA sequences.

Constructing Big Trees from Short Sequences Peter Erdős (V), K. Rice, Mike Steel (V), Laszlo Szekely (V), and Tandy Warnow (V) developed a new method, which they called the *Short Quartet Method*, for inferring evolutionary trees. The Short Quartet Method has great statistical power, is provably consistent throughout the parameter space, and uses only polynomial time. Erdős, Steel, Szekely and Warnow have also shown that their results extend to Markov-style mutation models.

Hybrid Tree Construction Methods S. Nettles, K. Rice, Tandy Warnow (V), and Shibu Yooseph (PD) have explored the problem of reconstructing very large evolutionary trees from biomolecular sequences. Using both analytical and experimental techniques, they showed that each of the existing popular methods has significant problems in obtaining accurate reconstructions of some trees. As suggested by their experimental results, they then proposed a new method that combines the existing methods into hybrid methods. They studied one such hybrid experimentally, and showed that in significant parts of the parameter space, it achieves results that are equal to or better than the best of its components. Their experimental performance study examined more than a thousand sets of sequences simulated on more than a hundred different model trees.

Better Methods for Solving Parsimony and Compatibility Evolutionary tree reconstruction is a challenging problem with important applications not only in Biology but also in Linguistics. In Biology, one of the most promising approaches to tree reconstruction is to find the "maximum parsimony" tree, while in Linguistics, the use of the "maximum compatibility" method has been very useful. However, these problems are NP-hard, and current approaches to solving these problems amount to heuristic searches through the space of possible tree topologies (a search which can, on large trees, take months to complete). Maria Bonet (PD), Mike Steel (V), Tandy Warnow (V), and Shibu Yooseph (PD) found a new technique, *Optimal Tree Refinement*, for reconstructing very large trees. Their technique was motivated by recent experimental studies which have shown that certain polynomial time methods reliably return *contractions* of the true tree. They have studied the use of this technique in solving both parsimony and compatibility problems, and obtained both hardness results and polynomial time algorithms.

DIMACS special foci cross-fertilize each other. There was considerable cross-fertilization between the special focus on discrete probability and that on molecular biology. The following results are examples of the impact of that cross-fertilization, and were already mentioned in Section 5.2.

Thresholds for Ordering Libraries of Variable Length Clones One of the many biological problems to which mathematicians have given attention in recent years is the Physical Mapping Problem from genetics. To attain a physical map, many copies of a genome are digested by various enzymes, creating a library of overlapping clones. One technique involves further digestion of the clones to get sets of digestion lengths. Hui Chen (PD) has considered the technique of testing the clones for the existence of various DNA subsequences called probes. If the probes are (relatively) long, they will appear at most once in the genome, and they are called unique probes. If the probes are (relatively) short, their appearances throughout the genome will be according to a Poisson process (assuming randomness of the base pairs in the genome), and these are called Poisson probes. In either case, clone overlaps can be detected by commonality of probes, and this information can be used to reconstruct the actual overlap structure of the clone library, and the ordering of the probes on the genome. This information constitutes the physical map, which can be used by biologists for further sequencing, or as a guide to other genetic research. Chen showed that a gapless distribution of clones, with distinct endpoints, has a unique ordering of the atomic intervals (up to complete reversal) satisfying the so-called consecutive 1's property if and only if there are no two atomic intervals the same. The consecutive 1's property provides a simple test of whether or not a graph is an interval graph, a central concept in the history of molecular genetics and in particular in physical mapping.

Sequence Reconstruction Problems of sequence reconstruction led to some interesting results by Richard Arratia (V), Bela Bollobas (V), Don Coppersmith, and Greg Sorkin. The original question they considered arises in SBH (Sequencing By Hybridization), which is the attempt to reconstruct a long sequence from the knowlege of which q-tuples it contains, without being given the knowledge of the order in which those q-tuples occur. For random sequences, through deBruijn graphs, this reduces to a question of enumerating Euler circuits. Arratia, et al. showed that the interlacement graph, which gives only the relation between pairs of repeated pairs of $(q\text{-}1)$-tuples, determines the number of Euler circuits. The tool for this leads to a new graph polynomial, which may be of independent interest.

5.4 Special Focus on DNA Computing

The Special Focus on DNA computing has given impetus to a strong collaboration between Princeton molecular and evolutionary biologist Laura Landweber and computer scientist Richard Lipton (PM, PU). Their interdisciplinary program of research has led to several projects in Landweber's laboratory which combine aspects of evolutionary biology with chemistry and computer science. Results are preliminary at this time, but we describe them below.

RNA Editing One of the biological phenomena Landweber has been exploring in the past few years is RNA editing. Through DNA computing, she has begun to think of this process as an example of a truly RNA-based computer, and using DIMACS support, has been collaborating with computer scientist Lila Kari to model RNA editing as a cellular Turing machine. Landweber and Kari met through DIMACS workshops on DNA computing.

Feedback Loops in DNA Computing Laura Landweber, Richard Lipton (PM, PU), and Andrew Ellington (V), a combinatorial chemist, have set up a productive collaboration. Together, they are trying several approaches to DNA computing, including building a small "circuit" or feedback loop based on ribozymes, or catalytic RNA. They have also incorporated novel DNA bases into their research, making use of an "expanded genetic alphabet" for DNA computing (which uses more base pairs than just A-T and G-C for extra versatility). Thus they hope to bring DNA computing into the area of cutting edge biotechnology research.

DNA Scrambling A cellular process called DNA scrambling which takes place in ciliated protozoa can be viewed as a hard path problem. The single celled organism routinely solves a much harder path problem than Adleman's seven city problem, by unscrambling 50 pieces of a gene through a similar assembly process in only 9-18 hours of replication. Laura Landweber and Hans Lipps are studying this phenomenon as the best example of a cellular computer that can solve a difficult problem - the assembly of a functional gene. Thus DNA computing can be viewed as a way to understand both natural history and the pathway for the assembly of genes. Their results so far have discovered that the unscrambling process does take place in the DNA polymerase alpha gene of Stylonychia lemnae, a ciliate which has the benefit of a working transformation system developed in Lipps' laboratory. By inserting novel scrambled genes into this ciliate, Landweber plans to test the complexity of problems which a ciliate DNA computer can solve.

5.5 Special Year in Massive Data Sets

Optimization Algorithms for Separable Functions with Tree-like Adjacency Variables and their Application to the Analysis of Massive Data Sets Ilya Muchnik (V) and Vadim Mottl (V) developed a new method for very large image, signal and sequence analysis. The base of the method is an extended Bellman function structure which can be efficiently optimized on trees. Many problems from image analysis involve massive data sets. Such a problem can be reformulated as a mapping problem - to transform an original image with a lot of a shape details into "a simple" shape image which is a picture filtered from the original one (in a smoothing problem) or a "symbolic map" which represents the image by a planar graph of a small number of connected fragments-regions each of which is labeled by "a pattern shape"(in a segmentation problem). Such a mapping problem is naturally formulated by using an objective function which depends on a huge number of variables associated with nodes and edges of the basic grid on which the original image is presented. A similar approach is useful for signal (1-dimensional image) analysis and for sequence analysis. The characterization of the above application problems as problems of global optimization is useful in helping us understand them, but on the surface solving these seems unrealistic because of their complexity. However, although the number of variables is large, they are organized in the geometrical lattice-graph with a very simple structure of neighborhoods for both node and edge variables. This allows decomposition into a sequence of optimization subproblems on trees. The crucial idea of the approach is that the optimization problem on a tree is solvable by a dynamic programming approach. The procedure is based on a recursive representation of the corresponding extended Bellman function. The results of Muchnik and Mottl have been applied to real data and the ideas have been extended to apply to sequence alignment problems.

Combinatorial Optimization in Clustering To find some part of a data set which is stable, has a particular property and can be separated from the complementary part of a very heterogeneous large data set, is a core of the apparatus of analysis which is called clustering. There are two main groups of clustering models - statistical and combinatorial. The statistical group is the most popular in data mining today. At the same time, combinatorial models have some value in analysis of massive data sets, because they focus on designing efficient procedures; their results can be interpreted more naturally because they have a simple logical characterization; the results have an extreme characteristic which achieves "the global best value"; and they work well without any assumption of a potential generator of the data. Y. Kempner, Boris Mirkin (V), and Ilya Muchnik (V) have focused on the simplest combinatorial clustering models oriented to massive data set applications. The base of several of them are seriation procedures. They have characterized a new type of set-function (which they call quasi-convex) for which seriations become both an efficient apparatus to find optimal clusters and a good illustration of their structure. Interestingly, the seriations give a dual characterization of the data and bring additional insight into the uncovered structure of the data.

Procedures for On-Line Detection of Seismic Signals Massive data set problems arise when we monitor seismic events and attempt to determine if they are natural earthquakes or underground nuclear tests. Similar problems arise in industrial seismic data analysis related to discovering oil and estimating the amount discovered. In an attempt to establish partnerships with seismologists, DIMACS invited Oleg Kedrov, Director of the Explosion Seismology Laboratory at the Institute of Seismology in the Joint Institute of Earth Physics in the Russian Academy of Sciences, for a month-long visit. Kedrov obtained results on the automatic detection of seismic signals at three-component stations and arrays. In collaboration with L.A. Polikarpova & G.M. Steblov, Kedrov described a new algorithm of on-line signals' detection at a 3-C station based on frequency-time (f-t) and polarization analysis of seismic oscillations. Analysis of detections in the f-t domain using polarization type of the waves and rectilinearity level of the oscillations allows one to reject about 95% of false detections.

A Greedy Algorithm for Logical Data Analysis The Logical Analysis of Data (LAD) is a relatively new methodology aimed at detecting structural information about datasets. It was first proposed by Peter Hammer (PM, RU) and developed by the RUTCOR LAD research group. The main characteristic of LAD is the detection of logical patterns which distinguish observations in one class from all the other observations. For a given set of binary data, a *pattern* is a conjunction of literals which covers at least one positive point and covers no negative points. One measure of a pattern is the coverage, i.e., the number of positive points it covers. Because the problem of finding a pattern that maximizes the coverage is NP-hard, one could not expect to find a polynomial time algorithm. Lijie Shi (GS, RU), Peter Hammer (PM, RU) and Alex Kogan (PM, RU) proposed a greedy algorithm for it. The idea is to try to find a literal by conjuncting it to the partial-pattern such that the resulting new partial-pattern will have the greatest number of positive points and simultaneously decrease the number of negative points as much as possible. Shi implemented the algorithm using C++ and determined that the pattern generated by it was always close to optimal.

Pattern Generation in Logical Analysis of Data In the LAD framework (see above), we are given a data set in which some of the data are classified as positive and the rest negative. LAD is a multimoduled algorithm that seeks to ferret out the variables in the data set that are most relevant to determining the data's classifications. LAD operates in three stages. It first binarizes all numerical variables so that all data are converted to 0 or 1. Secondly, LAD processes the binary data to find a set of boolean conjunctions that have certain properties with respect to the data. These conjunctions are called *patterns*. A pattern must output a 1 for at least one (and ideally many) positive data points, and a 0 for most (ideally all) negative data points. Finally, LAD selects a subset of the patterns that collectively outputs a 1 on a sufficiently large number of positive data points. This subset, called a *theory*, is used to classify future data points whose classification is unknown. The success or value of a theory lies in the accuracy of its classification predictions. An implementation of LAD in C++ has been succesfully tested on numerous data sets. Jason Schattman (GS, RU) implemented in C++ an alternative algorithm, proposed by Alex Kogan (PM, RU), for the pattern generation module of LAD. The current LAD implementation computes a complete enumeration of all patterns up to a fixed degree. Generating patterns of high degree in this way is very expensive computationally, for the number of possible patterns is exponential in the number of binary variables. The goal of this work was to implement a pattern finding algorithm that would find high-degree, high-quality patterns in a computationally efficient way. The algorithm builds patterns one conjunct at a time, adding conjuncts until the conjunction covers sufficiently few negative points. While the pattern is being built, the next conjunct is chosen from the set of candidate conjuncts using one of several heuristics. This approach to pattern generation made possible the identification of very high degree patterns, which could not have been found using current LAD in reasonable time. On most data sets, the algorithm required no more than a few seconds to find enough patterns to cover the entire positive data set. Most importantly, Kogan and Schattman found that many of the patterns found also had very high coverage of the positive data points, which translated into high prediction accuracy on future unclassified data.

Time-series Similarity Problems and Well-separated Geometric Sets. In data mining applications, one frequently needs to determine how similar two time series are. Bela Bollobas (V), G. Das, D. Gunopulos, and H. Mannila analyzed a model of time-series similarity that allows outliers, different scaling functions, and variable sampling rates. They obtained several deterministic and randomized algorithms for computing this notion of similarity. The algorithms are based on tools from computational geometry.

5.6 Continuing Impacts of Earlier Special Years

The success of a special year or even a single workshop cannot be measured immediately upon its conclusion, nor even a year or two later. Participants in a DIMACS special year or workshop may work on a problem during an event, and then continue to work or collaborate afterwards. Results, methods, or collaborations found or initiated during a special year or workshop may lead to questions that influence the field for many years. A variety of examples continue to illustrate this point.

Minimum Spanning Trees The minimum spanning tree problem is one of the oldest problems in computer science and arises in many practical contexts ranging from design of communication and transportation networks to measuring the homogeneity of bimetallic objects. The problem is a central one in the area of graph algorithms. In work tracing its motivation to the DIMACS Special Year in Combinatorial Optimization, Bernard Chazelle (PM, PU) has improved the previous bound from $O(E.\log\log^* E)$ to $O(E\alpha\log\alpha)$, where α is the inverse of Ackermann's function. The result is the first improvement on the deterministic complexity of the problem in more than 10 years. Also of significance, the technique is fundamentally different from everything that was done before on the problem (in particular, it is the first nongreedy method).

Robot Motion Planning In robot motion planning, a subject studied in DIMACS' Special Year in Discrete and Computational Geometry, the environment in which a robot can move often corresponds to a subset of R^d, called *free space*. Typically, it is easy to move a robot along a line segment contained in the free space. This leads to the question: in which cases can one find a constant number of points (*milestones*) in the free space, such that every point in the free space can be moved along a segment to a milestone? With pre-computed paths joining the milestones, this would give a fast and simple way to find a path of the robot between any two points in the free space. In two dimensions, Jiri Matoušek (V), a key player in the Special Year in Discrete and Computational Geometry, and Gil Kalai, had earlier shown how to find the set of milestones for any simply connected free space such that every point in the free space is connected, say, to at least 1/100 of the points in the free space by a line segment. An analogous result was conjectured in more dimensions. Pavel Valtr (PD) disproved this conjecture by constructing a "bad" free space in three dimensions.

Reducing the Complexity of Reductions The notion of "completeness" is one of the most useful concepts in theoretical computer science. NP-completeness is one of the main tools for indicating that a problem is likely to be computationally intractable, and completeness for other complexity classes (such as P, PSPACE, NL, etc.) gives precise characterizations of the complexity of many important real-world computational problems. (When defining "completeness" for smaller classes, it is necessary to use a more restrictive notion of reducibility. The most common and useful of these is AC^0 reducibility – which is based on first-order logic.) One of the most well-studied questions about complete sets is a decades-old conjecture (called the Berman-Hartmanis conjecture) stating that all NP-complete sets are polynomial-time isomorphic. (In an intuitive sense, this says that all the NP-complete sets are trivial re-encodings of each other.) In recent years, it had come to be widely-believed that if one-way functions exist, then the Berman-Hartmanis conjecture is false. Other versions of the isomorphism conjecture have also been studied, using more restrictive reducibilities, such as AC^0 reducibility. In a collaboration that grew out of the Special Year in Logic and Algorithms, Manindra Agrawal (V), Eric Allender (PM, RU), Russell Impagliazzo (V), Toniann Pitassi (V), and Steven Rudich (V) showed that for essentially any complexity class C, all of the sets complete for C under AC^0 reductions are AC^0-isomorphic. Thus a version of the Berman-Hartmanis conjecture is true. One of the surprising aspects of this result is that, even though one-way functions exist in AC^0 if they exist anywhere, Agrawal, et al. nonetheless showed that the presence of these one-way functions does not stop all complete sets from being AC^0-isomorphic.

No Feasible Interpolation for TC^0-Frege Proofs Proof complexity was one of the major thrusts of the Special Year in Logic and Algorithms. Two of the main directions of research in proof complexity are: (a) to show that certain tautologies require exponential-length proofs in given proof systems. (b) to show that certain proof systems have the nice property that if a short proof exists, then a short proof can be found quickly. (If direction (a) is completely successful, it will prove that P is not equal to NP. If direction (b) is completely successful, it will prove that P is equal to NP. Partial success has been made on each front, for weak proof systems.) Maria-Luisa Bonet (PD), Toniann Pitassi (V), and Ran Raz (V), in a collaboration directly traceable to the Special Year in

Logic and Algorithms, proved that, if factoring integers is intractible (which is a commonly-used hypothesis in cryptography), then for a certain important proof system (called the TC^0 Frege proof system), goal (b) is unattainable. That is, if factoring is hard, then there are tautologies with short TC^0 Frege proofs, but such that no short proof can be found quickly. Furthermore, under the same assumption, one of the main tools that has been used in work on goal (a) cannot be applied toward TC^0 Frege proofs. (This tool is called "interpolation" and it has formed the basis of many of the successful arguments showing that tautologies require long proofs.)

Making Nondeterminism Unambiguous Klaus Reinhardt and Eric Allender (PM, RU), in work related to the Special Year in Logic and Algorithms, proved the surprising result that NL/poly = UL/poly. Nondeterministic Logspace (NL) is one of the most widely-studied complexity classes. Among the well-known complete problems for NL are the problems of computing transitive closure and finding shortest paths in graphs. The question of whether NL=L is a small-space analog of the famous NP=P question, and it is another example of the central question of the power of nondeterminism versus determinism. Unambiguous computation is a mode of computation intermediate between determinism and nondeterminism. An umambiguous machine is a nondetermistic machine, but with the restriction that, on every input, the machine either has no accepting paths, or it has exactly one path. (Thus it is not possible for there to be two "ambiguous" ways to accept the input.) Unambiguity has been studied extensively, both in relation to parsing and context-free languages, as well as in relation to one-way functions. In most settings, it has been widely conjectured that unambiguous computation is strictly less powerful than nondetermism; in some settings it has even been possible to prove that this is the case. UL (Unambiguous Logspace) contains L and is contained in NL. All complexity classes come in two flavors: uniform and nonuniform. The uniform classes are defined in terms of machines, and the nonuniform classes are defined in terms of circuits. The nonuniform analog of the P=NP question asks if P/poly is equal to NP/poly, and it is generally viewed as being of equal importance. Similarly, the L/poly versus NL/poly question has received a great deal of attention, and is of interest from the standpoint of circuit complexity. The equality of NL/poly and UL/poly provides new insight into the nature of nondeterministic computation.

Verification of Open Systems In computer system design, we distinguish between closed and open systems. A *closed system* is a system whose behavior is completely determined by the state of the system. An *open system* is a system that interacts with its environment and whose behavior depends on this interaction. The ability of temporal logics to describe an ongoing interaction of a reactive program with its environment makes them particularly appropriate for the specification of open systems. Nevertheless, model-checking algorithms used for the verification of closed systems are not appropriate for the verification of open systems. Correct verification of open systems should check the system with respect to arbitrary environments and should take into account uncertainty regarding the environment. Orna Kupferman (V) and Moshe Vardi (V) demonstrated that this is not the case with current model-checking algorithms and tools. *Module checking* is an algorithmic method developed by Kupferman and Vardi that checks, given an open system (modeled as a finite structure) and a desired requirement (specified by a temporal-logic formula), whether the open system satisfies the requirement with respect to all environments. Kupferman and Vardi started this work during the Special Year in Logic and Algorithms and continued it after the special year. This work also inspired Alur, Henzinger and Kupferman to develop a new temporal logic that is adequate for the verification of open systems.

Existence of Reduction Hierarchies Orna Kupferman (V), Robert Kurshan (PM, BL), and Mihalis Yannakakis (PM, BL) worked on a problem that is directly related to the commercial practice of model-checking, and resulted from the Special Year in Logic and Algorithms. In the automata-theoretic approach to verification, one models programs and specifications by automata on infinite words. Correctness of a program with respect to a specification can then be reduced to the *language-containment* problem. In a concurrent setting, the program is typically a parallel composition of many coordinating processes, and the language-containment problem that corresponds to verification is

$$(\dagger) \quad \mathcal{L}(P_1) \cap \mathcal{L}(P_2) \cap \cdots \cap \mathcal{L}(P_n) \subseteq \mathcal{L}(T),$$

where $P_1, P_2, \ldots, P_n$ are automata that model the underlying coordinating processes, and T is the task they should perform. In 1994, Kurshan suggested the heuristic of *Reduction Hierarchies* for circumventing the exponential blow-up introduced by conventional methods that solve the problem ($\dagger$). In the reduction-hierarchy heuristic, one solves the problem ($\dagger$) by solving a sequence of easier problems, which involve only automata of tractable sizes. Complexity-theoretic conjectures (NP $\neq$ PSPACE) imply that there are settings in which the heuristic cannot circumvent the exponential blow-up. Kupferman, Kurshan, and Yannakakis demonstrated the strength of the heuristic, studied its properties, characterized settings in which it performs effectively, and suggested a method for searching for reduction hierarchies. In particular, they proved, independently of the NP $\neq$ PSPACE question, that reduction hierarchies of tractable sizes do not always exist.

Lower Bounds for the Polynomial Calculus. Alexander Razborov (V), in work resulting from his visit during the Special Year in Logic and Algorithms and the followup workshop, "Special Year on Logic and Algorithms: One Year Later," showed that polynomial calculus proofs (sometimes also called Groebner proofs) of the pigeonhole principle PHP_n^m must have degree at least $(n/2)+1$ over any field. This is the first non-trivial lower bound on the degree of polynomial calculus proofs obtained without using unproven complexity assumptions. As a straightforward application, he also showed that (speaking somewhat loosely) low degree polynomial calculus proofs are not capable of proving $NP \not\subseteq P/poly$.

5.7 Work of Postdocs, Students, Visitors

Many of the exciting results of our postdocs, students, and visitors is not connected to a particular special year. Here we give a few examples of such results, many obtained in collaboration with our permanent members or with other postdocs, students, and visitors.

The Maximum Number of Edges in a Connected Graph with a Bounded Number of Leaves in Each Spanning Tree Guoli Ding (V) and Paul Seymour (PM, PU) have obtained a complete answer to the question: what is the maximum number of edges a connected simple graph with n vertices can have if every spanning tree has at most t leaves? The answer is $n + t(t-1)/2 - 1$ if $t > n+2$. This problem has been open for many years, since at least the early 1980's.

Progress on the Classification of the Finite Simple Groups Ronald Solomon (V) and Richard Lyons (PM, RU) continued their collaboration, aimed at a unified proof of the classification of the finite simple groups, an ambitious project initiated by the late Daniel Gorenstein (PM, RU) and Lyons in the early 1980's. Specifically, Solomon and Lyons completed the final three sections of volume number 3 in their series of AMS Surveys and Monographs. This volume contains extensive and detailed information about the subgroup structure, automorphisms, presentations and other properties of the finite simple groups. These results are both of general interest to a wide spectrum of mathematicians and of crucial importance to the inductive proof of the classification theorem which will begin in volume number 4 of the series. Solomon and Lyons also made progress on parts of volume number 4, in particular finding an improved proof of a theorem of Janko.

Quantum Bit Commitment There are two main applications in quantum cryptography: quantum bit commitment and quantum key distribution. Quantum bit commitment is important because it was shown by Yao that it is sufficiently powerful to construct most post-cold war cryptographic applications. The first quantum bit commitment protocol was proposed by Bennett and Brassard in 1984 together with a way to break it. Nevertheless, the hope remained for many years that we might be able to build an unconditionally secure quantum bit commitment protocol. In fact, in 1993 a quantum bit commitment protocol was proposed together with a security proof. This proof was generally accepted until Dominic Mayers (PD) discovered in 1995 a flaw in the protocol and in the proof. Not only did Mayers prove that this particular protocol is not unconditionally secure, but later, in part while at DIMACS, he also proved that unconditionally secure bit commitment is impossible in general. With this result, he put an end to a hope that had existed for more than 13 years.

Quantum Key Distribution. Dominic Mayers (PD) has obtained the first proof that quantum key distribution is unconditionally secure.

Graph Drawing: Geometric Graphs with No k Pairwise Crossing Edges Graph drawing is a rapidly growing subject with many practical applications in electrical engineering, manufacturing, chip design, etc. Geometric graphs are a key concept in graph drawing. A *geometric graph* is a graph drawn in the plane so that the vertex set consists of points in the plane and the edge set consists of line segments joining the vertices. Two of the basic questions about geometric graphs ask how many edges (asymptotically) can be contained in a geometric graph (a) with no k pairwise disjoint edges, (b) with no k pairwise intersecting edges, where k is a constant. With condition (a), the question was partially answered by Pach and Törőcsik in 1993, who gave an upper bound linear in n, the number of vertices. Pavel Valtr (PD) strengthened this result by showing a linear bound also when condition (a) is essentially weakened. As a consequence, he obtained an upper bound $O(n \log n)$ on the number of edges in geometric graphs satisfying condition (b). This improved the previous upper bound $O(n \log^{2k-6} n)$ proved jointly by Agarwal, Aronov, Pach, Pollack, and Sharir.

Learning Boxes in High Dimension The learnability (under various learning models) of geometric concept classes has been studied in many papers. Particular attention has been given to the case of discrete domains of points and concept classes which are defined as unions of boxes in this domain. One of the reasons that unions of boxes seem to be interesting concepts is that they naturally extend DNF formulae. Amos Beimel (PD) and Eyal Kushilevitz (V) have found exact learning algorithms that learn several classes of (discrete) boxes in $\{0, \ldots, l\}^n$. In particular they learn the class of unions of $O(\log n)$ boxes in time $\mathrm{poly}(n, \log l)$, solving an open problem. They also learn the class of unions of disjoint boxes in time $\mathrm{poly}(n, t, \log l)$, where t is the number of boxes; previously this was known only in the case where all boxes are disjoint in one of the dimensions. In particular their algorithm learns the class of decision trees over n variables, that take values in $0,\ldots,l$-1, with comparison nodes in time $\mathrm{poly}(n, t, \log l)$, where t is the number of leaves, also solving an open problem. The method also learns the class of unions of $O(1)$-degenerate boxes (that is, boxes that depend only on $O(1)$ variables) in time $\mathrm{poly}(n, t, \log l)$ (generalizing the learnability of $O(1)$-DNF and of boxes in $O(1)$ dimensions).

Learning in the Presence of Noise Scott Decatur (PD) considered the problem of concept learning in Valiant's PAC learning model in which the data used for learning is noisy. Specifically, he introduced a new model of noise called *constant-partition classification noise* (CPCN) which generalizes the standard model of classification noise to allow different examples to have different rates of random misclassification, e.g., the rate of false positives different from the rate of false negatives. He then showed how to learn in the presence of CPCN for any concept class learnable by statistical queries. This set of classes includes *every* concept class known to be learnable in the presence of standard classification noise. His model is the first such non-uniform generalization of the standard classification noise model that allows efficient learning of this wide range of concept classes.

Improvement in the Erdős-Szekeres Theorem In 1933, Esther Klein raised the following question. Is it true that for every n there is a least number $g(n)$ such that among any $g(n)$ points in general position in the plane there are always n in convex position? This question was answered in the affirmative in a classical paper of Erdős and Szekeres. In fact, they showed that

$$2^{n-2}+1 \leq g(n) \leq \binom{2n-4}{n-2}+1.$$

This theorem turned out to be the starting pont of many new developments in graph theory and combinatorial geometry, and it was also a rediscovery of Ramsey's theorem which has many analogues, extensions, and applications in several fields of mathematics. The lower bound of the Erdős-Szekeres theorem, $2^{n-2}+1$, is sharp for $n = 2, 3, 4, 5$ and has been conjectured to be sharp for all n. However, the upper bound, $\binom{2n-4}{n-2}+1 \approx c\frac{4^n}{\sqrt{n}}$, was not improved for 60 years. Recently, Chung and Graham managed to improve it by 1. Shortly after, Kleitman and Pachter showed that $g(n) \leq \binom{2n-4}{n-2}+7-2n$. Geza Tóth (PD) and Pavel Valtr (PD) obtained a further improvement, roughly by a factor of 2. They showed that any set of $\binom{2n-5}{n-2}+2$ points in general position in the plane contains n points in convex position.

Probabilistic Checkable Proofs with Zero Knowledge Joe Kilian (PM, NEC), Erez Petrank (PD) and Gabor Tardos (V) constructed probablistic checkable proofs (PCPs) with strong zero-knowledge properties. First, they constructed polynomially bounded (in size) PCP's for NP which can be checked using poly-logarithmic queries, with polynomially low error, yet are statistical zero-knowledge against an adversary that makes U arbitrary queries, where U can be set to any polynomial. Second, they constructed PCPs for $\mathsf{NEXPTIME}$ that can be checked using polynomially many queries, yet are statistically zero-knowledge against any polynomially bounded adversary. These PCPs are exponential in size and have exponentially low error. Previously, it was only known how to construct zero-knowledge PCPs with a constant error probability.

5.8 Multi-Disciplinary, Multi-Institutional Projects

DIMACS has two multi-disciplinary, multi-institutional research projects in network modeling and parallel simulations, each funded separately from its main STC award. A third such project ended last year, and we have laid the groundwork for several new projects this year. Our projects of this kind are scientifically exciting and have taken us into new and exciting research areas. The two present projects, on simulations of integrated communications systems and scalable self-organizing simulations, can be directly traced to collaborations initiated during our Special Year on Massively Parallel Computing, and are now closely related to our continuing Special Year on Networks. These projects are described below in Sections 5.8.1 and 5.8.2. The newly developing projects are described in Sections 5.8.3 to 5.8.6. The project which ended last year was the LINK project, devoted to computational support for research and education in discrete mathematics. Plans to continue the project starting in the Spring of 1998 are described in Section 5.8.7.

The participants in these projects, both DIMACS members and outsiders, have repeatedly emphasized that the projects would not have happened without the organizational infrastucture provided by DIMACS. DIMACS brings scientific expertise and experience with large, complex projects that involve a variety of institutions and disciplines, and such projects provide opportunities for students, postdocs, faculty members, and industrial researchers to share in research experiences and learn more about each others' "cultures" and ways of doing science. Importantly, DIMACS provides a neutral site for collaboration: Industrial researchers find it easier to work on a joint high risk precompetitive project under the umbrella of an academic center which enhances the opportunities for interactions.

5.8.1 Simulations of Integrated Communications Systems

Project Summary: The centerpiece of this project is the design, prototype implementation and extensive testing of a standard description methodology for various telecommunication systems, with the following properties: 1) Efficiency and flexibility; 2) Suitability for (semi)automatic scalable parallel execution of simulations; 3) Portability and independence of the parallel simulation platform. To achieve these goals, the project involves research on parallel discrete event simulation algorithms and software, and on network modeling. The effort on modeling feeds back new ideas and corrections to software design.

The projected impact of this project is best described by analogy with the extremely successful use of computer–aided design tools in the integrated circuit (VLSI) industry, where description methods combining design with simulations (expressed in description languages such as VHDL, SDL or Verilog) have allowed designs of ever higher levels of complexity. In contrast, network protocol design and engineering – which deals with systems of larger complexity than VLSI – has not yet fully adopted such tools, and not for large scale designs, primarily because simulations are: 1) Much more difficult (involving modeling of network user behavior and traffic in addition to multiple layers of protocols), and 2) Much more computationally demanding (the stochastic nature of the problems coupled with the large size of interesting systems often leads to intractably long simulations).

The observed inefficiencies and recurring failures of very large data networks such as the global Internet show the desirability of such tools. This project is funded by the National Science Foundation for a three year period over 1995 - 1998, for a total of $1.5M.

Participating Scientists: Sandeep Bhatt (PM, BC and RU), Richard Fujimoto (Georgia Tech), Albert Greenberg (PM, ATTL), Jim Kurose (University of Massachusetts), David Nicol (William and Mary / Dartmouth), Andrew Ogielski (PM, RU), Diane Souvaine (PM, RU), Don Towsley (University of Massachusetts), and Paul Wright (PM, ATTL).

Collaborations: This project brings together experts in network modeling, design and analysis, distributed algorithms, and high performance discrete-event simulation to focus on the development of high-performance simulation tools for broadband networks and wireless communication services. This project directly brings together the expertise of scientists from two of the nation's leading telecommunications companies, in partnership with leading academic sites; and includes graduate and postdoctoral training that benefits students through exposure to industrial scale problems and internships.

Project Description: Early in the project the investigators developed a prototype network-centric description framework expressed as a a meta-language extending C++, called TeD (Telecommunications Description Language). TeD has parsimoniously incorporated the core modeling concepts identified over the years of research on hardware description languages, while at the same time providing facilities for parallel simulations without involving the modelers in the internal complexities of parallel simulators. Recognizing that modeling language design is an iterative process requiring user feedback, TeD and a compiler to translate TeD descriptions to executable parallel simulation code have been delivered to model developers in continually improving six releases (from 1.0 to the latest 1.6). Curently TeD compiles to the Georgia Time Warp (GTW) parallel simulation engine. The parallel simulations are performed on multiprocessor SGI and UltraSparc machines at Bellcore, Georgia Tech and University of Massachussets.

The investigation of the TeD methodology involves complex network modeling and parallel simulations subprojects: routing in broadband ATM PNNI internetworks; reliable multicast services; multipath propagation modeling in wireless networks.

The Private Network-Network Interface (PNNI) protocol suite is an international draft standard proposed by the ATM Forum. It is a complex set of ATM routing and signaling mechanisms and represents one of the most sophisticated signaling protocols devised to date, aimed at supporting QoS-based routing claiming unprecedented levels of scalability. This project created implementations of PNNI protocols and network configurations, with nodes stochastically generating requests for connection (VC) setup and teardown to test the actual performance of the protocols. Simulations were investigated on sample network configurations representative of regional telephony backbone networks, with hundreds of complex nodes, and demonstrated an unparalleled capability of simulating in parallel over one million of VC setups per wallclock hour.

Multicast is recognized as a fundamental operation in evolving Internet and in future networks. We identified the parallelism in modeling of multicast protocols and ran simulation experiments that demonstrate that. The available parallelism is a function of the degree to which the broadcast trees for multicasts overlap spatially and temporally. We have used TeD to evaluate two different approaches to providing reliable, scalable multicast communication: The *sender-initiated* and *receiver-initiated* approach. We have successfully built TeD performance models for both a specific sender-initiated protocol and two receiver-initiated protocols. We have successfully validated the simulations predicted throughput values against those predicted by analysis for up to 500 receivers.

Our efforts in the wireless area have focused on the design of modeling software that can provide radio propagation prediction at various levels of detail. This work is being undertaken collaboratively with researchers at WINLAB at Rutgers University. The project investigated parallel implementations of radio propagation models and interference calculations, and developed tools for ray-tracing calculation of multipath radio propagation in specified geographical environments, improving various computational geometry algorithms in the process.

As a complement to this project, we organized the DIMACS Workshop on End-to-End Network Modeling and Simulations at Princeton (Oct 23-24, 1997). The Workshop brought together the leading researchers on modeling the long–range dependent traffic and user behavior, on Internet modeling, on network simulations, and wireless network modeling.

5.8.2 Scalable Self-Organizing Simulations (S3)

Project Summary: This is a project to build and distribute to the networking research community and the DoD a radically new, highly efficient network modeling and parallel simulation software package, building on the research carried out in the NSF project described in Section 5.8.1, and extending it in several important directions. The goal of this project is to achieve a generational transition in performance and manageability of simulations of telecommunication systems. This project is funded by DARPA for a three year period over 1996-1999, for a total of $3.1M.

Participating Scientists: Sandeep Bhatt (PM, BC and RU), Marina Chen and Mark Crovella (Boston University), James Cowie (Cooperating Systems Corporation), Richard Fujimoto, Kalyan Perumalla and Ellen Zegura (Georgia Institute of Technology), Don Leskiw (Ultra Corporation), David Nicol (Dartmouth), Andrew Ogielski (PM, RU).

Collaborations: The daunting complexity of modern communications technologies on the one hand, and of the large software systems required on the other, demands that a project of this sort be carried out by an interdisciplinary team of experts in parallel simulations, networking, software engineering, and defense applications. We have put together a team of scientists and engineers from six institutions to work on the project, including participants from two small companies. Academic collaborators involve in the project postdoctoral fellows and many graduate research assistants.

Project Description: Project S3 aims to provide the fastest parallel simulations, convenient modeling tools, and databases of scalable models for a range of networking problems, encompassing very different types of simulations such as wired network traffic, complex protocols stacks, packet interference in mobile wireless networks and military communications scenarios.

These goals require the common modeling and runtime environment (SSF, the **S**calable **S**elf–organizing **F**ramework) that allows *dynamic* hierarchical modeling suitable for parallel simulations, and allows one to simulate software models – without changes – both on conservatively and optimistically synchronized, radically new parallel simulators. The SSF API is the core interface common to all the individual S3 subprojects. Further speedup of simulations is achieved by researching and implementing runtime aggregation and deaggregation techniques, consisting of automated substitution of coarser (aggregate) submodels where absolute fidelity is not required.

The S3 project develops a truly process oriented, conservatively synchronized simulation engine **Nops**, the **No**rthern **p**arallel **s**imulator. Nops uses a modification of the threaded parallel runtime system, Cilk, developed at MIT. Cilk automates load-balancing through a process migration mechanism called "stealing". The first tests of Nops show excellent performance, providing a very strong proof of concept for this novel architecture of a parallel simulator.

The second parallel simulator, **MegEnta** (the name refers to a Million Entity Framework) consists of two distinct modules: TiCOON – A high-performance parallel simulation executive based on a novel micro-kernel approach which incorporates a Time Control mechanism that can simultaneously support different synchronization schemes; and HOPPERS – A novel ultra light-weight process abstraction which supports shared-memory semantics. Multiple instances of Nops and MegEnta parallel simulators can be linked over the network to create distributed federations of simulators.

Usability of parallel network simulators is a major goal of the S3 Project. Usability emerges from the separation of modeling work from the internal complexities of parallel simulators, and from the availability of user–friendly tools: for model browsing and assembly (CAD), for model debugging, for parallel performance monitoring and load management. The modeling environment and tools should be used and extended by end-users who are not experts in parallel discrete-event simulation, and whose primary goal is to describe and model their particular application domain, not rebuild a customized simulator for it.

SSF owes its existence to TeD, the pioneering modeling API prototype that made the critical leap from low-level, event-oriented simulation to high-level, process-oriented hierarchical modeling. The first implementations of TeD, atop Georgia Time Warp, and using CPP as the external language, have successfully shown that modelers can take advantage of parallel resources to run very large, complex simulations. TeD, in turn, has been designed to concisely express the key modeling concepts identified over the years of development of various hardware design languages, such as VHDL, SDL or Verilog.

The project stresses the necessity of verification of software design and implementation in realistically large and detailed modeling projects utilizing the modeling environment and simulators built by the S3 partners. This *co–design* process, involving networking engineers and graduate students, provides real life tests, exposes weaknesses, and feeds back corrections and new ideas to simulator and tool developers. The leading theme is versatility; thus we cover the extremes of the modeling problems: wireless networks, wired ATM internetworks, the global Internet, and the battlefield scenario–driven military communications models. In this way we gain confidence in the power of our approach to usable, versatile, scalable high performance simulations of telecommunications systems.

5.8.3 Environmental Monitoring

As part of our pre-special year on Massive Data Sets, we developed partnerships with a variety of individuals and organizations interested in environmental monitoring and cancer risk evaluation. The group is interested in the development of efficient mathematical methods of pattern extraction from complex empirical data (clustering); the development of methods that start with data about individual entities and produce aggregative indices; and the development of methods to work with heterogeneous data sources gathered under different spatial resolutions, time scales, and techniques.

Participants in the group include Teuvo M. Airola, Director of the Grant F. Walton Center for Remote Sensing and Spatial Analysis and Chair of the Dept. of Environmental Resources at Rutgers University; V. Vapnik of AT&T Labs; Helene Kulsrud of the Institute for Defense Analyses-Center for Communications Research in Princeton, NJ; David Ozonoff, Chair of the Department of Environmental Health at Boston University School of Public Health, Medical Director of the Boston Environmental Hazards Center (BEHC), and Director of the Superfund Center at Boston University; Daniel Wartenberg of the Dept. of Environmental and Community Medicine at Robert Wood Johnson Medical School/University of Medicine and Dentistry of NJ; Boris Klebansky, President of BK Computer Corp.; and DIMACS members/long-term visitors Jon Bentley (PM, BL), Joan Feigenbaum (PM, ATTL), Boris Mirkin (V), Ilya Muchnik (V), Leonid Gurvits (V), Fred Roberts (PM, RU), and Bob Sedgewick (PM, PU).

5.8.4 Massive Data Sets in the Biological Sciences

Also as part of our Massive Data Sets pre-special year and as an outgrowth of our Special Year on Mathematical Support for Molecular Biology, we put together a group of researchers interested in massive data set problems in the biological sciences. While biological databases are not traditionally thought of as "massive" – for example, they are quite small in comparison with the records of all long distance phone calls handled by AT&T in a month – those who work with them share some of the same problems as those who work with other very large databases. Furthermore, many types of biological data exhibit an "Inverse Moore's Law." Moore's Law is the observation that computational power has doubled every 18 months for the last few decades. Since some types of biological data sets double at a rate faster than 18 months, any computational problems we have now in dealing with the data will only get worse. Among the specific research issues targeted by this group are the need for homology similarity estimates between the two largest RNA polymerase polypeptides beta and beta' (these homologs are very large for bacteria, achaea, and eukaryotes); the desire to do a correlation analysis among sequence and structural features along the whole genome based on NMR data; the development of a classification of bacterial promoters; and the uncovering of evolutionary phylogeny structure for bacteria based on the genome duplication model developed in recent years by DIMACS reseachers.

Participants in the group include Joachim Messing, Director of the Waksman Institute for Molecular Genetics at Rutgers; Helen Berman, director of the Nucleic Acids Database Project in the Center for Molecular Biophysics and Biophysical Chemistry at Rutgers; Richard Ebright of the Waksman Institute; Steve Anderson of the Molecular Biology and Biochemistry Department at Rutgers and a member of the Center for Advanced Biotechnology and Medicine (CABM), a joint project of Rutgers and the University of Medicine and Dentistry of NJ; Gaetano Montelione of the Molecular Biology and Biochemistry Department at Rutgers and of CABM; Casimir Kulikowski, the coordinator of the Rutgers Laboratory in Structural Bioinformatics; Julio Collado-Vides, Director of the Center for Molecular Biology at the State University in Cuernavaca, Mexico; Roderic Guigo of the University of Barcelona; and DIMACS members and long-term visitors Fred Roberts (PM, RU), Boris Mirkin (V), Ilya Muchnik (V), Martin Farach (PM, RU), Bob Sedgewick (PM, PU), Jon Bentley (PM, BL), and Joan Feigenbaum (PM, ATTL).

5.8.5 Collaborations between Mathematical and Biological Scientists on Problems of Molecular Biology

As an outgrowth of the Special Year on Mathematical Support for Molecular Biology, we have taken on two major partners, the Biomathematical Sciences Department at the Mt. Sinai School of Medicine and the Center for Molecular Biophysics and Biophysical Chemistry (MBBC) which is joint between Rutgers and the University of Medicine and Dentistry of NJ. We are aiming at fostering serious research interactions between mathematical and biological scientists on such areas as stress-induced denaturation in DNA; algorithms for computer analysis of DNA sequences; structure analysis and prediction for nucleic acids; lysomal biology and diseases in which the goal is to determine the molecular and cell biology of human lysomal enzymes in health and disease; algorithmic aspects of phylogeny reconstruction; coalescent models and the estimation of historical population genetic parameters; the influence of chemical architecture on the conformation, properties, and interactions of nucleic acids; protein folding and structural prediction; consensus methods in molecular biology; and modeling of ultraselfish genes in mice.

Participants in the group are Craig Benham, Mt. Sinai Biomath; Gary Benson, Mt. Sinai Biomath; Helen Berman, MBBC; John Blaho, Mt. Sinai Microbiology; Robert Desnick, Mt. Sinai, Head of Human Genetics; Jody Hey, Rutgers Department of Evolutionary Genetics; Joseph Naus, Rutgers Statistics; Wilma Olson, Director of MBBC; Shalom Rackovsky, Mt. Sinai Biomath; Lee Silver, Princeton Molecular Biology; and DIMACS members and visitors Farid Alizadeh (PM, RU), Dannie Durand (V), Martin Farach (PM, RU), Richard Lipton (PM, PU), Fred Roberts (PM, RU), Larry Shepp (PM, RU), Peter Winkler (PM, BL), Andy Yao (PM, PU).

5.8.6 Visualization

As an outgrowth of DIMACS activities in massive data sets and graph drawing, we have started up a group dealing with problems of visualization. Visualization is a field devoted to helping us understand complex systems by exploiting technology and human visual psychology. Visualization transforms the symbolic into the geometric, enabling researchers to "see the unseen." It is a tool both for interpreting image data fed into a computer and for generating images from complex, multi-dimensional data sets. The goal of visualization is to leverage existing scientific methods by providing new insight through visual methods. The group is concerned with network visualization, visualization of tables and matrices, and selections of "representatives" in image data. How one visually presents a network can make a large difference in the ability to understand its workings, manipulate it, interdict it, or improve its functioning. The diverse applications of networks have led to a variety of computer tools and mathematical methods for network visualization. We will be interested in developing algorithms for "efficiently" presenting a network. Data of all kinds is often stored or presented in tables or matrices. In a typical application, the data could represent similarity judgments or proximity estimates. The tables often have millions of entries. Since humans can grasp information and relationships visually more rapidly than by other means, an important question is whether it is possible to present the data from this table visually in such a way that will help to spot patterns or will make it easier to analyze it. We can envision the storage of libraries of visually represented data, for example photographs, health information (with hundreds of thousands of individual items of height, blood pressure, etc.), and explanatory diagrams. One would like to automatically select sets of determined sizes that "represent" the sorts of things in which we are interested. We would like to develop methods to generate automatic catalogs of representative images.

Participants in this group include Sven Dickinson from Rutgers Computer Science, Manesh Parashar, Deborah Silver, and Norman Zabusky from Rutgers College of Engineering, and DIMACS members and long-term visitors James Abello (PM, ATTL), Sandeep Bhatt (PM, BC), Boris Mirkin (V), Ilya Muchnik (V), and Andy Ogielski (PM, RU), and we are in the process of adding psychologists from Rutgers and the University of Illinois.

5.8.7 The LINK Project

LINK is a software system designed to be a general-purpose environment for experimentation with discrete mathematics. The environment is friendly enough so that non-technical users will be able to visualize problems easily, yet powerful enough for computer scientists and mathematicians to express complicated computations. The system is designed to support educational pursuits, to be used as a research tool, and to be useful in solving real-life problems. LINK was funded by an NSF grant to DIMACS for a three year period over 1993 to 1996, and it also received support from the Office of Naval Research and Los Alamos National Laboratory. Participating scientists in the first stage of LINK were Jonathan Berry (PD, Elon College), Nathaniel Dean (PM, BL), Mark Goldberg (Rensselaer Polytechnic Institute), Greg Shannon (Spanning Tree Technologies), and Steve Skiena (SUNY Stony Brook).

LINK will play a central role in two summer programs at DIMACS in the summer of 1998. Berry and Dean will be two of the principal lecturers in the summer conference aimed at "reconnecting" instructors at two and four year colleges to the mathematical sciences enterprise (see Section 6.2.2). Berry and Dean will lecture on recent research in network visualization, and will introduce LINK as a hands-on tool for participants to experiment with research problems and to help develop educational materials to be brought into the classroom. Berry will also develop the next stage of LINK, for the PC, so that it can be used in the computer labs at DREI'98 as a way for high school teachers to master concepts in graph theory. (See Section 6.5 for a description of the DREI program.) With DIMACS support, Berry will be working on the PC version of LINK during the Spring of 1998.

6 Education Programs

DIMACS' educational programs divide into two kinds, K-12 programs and programs for undergraduates. (Our graduate student and postdoc programs are also educational in many aspects, but are presented in this report as part of our description of the research activities of the center.)

DIMACS' two most visible K-12 programs, the Leadership Program in Discrete Mathematics and the Young Scholars Program, are described in Section 6.1. Other K-12 programs for teachers include the "Workshops in Your District Program," through which trained participants in the Leadership Program have gone to districts to provide workshops to teachers; we have reached thousands of teachers in this way. We also have a newsletter for high school teachers, "In Discrete Math," that is circulated free twice a year to approximately 3500 teachers. DIMACS has played a central role in the adoption of a new Curriculum Framework in Mathematics in New Jersey, cited by the governor in her 1997 "State of the State" message, and in the adoption of a discrete mathematics component by the Manhattan public schools. A DIMACS volume, *Discrete Mathematics in the Schools*, was published this year, and it has been called "truly a great resource and a source of inspiration" (Viera Proulx, Northeastern University).

DIMACS runs a variety of undergraduate programs and has launched three major new initiatives in the undergraduate arena: A Modules series; a program to "Reconnect" two and four year college faculty to the mathematical sciences enterprise through exposure to current research topics in DM/TCS relevant to the classroom; and a summer program at Evergreen State College aimed at presenting to computer science faculty recent results of relevance to the undergraduate curriculum. DIMACS' new undergraduate programs are described in Section 6.2. Its successful and growing program of research experiences for undergraduates is described in Section 6.3.

The center has made major strides toward integration of its research and education programs. Our efforts at integration are described in Section 6.4, and in particular our major program, the DIMACS Research and Education Institute (DREI), which has integration as its primary theme, is described in Section 6.5.

6.1 K-12 Programs

6.1.1 Leadership Program in Discrete Mathematics

The Leadership Program in Discrete Mathematics (**LP**) has evolved through three phases: Phase I for high school teachers (1989-1991), Phase II for high school and middle school teachers (1992-1994), and now Phase III for K-8 teachers (1995-1998). In the present phase, under an extra grant from NSF and support from DIMACS and from its STC grant, teachers attend a two-week program during the summer, attend four follow-up days (six are available) during the year, and then return for a week the following summer.

The content material for the first summer (two weeks) includes applications of graphs and patterns in numbers and geometry. Participants begin by coloring maps and end by generating spirolaterals and fractals. These topics are presented during morning workshops lasting two hours, and consist of instruction together with hands-on activities and group work designed to engage the teachers in the material and in the problem-solving process. Also mentioned during these workshops, and emphasized in other sessions later in the day, are ways that the material can be brought into the classroom.

The **LP** is successful if the teachers return to their schools empowered to become "discrete math leaders." During the program it is emphasized that they have become experts of a sort and that they are encouraged and expected to try this material in their classrooms, and to give workshops for other teachers in their school or district, telling them about discrete math. Thus the **LP** includes special sessions on problem solving, assessment of mathematical learning, an overview of discrete mathematics (so that they can successfully answer the common question "what *is* discrete math?") and leadership issues. It is fair to say that the **LP** is as leadership-oriented as content-oriented. A typical year's cohort in the K-8 **LP** program gives workshops attended by at least 5000 teachers.

Study group sessions follow the morning workshop. During this time, the teachers are broken into grade-heterogeneous groups to work on problems related to the morning workshops. These problems will be discussed further during the afternoon consolidation sessions and the following morning, before the next workshop. Afternoons at the **LP** find the teachers in grade-homogeneous groups exchanging ideas for implementing the material in the classroom and incorporating it into their already full syllabi. Teachers are also exposed to e-mail (and everyone is given a DIMACS account), the Internet and LOGO–Turtle Math.

The 1997 **LP** involved two two-week residential programs (both at Rutgers) and two twelve-day commuter programs, one at Rutgers (as in 1995 and 1996) and one in Scottsdale, AZ. Those programs represent 125 teachers participating in the **LP** for the first time. In addition, there were five one-week second-summer follow-up programs for the 1996 participants; of these, two were residential programs at Rutgers, and three were commuter programs, at Rutgers, Virginia Beach, VA, and Warwick, RI. Altogether, 104 of the 161 teachers in the 1996 summer programs attended one of these follow-up programs. In addition, 24 of the 1995 and 1996 participants attended a one-week "workshop workshop" program, at which they developed workshops that would be delivered to K-8 teachers in their schools as part of our "Workshops in Your District" outreach program. We also conducted a one-week program for the high school and middle school "veterans" from the 1989-1994 programs, which was attended by 24 teachers, and a two-day "crash course" in discrete mathematics, which was attended by 30 high school teachers. Thus, we have been able to continue our high school teacher programs even though we did not have a specific extra grant to fund these activities. Both of these programs will continue in 1998.

In order to ensure that our off-site programs are comparable in quality to our programs at Rutgers, we have developed high-quality materials which were used at all different sites by a number of different workshop leaders. These materials include detailed instructor's notes, overhead transparencies, handouts, homework problems, and supplementary resource materials. All new instructors are provided with some training in the use of these materials and in the philosophy of the **LP**. Moreover, in order to compensate for the extra time that is available to participants in the resident programs, all of the 1997 commuter programs were extended by 2 days, so that they involved 12 days of activities, as noted above.

The 1998 schedule will involve two two-week residential programs at Rutgers and four twelve-day commuter programs, all of which will be off-site weeks. Commuter programs have been scheduled for Scottsdale, AZ, Louisville, KY, New York City, and Vermont. The extra commuter program will compensate for a third commuter program that was scheduled for 1997 but for logistical reasons did not take place. We anticipate a total of about 210 teachers in these six programs.

In the summer of 1998, we will also have the returning teachers from the summer of 1997, and veterans from all programs before that. There will also be another two-day "crash course" for high school teachers.

The materials that have been developed for the leadership program workshops described above are sufficiently detailed and have been sufficiently field-tested and evaluated, that it will be possible for other mathematicians to use them as the basis for independently-conducted ("franchised") parallel high-quality institutes in 1999, after suitable training provided by the leadership program.

The **LP** is directed by Joseph Rosenstein (PM, RU).

6.1.2 Young Scholars Program in Discrete Mathematics

The Young Scholars Program in Discrete Mathematics **YSP** is a four-week residential program for 40 mathematically talented students entering grades 11-12, with four follow-up sessions during the year. Week-long workshops are given by college and university faculty on topics ranging from number theory to fractals, designed to excite and challenge the students, motivate them to continue their study of mathematics, and consider careers in the mathematical sciences. Rutgers undergraduates and graduate students serve as group leaders and mentors. Maintaining diversity is a priority: each year the program has maintained a 50/50 balance between females and males, and approximately 20% of the students are minorities.

An important feature of the **YSP** (since 1994) is the research component in which DIMACS-affiliated researchers develop projects and work on them with groups of 4-6 students during the last three weeks of the program. Eighteen researchers have now participated, with project topics including secret communication, scheduling, polyomino tilings, and text compression. Both students and researchers have found the projects valuable. The researchers for the 1997 program were Las Goonetilleke ("Simulating the Truth"), Robert Hochberg (PM, University of Delaware, "Dominos Vobiscum"), Paul O'Donnell (GS, RU, "Coloring the Plane"), Gretchen Osteimer (RU, "A Prime Forgery"), George Popel (Middlesex County College, "Nones and Ones"), and Devan Putnam (RU, "Stable Marriages"). The researchers for the 1996 program were Adam Buchsbaum (PM, ATTL, "Paying the Piper—Balancing the Costs of Different Tasks"), Matt Franklin (PM, ATTL, "Search Games on Graphs"), Deborah Franzblau (PM, RU, "Sticks and Stones or Knucklebones: How do You Play that Game?"), Joseph Gerver (Rutgers Camden, "Soap Bubbles and Minimal Surfaces"), Robert Hochberg (PM, RU, "Ominous Ominos"), Stephen Penrice (PD, "Can you Design a Fair Draft Lottery?"), Nick Reingold (PM, ATTL, "Dotland"), and Nina Shteingold (GS, RU "Trinkets and More").

For the past three years there has been a significant technology component to the Young Scholars Program. Each year an entire week is devoted to doing mathematics with computers and graphing calculators, together with an exciting "LEGO LOGO" project. Janice Kowalczyk (an educator from Rhode Island, now working with the Leadership Program full-time) runs three days of workshops on programming in the LOGO language and how the interface works with LEGO motors and products, at the end of which the students have a competition to design and build a robot to perform some specified task. This project integrates mathematics, programming, design and construction.

The program has run since 1990 at Rutgers University with extra funding from the NSF as well as funding from DIMACS and from its STC grant. Although the NSF has phased out its Young Scholars Program, the DIMACS Young Scholars Program in Discrete Mathematics will continue, with funding from DIMACS, funding from corporate sources, and increased fees from participants. The program has been co-directed by Robert B. Davis (RU), Michael O'Nan (PM, RU) and Joseph G. Rosenstein (PM, RU).

6.2 Undergraduate Education Initiatives

Starting in the Fall of 1995, we totally reassessed the goals and purposes of the center, and in February 1996 decided that DIMACS should play much more of a national leadership role relative to the inclusion of DM/TCS in the undergraduate curriculum and in introducing undergraduates and college teachers to the topics of DM/TCS and their applications and began a serious assessment of alternative new programs. This assessment led us to three major new programs that we implemented this year. These are described in the following three subsections. While we were carrying out this assessment and in the time since, we also continued a number of other programs. Here are some of the things that we did:

- We continued our successful REU program and made some changes in it. See Section 6.3.

- We continued to host visitors from such four year colleges as Drew, Holy Cross, Lafayette, and Wellesley.
- We involved in a leadership role at DIMACS faculty from many two and four year colleges, including Drew, Swarthmore, Wheaton, Middlesex County College, Union County College, and County College of Morris.
- We discussed programs with the NJ Chapter of the Mathematical Association of America and the Mathematical Association of Two Year Colleges of New Jersey, and partnered with the latter organization in developing our undergraduate initiatives for two year colleges.
- We organized a "working group" on curriculum issues in logic and CS under the direction of Moshe Vardi of Rice University. This group has organized a symposium on logic and CS at the February 1998 meeting of the ACM Special Interest Group in Computer Science Education (SIGCSE).

6.2.1 DIMACS Module Series

DIMACS has launched a new Module Series and work on the initial modules in the series is progressing. The modules in the series will be self-contained text and problem material that can be used for one or more class meetings in an undergraduate course or, in some cases, in a high school course. The idea is to make recent advances in DM/TCS known speedily to undergraduates and undergraduate instructors and to some extent high school students and teachers. The modules will state their topics carefully and pinpoint particular theorems and lots of examples (perhaps of special cases) for which proofs/algorithms can be given and exercises written that undergraduates can solve. There will also be pointers to more extensive results and open questions. The modules will be somewhat like the UMAP Modules published over the last 20 years by COMAP, the Consortium for Mathematics and its Applications Project, and widely used, but DIMACS modules will usually be related to the special research foci and the many outreach activities at the center. We expect that DIMACS modules will take advantage of advances in technology. They will all be available electronically, and some may be interactive. We are planning to bundle modules into DIMACS volumes, to be published in the DIMACS book series, published by the American Mathematical Society. AMS is amenable to co-publishing agreements with other publishers such as NCTM and MAA in order to maximize exposure of a volume and it is likely that the volumes of DIMACS modules will be co-published. Modules will be available electronically without charge, and in hard copy at cost. Modules will be prepared, tested, critiqued, and improved in a series of steps that is being developed by the Editorial Board. This will involve classroom testing and several rounds of feedback.

The Module Series has a chief editor, five associate editors, and an Editorial Board. The Editorial Board consists of leaders in DM/TCS research with special interest in undergraduate education or high school education, leaders from the two year college community, and experts on bringing DM/TCS into the schools.

DIMACS Module Series Editorial Board

Editor in Chief

Stephen Maurer, Swarthmore College

Associate Editors

- Susanna Epp, DePaul University

6.2.2 The "Reconnect" Project

DIMACS has initiated a project that seeks to "reconnect" to the mathematical sciences enterprise two and four year college faculty who don't have the time to keep up with research developments. It will expose them to current research topics in discrete mathematics and theoretical computer science that are relevant to their teaching by involving them in a research center where much of the relevant research is being carried out. We aim to enhance the ability of teachers to transform their classrooms into places that connect up to modern uses of mathematics and computer science and to help them produce classroom materials that reflect current research. We will do this through summer conferences and a sequence of two-day conferences. The summer conferences, national in scope and directed at two- and four-year college faculty with some prior exposure to discrete mathematics and theoretical computer science, will present recent research results in topics such as computational molecular biology, network visualization, clustering, and visibility in geometry, and divide the participants into writing groups. The two-day conferences, regional in scope and directed at two year college faculty with little prior exposure to discrete mathematics and theoretical computer science, will present an introduction to these fields, with connection to topics of current research interest, and will also involve the participants in writing materials in a six month period between two two-day conferences. Participants in both the two-day and two-week conferences will be expected to use the materials they have written as vehicles to bring discrete mathematics and theoretical computer science into their classrooms. Participants will be encouraged to make their materials available to a broader audience through developing and publishing them in the DIMACS Module Series. Participants will also be encouraged to become involved in other DIMACS activities, such as workshops and the DIMACS Research and Education Institute (DREI).

To launch this project, we have held a series of meetings with faculty involved in teaching of undergraduates, and have had extensive contacts with such faculty members in New Jersey, in particular from the two year college community. Our project has the cooperation and endorsement of the Mathematical Association of Two Year Colleges of New Jersey. We have been told that the Undergraduate Faculty Enhancement program at NSF will support our Reconnect project. The two day conferences will begin in the Spring of 1998, and the first summer conference will be in August 1998. That conference will have two themes, computational biology and network visualization.

Two Day Conference Organizing Committee

- Joseph Rosenstein, Rutgers University, Organizer
- Jean Lane, Union County College, Co-Organizer

Summer Conference Organizing Committee

- Rochelle Leibowitz, Wheaton College, organizer
- Reggie Luke, Middlesex County College, advisor
- Theresa Michnowitz, Jersey City State College, advisor

"Reconnect" Project Advisory Committee

- Susanna Epp, DePaul University
- Stephen Greenfield, Rutgers University
- Robert Hochberg, University of Delaware

- Jean Lane, Union County College
- Rochelle Leibowitz, Wheaton College
- Linda Lesniak, Drew University
- Reggie Luke, Middlesex County College
- Joseph Malkevitch, York College (CUNY)
- Stephen B. Maurer, Swarthmore College
- Marilyn Mays, North Lake College
- Catherine McGeoch, Amherst College
- Theresa Michnowicz, Jersey City State College
- Charles Miller, Camden County College
- Ruth O'Dell, County College of Morris
- Warren Page, New York City Technical College (CUNY)
- George Popel, Middlesex County College
- Fred Roberts, Rutgers University, chair
- Joseph Rosenstein, Rutgers University
- Michael Saks, Rutgers University
- Charles Sims, Rutgers University
- William Steiger, Rutgers University
- Kenneth Wolff, Montclair State University

6.2.3 Evergreen College Program

DIMACS' first effort at a program with a theme similar to its new "Reconnect" project was its participation in a program in the summer of 1997 aimed at introducing college faculty to research topics they can bring into the undergraduate classroom. The program was run, with NSF support, at Evergreen State College in Washington State during the week of July 7-11, and featured four topics: Software engineering, functional programming, neural networks, and computational geometry. The latter topic was added to the program by DIMACS at the request of the Evergreen organizers, and it was organized by Walter Meyer of Adelphi University, advised by Bernard Chazelle (PM, PU), the co-director of DIMACS.

The program on computational geometry featured presentations on graph drawing algorithms by Bill Lenhart, visibility problems by John Hershberger, and Voronoi diagrams by Scot Drysdale. It was aimed at those who teach computer graphics, computer vision, robotics, and theoretical computer science (algorithms and data structures).

Because of the success of the Evergreen College program, the Evergreen organizers have urged DIMACS to run a similar program in New Jersey.

6.3 REU Program

DIMACS has run a successful NSF Research Experiences for Undergraduates (REU) program since 1993. The program introduces mathematically talented students to research in DIMACS-related fields under the guidance of research mentors, both to discover new results and provide a basis for making career and graduate school decisions. Students come to work one-on-one with senior DIMACS researchers for eight weeks during the summer, and enjoy the stimulating and supportive environment at DIMACS. Students report on their projects at weekly student-faculty lunches, and write a short report at the end of the summer. Most of the students discover original results, and some are impressive. For example, working in the area of protein folding, Peter Coles managed to improve upon the Batzoglou-Decatur algorithm for protein folding on a triangular lattice. Preliminary results show that the algorithm has a worst case yield of just under 61%. This is not significantly higher than 60%, but it does show that 60% is not a fundamental limiting ratio for these types of layering algorithms, as some had speculated.

During the summer of 1997, we improved upon our series of popular evening lectures, this year bringing in five distinguished scientists: Jozsef Beck (PM, RU), Joel Spencer (New York University), Ron Graham (PM, ATTL), Peter Winkler (PM, BL) and Shimon Even (Technion). Each of these talks was followed by an informal "meet the speaker" dinner during which time the students (and other attendees) could interact with the presenter. We also housed our students on-campus for the first time this summer. The increased camaraderie and opportunities for collaboration were apparent.

The 1997 DIMACS REU program hosted 11 students, our largest program to date. They came from 10 different institutions.

Student	Mentor	Project area	Home Institution
Peter Coles	Scott Decatur (PD)	Protein Folding	Princeton
Bryan Cooley	Andy Ogielski(PM, RU)	Simulation of Wireless Networks	Nebraska
Haggai Elitzur	Charles Sims(PM, RU)	Computational Group Theory	Berkeley
Lynette Hock	Wilma Olson(RU, Chemistry)	DNA Structure	Nebraska
Andrew Hutchings	Farid Alizadeh(PM, RU)	Semidefinite Programming	Harvey Mudd
Gabriel Loh	Andy Ogielski(PM, RU)	Simulation of Wireless Networks	Cooper Union
Alantha Newman	Mona Singh(PD)	Protein Structure Prediction	Columbia
Vinhthuy Phan	Diane Souvaine(PM, RU)	Wireless Communication	Brandeis
Jing Qi	Martin Farach(PM, RU)	Protein Structure Prediction	Rutgers
David Rosenberg	Sven Dickinson(RU)	Vision Based Robot Navigation	Yale
Mike Shubov	Adi Ben-Israel(PM, RU)	Decision Making	Texas Tech.

This year, we expect 8 of the 11 students to continue working on their projects during the school year, either with their advisor or with an advisor at their own universities. We have asked our students to put their projects up on our REU web pages, and they can be found at `http://dimacs.rutgers.edu/REU/index.html`.

Some of the students have begun papers describing their summer research, and most have given talks or informal presentations at their schools or at undergraduate research conferences.

6.4 Integration of Research and Education

One of DIMACS' goals is to integrate our research and education programs. We believe that involving our permanent members, postdocs, and graduate students in the educational programs of DIMACS can help instill in them a career-long commitment to helping in the education of young people, and that involving teachers with our research programs can help make education in the mathematical sciences more relevant and alive, and create long-term partnerships between researchers and educators. Our education and research programs are linked through the incorporation of educational outreach activities in the planning and implementation of special years (we are now asking special year organizers to plan on educational components for their years if possible) and the direct participation of special year researchers in educational outreach activities. Our new undergraduate initiatives described in Section 6.2, the Module series, the "Reconnect" project, and our Evergreen College program, are all aimed at bringing current research into the classroom. We are asking each of our postdocs to participate in either an outreach activity or an educational activity while at DIMACS. In the past, and in particular in the past year, DIMACS postdocs, visitors, permanent members, and graduate students have participated in educational activities by acting as mentors in our research experiences for undergraduate programs and in our Young Scholars Program for high school students, by giving lectures or teaching courses in our Leadership Program for precollege teachers, and by supervising undergraduate theses. Participants in these programs have also included some of the most senior DIMACS leaders from all five institutions. A small portion of this involvement is described in Sections 6.1 and 6.3.

In the summer of 1996, we inaugurated a four year DIMACS Research and Education Institute (DREI), under a $1.9M grant from the NSF program of Regional Institutes in the Mathematical Sciences, RIMS. The program is aimed at "integrating" researchers, postdocs, students, and teachers. Our program, in collaboration with the Institute for Advanced Study/Park City Mathematics Institute, is described in Section 6.5.

6.5 DIMACS Research and Education Institute

The DIMACS Research and Education Institute (DREI) is aimed at integrating research and education. DREI has many purposes. It is to be a research meeting, gathering people from the "cutting edge" of its disciplines in academia and industrial research, both to present their results and to foster further advances. It should also educate high school faculty members in some aspects of the topic of emphasis, so they will be able to inform some of their students about the subject matter, and at the same time reinvigorate their teaching styles by discussions of pedagogical issues with other participants. Most particularly, DREI should encourage serious intellectual interaction among all the participants. Teachers can get an understanding of current research topics and their practical and social significance, and of what it is like to pursue research. Researchers can learn more about the educational enterprise and, hopefully, become more interested in education at all levels.

DREI is organized around a three week summer program, with research workshops and a high school teachers program. There are also followup activities later in the year. The theme changes each year. The topic for DREI '97 was "Cryptography and Network Security". The "residential" aspect met during three weeks (July 28 through August 15, 1997) at Rutgers. Followup activities, principally to encourage and support efforts by high school faculty members involved to introduce new content and methods in their courses, were scheduled to occur three times during the 1997-98 academic year. Followup activities for DREI '96, which had a theme of "Computational Geometry," were held on four weekends in 1996-97.

The convenor of DREI '97 was Stephen Greenfield (RU). Joan Feigenbaum (PM, ATTL) was the research program director, and organized the three research workshops. The three weeks of the research program were devoted to the topics: "Underlying Mathematics," organized by Eric Bach of the University of Wisconsin (Madison); "Computer Science," organized by Steven Rudich of Carnegie Mellon University; and "Network Security," organized by Ross Anderson of Cambridge

University. The education program chair and chief speaker for the high school teachers program was Steven Rudich. During week 2, several of the lectures to high school faculty were given by Joe Kilian (PM, NEC). The education efforts were aided by the presence during two of the three weeks of Henry Kepner of the University of Wisconsin (Milwaukee), a math education faculty member who is very experienced in coping with the challenges of new pedagogy and content in the high school math classroom. High school lead teachers who organized small group sessions and helped with homework were Charles Biehl (The Charter School, Wilmington, Delaware), Brother Patrick Carney (The Montclair Kimberly Academy, New Jersey), and Joan Vital (Norton Public Schools, Massachusetts). The first two of these lead teachers had assisted in DREI '96 and several other activities involving college/high school instructional interaction, and the third has also had a long involvement with similarly successful efforts.

Approximately 200 people were formally registered attendees at DREI '97, including 30 high school faculty members, and probably an additional 50 people attended various DREI '97 activities. There was a remarkably high percentage of industrial research participants, reflecting the intensity and importance of innovation in the area of cryptography to industry. In the morning, there were research talks and the high school program lectures by Rudich. There were plenary talks on technical material at 11 every day. These talks were aimed at an audience with extremely heterogeneous backgrounds, ranging from high school faculty members whose college studies in mathematics and computer science may have been many years past, to people participating in current research efforts. The great majority of these talks were understood in large part by all of the audience. One of this year's successful innovations occurred after lunch almost every day: the plenary speakers each spent at least half an hour discussing the contents of their talks, their backgrounds, and their ambitions in "private" with high school faculty members. This effort seemed very popular. Those speakers who concentrated on the less "friendly" aspects of the subject were often very candid and understandable in these smaller question-and-answer sessions. They'd often discuss *why* they were looking at a problem, and what its connections were with other efforts. Fairly frequently their comments about their professional work were supported by personal references which seemed to make the subject come alive much more for the high school faculty members. Indeed, at times the audience for these post-plenary sessions was supplemented by a number of researchers who wanted to get similar insights but were perhaps a bit reluctant to request them at a formal plenary talk.

Afternoons were spent in various ways. Research talks were given among the research community. The high school faculty members used computers to study topics related to DREI; participated in small groups ("homework sessions") to reinforce the material of the lectures; and discussed the feasibility of using the DREI material in the classroom. A number of researchers who had not given plenary talks asked for time with the education group. They used this time to explore ideas which they felt could lead to interesting activities in the high school classroom.

During the week, most evenings were occupied by DREI activities. Evenings were frequently about public policy issues such as key escrow, copyright, etc. Controversial topics related to cryptography are now in the news quite a lot, so it was rather common to have vigorous debates which included world-renowned experts as well as novices. Some of these interactions were truly fulfilling of DREI's aims. We had several presentations about educational issues that led to rather heated discussions – an encouraging sign of involvement.

Three followup sessions were planned, and one has already taken place (November 1). About 25 of the 30 high school faculty members returned to DIMACS for this meeting. The first followup meeting featured two speakers from SWRIMS, the Southwest Regional Institute in the Mathematical Sciences, an NSF-financed project with aims similar to those of DREI.

Plans are well underway for DREI '98, which will revolve around the theme "Graph Theory." Linda Lesniak of Drew University is serving as convenor and research program director, and Glenn

Hurlbert of Arizona State University as education program coordinator and primary speaker to the high school teacher group.

DREI'96 was extensively evaluated by professional evaluators Norman Webb and Kathryn Mertz. Their detailed report concluded: "Participants were highly satisifed with their institute experience. ... Overall, the institute met its goals and served for these separate groups in the mathematics community to engage each other and to learn from each other." We are awaiting a similar detailed evaluation of DREI'97.

A Management

- Director: Fred S. Roberts, Rutgers University
- Co-Director: Bernard Chazelle, Princeton University
- Associate Director:

  Stephen Mahaney, Rutgers University thru October 13, 1997
  Julia Abrahams, Rutgers University beginning January 1, 1998

- Associate Director for Education: Joseph Rosenstein, Rutgers University

B Executive Committee

- Sandeep Bhatt, Bellcore
- Bernard Chazelle, Princeton University
- Peter Hammer, Rutgers University
- David Johnson, AT&T Labs - Research
- Stephen Mahaney, Rutgers University
- Fred S. Roberts, Rutgers University
- Joseph Rosenstein, Rutgers University
- Michael Saks, Rutgers University
- William Steiger, Rutgers University
- Robert Tarjan, Princeton University
- Mihalis Yannakakis, Bell Labs

C External Advisory Board

- Peter Arzberger, University of California/San Diego
- Jennifer Chayes, Microsoft Corp.
- Cynthia Dwork, IBM Almaden Research Center
- Avner Friedman, University of Minnesota (Chair)
- Richard Karp, University of Washington
- Daniel Kleitman, M.I.T.
- Laszlo Lovasz, Yale University
- Jill Mesirov, Boston University
- Joseph O'Rourke, Smith College
- Christos Papadimitriou, University of California/Berkeley
- William Pulleyblank, IBM Watson Research
- Dana Scott, Carnegie Mellon University
- Eva Tardos, Cornell University
- Fred Wan, University of California/Irvine
- Herbert Wilf, University of Pennsylvania

D Permanent Members of DIMACS and their Research Interests

James Abello, AT&T Labs - Research
Research Areas: Discrete and computational geometry and applications; very large graph theory; visualization and animation.

Alfred Aho, Bell Labs
Research Areas: Pattern matching algorithms; digital libraries; software engineering.

Bill Aiello, Bellcore
Research Areas: Cryptography and distributed computing.

Farid Alizadeh, Rutgers University
Research Areas: Combinatorial optimization; convex optimization; computational biology.

Eric Allender, Rutgers University
Research Areas: Computational complexity theory.

Rajeev Alur, Bell Labs
Research Areas: Formal methods; verification; logic and concurrency.

Matthew Andrews, Bell Labs
Research Areas: Simulators for disk drive scheduling and ATM private network-network interface routing.

Sanjeev Arora, Princeton University
Research Areas: Complexity theory; probabilistically checkable proofs; approximation algorithms; computational uses of randomness.

Brenda Baker, Bell Labs
Research Areas: Analysis of algorithms and software tools.

Alexander Barg, Bell Labs
Research Areas: Coding theory; combinatorics; finite algebra; computer science.

Eric Baum, NEC Research Institute
Research Areas: Cognition; artificial intelligence; neural nets; learning theory; DNA computing.

Jozsef Beck, Rutgers University
Research Areas: Combinatorics; combinatorial games; number theory.

Adi Ben-Israel, Rutgers University
Research Areas: Matrix theory; convexity and optimization; mathematical programming; mathematical economics.

Jon Bentley, Bell Labs
Research Areas: Experimental algorithms.

Sandeep Bhatt, Bellcore
Research Areas: High performance parallel and distributed computing; simulations of communications systems; algorithms and data structures.

Dan Bienstock, Bellcore
Research Areas: Optimization and parallel computing.

Matt Blaze, AT&T Labs - Research
Research Areas: Security.

Enrico Bombieri, Institute for Advanced Study
Research Areas: Probability.

Endre Boros, Rutgers University
Research Areas: Discrete and combinatorial optimization; finite geometries.

Adam L. Buchsbaum, AT&T Labs - Research
Research Areas: Algorithms; data structures; graph problems; combinatorics.

Richard Bumby, Rutgers Univesity
Research Areas: Number theory.

Robert Calderbank, AT&T Labs - Research
Research Areas: Coding theory.

Tamra Carpenter, Bellcore
Research Areas: Telecommunications applications; interior point methods; stochastic optimization.

Bernard Chazelle, Princeton University
Research Areas: Computational geometry; data structures.

Gregory Cherlin, Rutgers University
Research Areas: Logic; model theory; model theoretic algebra.

Fan Chung, Univ. of Pennsylvania
Research Areas: Combinatorics.

Vasek Chvatal, Rutgers University
Research Areas: Algorithms; combinatorics; graph theory; operations research.

Ken Clarkson, Bell Labs
Research Areas: Discrete and computational geometry; applications.

Edward G. Coffman, Bell Labs
Research Areas: Combinatorial scheduling and packing theory; probabilistic analysis of algorithms; stochastic modeling; stochastic optimization.

Edith Cohen, AT&T Labs - Research
Research Areas: Algorithm design and analysis.

William Cohen, AT&T Labs - Research
Research Areas: Machine learning and information retrieval.

John Conway, Princeton University
Research Areas: Combinatorics; games.

W.M. (Bill) Coughran, Jr., Bell Labs
Research Areas: Scientific computing.

Ingemar Cox, NEC Research Institute
Research Areas: Computer vision.

Valerie De Bellis, Rutgers University
Research Areas: Mathematics education; interactions between affect and cognition during mathematical problem solving.

Nathaniel Dean, Bell Labs
Research Areas: Graph algorithms.

Richard DeMillo, Bellcore
Research Areas: Software reliability; security.

David Dobkin, Princeton University
Research Areas: Graphics; analysis of algorithms; geometry.

Nick Duffield, AT&T Labs - Research
Research Areas: Applied probability; queueing; telecommunications; statistical mechanics.

Jonathan Eckstein, Rutgers University
Research Areas: Parallel branch and bound algorithms; proximal algorithms; monotone operators.

Kousha Etessami, Bell Labs
Research Areas: Descriptive complexity and finite model theory; applications of logic in computer science; temporal logic; model checking; computer-aided verification.

Martin Farach, Rutgers University
Research Areas: Algorithms; computational molecular biology.

Joan Feigenbaum, AT&T Labs - Research
Research Areas: Security.

Anja Feldmann, AT&T Labs - Research
Research Areas: Analysis and design of algorithms; on-line and parallel algorithms; high-speed networks; traffic characterization.

Peter Fishburn, AT&T Labs - Research
Research Areas: Utility and decisionmaking; social choice; graph theory; foundations of probability.

Steven Fortune, Bell Labs
Research Areas: Computational geometry.

Matt Franklin, AT&T Labs - Research
Research Areas: Security.

Michael Fredman, Rutgers University
Research Areas: Data structures and algorithms; computational complexity.

Yoav Freund, AT&T Labs - Research
Research Areas: Machine learning theory and applications.

Emden Gansner, AT&T Labs - Research
Research Areas: Graph drawing and algorithms; programming languages.

Michael Garey, Bell Labs
Research Areas: Algorithms and complexity; discrete math; scheduling theory.

Israel Gelfand, Rutgers University
Research Areas: Artificial intelligence; mathematics; neuroanatomy; cell biology.

Phil Gibbons, Bell Labs
Research Areas: Parallel computation; algorithms; databases.

C. Lee Giles, NEC Research Institute
Research Areas: Novel applications of AI; neural networks; machine learning to the WWW; computer systems; telecommunications; finance.

Simon Gindikin, Rutgers University
Research Areas: Theory of representations; integral geometry; several complex variables; mathematical physics; molecular biology problems.

Ram (Ramanathan) Gnanadesikan, Rutgers University
Research Areas: Statistical data analysis; graphics; pattern recognition.

Gerald Goldin, Rutgers University
Research Areas: Mathematical physics; lie theory; mathematics education; psychology of problem solving.

Ronald Graham, AT&T Labs - Research
Research Areas: Combinatorics; graph theory; number theory, theoretical computer science.

Albert Greenberg, AT&T Labs - Research
Research Areas: Networks; performance analysis.

Michael Grigoriadis, Rutgers University
Research Areas: Algorithms for network optimization.

Stuart Haber, Surety Technologies
Research Areas: Security and cryptography.

Andras Hajnal, Rutgers University
Research Areas: Set theory; discrete mathematics.

Peter. L. Hammer, Rutgers University
Research Areas: Logical analysis of data and discrete applied mathematics.

Haym Hirsh, Rutgers University
Research Areas: Artificial intelligence; machine learning.

Rob Hochberg, University of Delaware
Research Areas: Combinatorics and graph theory.

Alan J. Hoffman, Rutgers University
Research Areas: Combinatorics and optimization; linear algebra.

Frank Hwang, Chiaotung University
Research Areas: Discrete math; optimization.

Guy Jacobson, AT&T Labs - Research
Research Areas: Algorithms; data compression; games.

Suresh Jagannathan, NEC Research Institute
Research Areas: Programming languages; distributed and parallel systems.

Ravi Jain, Bellcore
Research Areas: Theoretical computer science.

Markus Jakobsson, Bell Labs
Research Areas: Cryptography/Security.

David S. Johnson, AT&T Labs - Research
Research Areas: Combinatorial optimization; complexity theory.

Jeffry Kahn, Rutgers University
Research Areas: Extremal problems; probabilistic methods; and geometry.

Narendra Karmarkar, Bell Labs
Research Areas: Linear programming.

Michael Kearns, AT&T Labs - Research
Research Areas: Computational learning theory.

Leonid Khachiyan, Rutgers University
Research Areas: Discrete optimization; mathematical programming; complexity.

Sanjeev Khanna, Bell Labs
Research Areas: Design and analysis of algorithms; combinatorial optimization; application of combinatorial optimization techniques to computer systems; databases; network design, and robotics; structural complexity theory; combinatorics; graph theory.

Joe Kilian, NEC Research Institute
Research Areas: Computational complexity; cryptography.

Nils Klarlund, AT&T Labs - Research
Research Areas: Automatic verification; programming languages and programming logics.

Alexander Kogan, Rutgers University
Research Areas: Boolean functions; discrete optimization; artificial intelligence.

Janos Komlos, Rutgers University
Research Areas: Combinatorics; probability; theoretical computer science.

Harold Kuhn, Princeton University
Research Areas: Game theory.

Robert Kurshan, Bell Labs
Research Areas: Computer aided verification.

Jeffrey C. Lagarias, AT&T Labs - Research
Research Areas: Number theory; optimization; discrete geometry; computational complexity.

Kevin Lang, NEC Research Institute
Research Areas: Neural networks; learning theory; parallel computation.

Andrea Lapaugh, Princeton University
Research Areas: Combinatorial algorithms; VLSI design.

Joel Lebowitz, Rutgers University
Research Areas: Statistical mechanics of equilibrium and nonequilibrium processes; discrete and continuous time stochastic particle systems; computer simulation of statistical mechanical model systems; foundations of statistical mechanics.

David Lee, Bell Labs
Research Areas: Network protocols.

Richard Lipton, Princeton University
Research Areas: Architecture; complexity; VLSI.

Nicholas Littlestone, NEC Research Institute
Research Areas: Computational learning theory.

Carsten Lund, AT&T Labs - Research
Research Areas: Complexity theory; interactive proofs.

Richard Lyons, Rutgers University
Research Areas: Finite simple groups.

Stephen R. Mahaney, Rutgers University
Research Areas: Computational complexity

Dahlia Malkhi, AT&T Labs - Research
Research Areas: Security in distributed and networked systems; mobile code security; fault tolerant software solutions and analysis; distributed coding; advanced communication networks and protocols.

Yossi Matias, Bell Labs
Research Areas: Algorithms; parallel computation; query processing and optimization; data structures; internet applications; probabilistic methods.

Stephen B Maurer, Swarthmore College
Research Areas: Discrete math; algorithmics; modeling; mathematical exposition.

Alain Mayer, Bell Labs
Research Areas: Cryptography: electronic commerce; computer and network security; Internet technology; network and security management; network protocols; multimedia; distributed algorithms and systems.

Michael Merritt, AT&T Labs - Research
Research Areas: Distributed algorithms; cryptographic protocols; formal methods.

Milena Mihail, Bellcore
Research Areas: Network algorithms; matroids.

Clyde Monma, Bellcore
Research Areas: Combinatorial optimization; applications to telecommunications.

S. (Muthu) Muthukrishnan, Bell Labs
Research Areas: Experimental algorithms; algorithms in compilers; programming languages etc; probabilistic methods; information retrieval; pattern matching; computational biology.

Arun Netravali, Bell Labs
Research Areas: Digital image processing.

Michael O'Nan, Rutgers University
Research Areas: Permutation groups; simple groups.

Andrew Odlyzko, AT&T Labs - Research
Research Areas: Cryptography; number theory; combinatorics; coding theory; analysis; probability theory.

Andrew Ogielski, Rutgers University
Research Areas: Simulation.

Rafail Ostrovsky, Bellcore
Research Areas: Cryptography; computer security; distributed protocols and algorithms.

Steven Phillips, AT&T Labs - Research
Research Areas: Algorithms in telecommunications.

Andras Prekopa, Rutgers University
Research Areas: Stochastic processes; stochastic optimization; linear and nonlinear programming.

Anuj Puri, Bell Labs
Research Areas: Computer-aided verification; real-time systems; hybrid systems; computer networks; transportation systems; optimization.

Sivaram Rajagopalan, Bellcore
Research Areas: Cryptography; average-case complexity.

Sridhar Ramaswamy, Bell Labs
Research Areas: Databases performance; I/O algorithms; data mining.

Satish Rao, NEC Research Institute
Research Areas: Parallel computing and combinatorial optimization.

Jim Reeds, AT&T Labs - Research
Research Areas: Security; cryptography; geometry; probability theory; cellular telephony; history of cryptography.

Marty Reiman, Bell Labs
Research Areas: Applied probability: queueing theory, stochastic control, diffusion processes; applied to broadband networking and other computer/communication systems.

Nick Reingold, AT&T Labs - Research
Research Areas: Algorithms; network design.

Michael Reiter, AT&T Labs - Research
Research Areas: Research secure systems.

Mauricio G. C. Resende, AT&T Labs - Research
Research Areas: Combinatorial optimization; interior point methods.

Eric Ristad, Princeton University
Research Areas: Computational linguistics; statistical modeling.

Fred S. Roberts, Rutgers University
Research Areas: Mathematical models in the social, behavioral, biological and environmental sciences and of problems of communications and transportation; graph theory; measurement theory; utility, decisionmaking, and social choice.

Joseph G. Rosenstein, Rutgers University
Research Areas: Mathematics education; linear orderings; recursive model theory.

Michael H. Rothkopf, Rutgers University
Research Areas: Practice of operations research; models of bidding; energy economics.

Avi Rubin, AT&T Labs - Research
Research Areas: Cryptography; network security; secure systems; web security.

Michael Saks, Rutgers University
Research Areas: Computational complexity; algorithms; extremal set theory; partially ordered sets.

Iraj Saniee, Bellcore
Research Areas: Network and combinatorial optimization; SONET and ATM network design.

Peter Sarnak, Princeton University
Research Areas: Data structures; graph algorithms.

Serap Savari, Bell Labs
Research Areas: Data compression; communication; algorithms; operations research.

Rob Schapire, AT&T Labs - Research
Research Areas: Learning theory and applications.

Robert Sedgewick, Princeton University
Research Areas: Analysis of algorithms; algorithm visualization.

Ravi Sethi, Bell Labs
Research Areas: Programming languages.

Paul Seymour, Princeton University
Research Areas: Graph theory.

David Shallcross, Bellcore
Research Areas: Discrete optimization.

David F. Shanno, Rutgers University
Research Areas: Linear and nonlinear programming; numerical analysis; parallel computing.

Saharon Shelah, Rutgers University
Research Areas: Logic; model theory.

Lawrence Shepp, Rutgers University
Research Areas: Probability; tomography; combinatorics.

Peter Shor, AT&T Labs - Research
Research Areas: Quantum computing; algorithms; computational geometry.

Charles Sims, Rutgers University
Research Areas: Computational group theory and algebraic algorithms.

Jaswinder Singh, Princeton University
Research Areas: Parallel computing applications and systems; computational biology.

Neil J. A. Sloane, AT&T Labs - Research
Research Areas: Mathematics; statistics.

Warren Smith, NEC Research Institute
Research Areas: Fundamental limits on computation set by laws of physics; algorithms; chess programming; optimal shapes.

Eduardo Sontag, Rutgers University
Research Areas: System and control theory; learning and neural networks.

Diane Souvaine, Rutgers University
Research Areas: Design and analysis of geometric and graph-theoretic algorithms.

William Steiger, Rutgers University
Research Areas: Algorithms; computational geometry; probability; parallel computation.

Ken Steiglitz, Princeton University
Research Areas: Computing with particles; economic simulation; computer music.

Martin Strauss, AT&T Labs - Research
Research Areas: Massive data sets; security; complexity theory.

Torsten Suel, Bell Labs
Research Areas: Algorithm design and analysis; experimental algorithms; parallel computation; databases.

Don X. Sun, Bell Labs
Research Areas: Industrial experimental designs; quality improvement through experimentations and data analysis; combinatorial design of experiments; signal processing and speech recognition; pattern classification and artificial intelligence; graphical methods for large data analysis.

Ram Swaminathan, Bell Labs
Research Areas: Algorithms; combinatorics; probability; geometry; complexity.

Mario Szegedy, AT&T Labs - Research
Research Areas: Algorithms and complexity; approximations.

Endre Szemeredi, Rutgers University
Research Areas: Number theory; extremal graphs; theoretical computer science.

Robert Tarjan, Princeton University
Research Areas: Data structures; graph algorithms; complexity.

Simon Thomas, Rutgers University
Research Areas: Model theory; infinite groups.

Robert J. Vanderbei, Princeton University
Research Areas: Optimization.

Wolmer Vasconcelos, Rutgers University
Research Areas: Commutative algebra and computational algebra.

Kiem-Phong Vo, AT&T Labs - Research
Research Areas: Discrete/graph data structures and algorithms; graph visualization; pattern matching; data compression; software architecture and design; reliability; libraries.

Jeff Westbrook, AT&T Labs - Research
Research Areas: Data structures and algorithms; on line algorithms; network design.

Ward Whitt, AT&T Labs - Research
Research Areas: Operations research; telecommunications systems; probability and stochastic processes and queues.

Gordon Wilfong, Bell Labs
Research Areas: Computational geometry; algorithms.

Albert C. Williams, Rutgers University
Research Areas: Linear, nonlinear, and integer programming; applications of operations research.

Peter Winkler, Bell Labs
Research Areas: Discrete math; probability; theory of computing.

Avishai Wool, Bell Labs
Research Areas: Quorum systems; fast communication networks; combinatorial optimization.

Paul Wright, AT&T Labs - Research
Research Areas: Probabalistic methods; stochastic processes.

Rebecca Wright, AT&T Labs - Research
Research Areas: Cryptography; network security.

Mihalis Yannakakis, Bell Labs
Research Areas: Algorithms; complexity; databases; testing.

Andrew Yao, Princeton University
Research Areas: Computational complexity; analysis of algorithms.

Ann Yasuhara, Rutgers University
Research Areas: Recursive functions; logic; theoretical computer science.

Lisa Zhang, Bell Labs
Research Areas: Design, analysis and implementation of algorithms.

Shiyu Zhou, Bell Labs
Research Areas: Design and analysis of data structures and algorithms; computational complexity.