DIMACS Special Focus on Computational and Mathematical Epidemiology: 2004 Annual Report

September 1, 2003-August 31, 2004 (prepared May 2004)

Participant Individuals:

Fred S. Roberts, Principal Investigator
Contribution to project: involved in all the working groups and all aspects of the project.

S. Muthukrishnan, Senior Personnel
Contribution to project: Organizer of Working Group on Data Mining and Epidemiology

Martin Farach-Colton, Senior Personnel
Contribution to project: has been playing a role in the connections of the epidemiology work to issues of bioinformatics/genetics.

Donald Hoover, Senior Personnel
Contribution to project: Organizer of Working Group on Adverse Event/Disease Reporting, Surveillance and Analysis I and II.

Joel Leibowitz, Senior personnel
Contribution to project: Supervised research of Graduate Student Jaewook Joo, served on Project Advisory Committee.

David Madigan, Senior Personnel
Contribution to project: Organizer of Tutorial on Epidemiology for Mathematical Scientists. Organizer of Working Group on Adverse Event/Disease Reporting, Surveillance and Analysis I and II. Organizer of Tutorial on Statistical and Other Analytic Health Surveillance Methods.

Marc Lipsitch, Special Focus Organizer
Contribution to project: Member of Organizing Committee

Megan Murray, Special Focus Organizer
Contribution to project: Member of Organizing Committee

David Ozonoff, Subcontract
Contribution to project: Organizer of tutorial on epidemiology for mathematical scientists. Organizer of working group on data mining and epidemiology. PI of subcontract to Boston University

Burton Singer, Consultant
Contribution to project: Member of Organizing Committee

Daniel Wartenberg, Consultant

Contribution to project: Organizer of Tutorial on Epidemiology for Mathematical Scientists.

Sunetra Gupta, Consultant
Contribution to project: Member of Organizing Committee

David Krakauer, Consultant
Contribution to project: Member of Organizing Committee

Simon Levin, Consultant
Contribution to project: Member of Organizing Committee, Organizer of the DIMACS Working Group on the Ecology and Evolution of Influenza and Related Viruses

Ilya Muchnik, Senior personnel
Contribution to project: Organizer of Working Group on Data Mining and Epidemiology. Developing a cancer registry modeling project with Dr. James Abello and graduate student David Millman.

Graham Cormode, Postdoc
Contribution to project: Working on data mining and epidemiology, primarily with Dr. Muthukrishnan. Organizer of the Working Group on Data Mining and Epidemiology.

Patrick DeLeenheer, Postdoc
Contribution to project: Working on chemostat models.

Jaewook Joo, Graduate student
Contribution to project: Working on spatio-temporal modeling of disease under the supervision of Professor Joel Lebowitz.

Yihua Wu, Graduate Student
Contribution to project: Working with Dr. Muthukrishnan on building skeleton decision trees for mining massive data such as might arise from using wireless sensors to monitor people in *loco*.

Smriti Bhagat, Graduate Student
Contribution to project: working with Dr. Muthukrishnan on bio-sensors, in particular, the ECG and EMG sensor by the Cleveland Medical Devices Inc. The overall objective is to use the system for physiological and epidemiological studies in conjunction with other sensors.

German Enciso, Graduate Student
Contribution to project: working with Dr. Eduardo Sontag in the study of the long-term behavior of dynamical systems that are frequent in molecular biology.

Rohan Fernandes, Graduate Student
Contribution to project: working with Dr. Martin Farach-Colton, on problems related to sensor networks.

Samir Goel, Graduate Student

Contribution to project: working on developing an etiquette protocol for ultra low power operation in sensor networks with Dr. Tomasz Imielinski.

Stephen Hartke, Graduate Student
Contribution to project: Working on generalizations of the 'voter' model for disease spread and on the analogies between spread of opinion models and spread of disease models and models of vaccination strategies using graph structures with Dr. Fred Roberts.

Jaewon Kang, Graduate Student
Contribution to project: The recent advancements in MEMS technology, processor design, and wireless communication have enabled a wide range of monitoring application using networks of small sensor devices. For example, a sensor network can be deployed to study the behavior of an endangered species. He is investigating several strategies to alleviate congestion in a crisis state.

Liming Wang, Graduate Student
Contribution to project: study of dispersal and space-dependence in population models, applied in a molecular biology (cell-cycle) context. This work is performed under the supervision of Dr. Eduardo Sontag.

David Millman, Graduate Student
Contribution to project: Working on a cancer registry modeling project with Dr. James Abello and Dr. Ilya Muchnik.

James Abello, long term visitor
Contribution to project: Organizer of the Working Group on Data Mining and Epidemiology, Second Meeting. Working on a cancer registry modeling project with Dr. Ilya Muchnik and graduate student David Millman.

Michael Capalbo, long term visitor
Contribution to project: working on random graph models of spread of disease.

Partner Organizations:

Princeton University: Collaborative Research; Personnel Exchanges. Partner organization of DIMACS. Individuals from the organization participated in the program planning.

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

Telcordia Technologies: Collaborative Research; Personnel Exchanges. Partner organization of DIMACS. Individuals from the organization participated in the program planning.

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

Lucent Technologies, Bell Labs: Collaborative Research; Personnel Exchanges.

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

Rutgers, The State University of New Jersey: Collaborative Research

Office of Naval Research: Financial Support

Alfred P. Sloan Foundation: Financial Support

Burroughs-Wellcome Fund: Financial Support

New Jersey Commission on Science and Technology: Financial Support

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

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

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

American Statistical Association: Financial Support
Supported Jana Asher, graduate student from Department of Statistics, Carnegie Mellon University, to attend and give talk at DIMACS Working Group on Adverse Event/Disease Reporting, Surveillance and Analysis from October 16 - 18, 2003.

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

Centers for Disease Control: Collaborative Research; Personnel Exchanges. Individuals from the organization participated in the program planning and working group/workshop organization.

Other Collaborators:

Rakesh Agrawal, IBM Almaden: Organizer of the Working Group on Privacy / Confidentiality of Health Data
Larry Cox, CDC: Organizer of the Working Group on Privacy / Confidentiality of Health Data
Zhilan Feng, Purdue University: Organizer of Working Group on Genetics and Evolution of Pathogens
John Glasser, CDC: Organizer of the Working Group on Methodologies for Comparing Vaccination Strategies
Joe Fred Gonzalez, CDC: Organizer of the Working Group on Privacy / Confidentiality of Health Data
Harry Guess, University of North Carolina: Organizer of the Working Group on Privacy / Confidentiality of Health Data

Herbert Hethcote, University of Iowa: Organizer of the Working Group on Methodologies for Comparing Vaccination Strategies
Natalia Komarova, Rutgers University and Institute for Advanced Study: Organizer of the DIMACS/IAS Workshop on Genomic Instability in Cancer: Biological and Mathematical Approaches
Arnold Levine, Institute for Advanced Study, and RWJSM, Organizer of the DIMACS/IAS Workshop on Genomic Instability in Cancer: Biological and Mathematical Approaches
Kenton Morgan, University of Liverpool: Organizer of the Working Group on Data Mining and Epidemiology
Henry Rolka, CDC: Organizer of the Working Group on Adverse Event/Disease Reporting, Surveillance, and Analysis
Tomas Sander, HP Labs: Organizer of the Working Group on Privacy / Confidentiality of Health Data

Research and Education Activities:

This project is organized around a series of working groups that meet and the resulting collaborations. There are also workshops and tutorials. In addition, a regular seminar series brought participants together. This seminar series was jointly sponsored by the **Bio**logical, **Ma**thematical, and **P**hysical **S**ciences Institute for Quantitative Biology (BIOMAPS).

Workshops, Working Group Meetings and Tutorials

Tutorial on Statistical and Other Analytic Health Surveillance Methods, June 17 - 20, 2003

Organizers:
David Madigan, Rutgers University
Henry Rolka, CDC
Martin Kulldorff, University of Connecticut

Surveillance is a core activity for public health practice. Standard tasks include outbreak detection, post-marketing drug adverse event surveillance, immunization program evaluation and institutional comparisons. Recent concerns about bioterrorism have attracted increased attention to new analytic surveillance methods applied to non-traditional data sources such as over-the-counter drug sales, ambulance dispatches and pre-diagnostic patient encounter summaries. Both the novel methods and new data sources raise significant statistical and computing challenges. The need for innovative statistical methods for surveillance of multiple data streams and the evaluation of utility for new methodologies are especially pressing issues.

This DIMACS tutorial provided an overview of the challenges and addressed an array of statistical surveillance tools and techniques. The first two days of the tutorial focused on hands-on use of specific methodologies through case-studies and elementary introductions. The second two days concerned themselves more with current research activities and new research challenges.

The tutorial addressed the following topics:

Scan statistics
Statistical process control
Changepoint methods
Bayesian approaches
Rule-based surveillance
False discovery rate and sequential testing
Text data surveillance

Working Group on the Ecology & Evolution of Influenza and Related Viruses, June 29-July 2, 2003

Organizers:
Jonathan Dushoff, Princeton University
Simon Levin, Princeton University
Joshua Plotkin, IAS/Princeton University

Influenza viruses are negative-stranded RNA viruses that cause significant human mortality and morbidity worldwide. The study of influenza epidemiology has profound significance for human health. In addition, influenza viruses offer a remarkable system for evolutionary study. Over the past two decades, the influenza A viral genome has undergone roughly the equivalent of 20 million years of mammalian evolution.

Influenza viral evolution depends, through frequency-dependent selection, upon influenza's epidemiology - that is, the patterns of host infection. Conversely, annual outbreaks of influenza depend upon the evolution of viral strains. Thus, an understanding of influenza viruses requires a synthetic approach combining both the evolution and the ecology of the pathogen.

This working group brought together experts who work on influenza from a range of viewpoints. In order to facilitate a synthesis of perspectives, the working group was organized around half-day sessions, each with one or two short talks followed by active discussion. The five primary topics of inquiry were:

The structure and mechanisms of viral antigenic drift
Pandemics and interactions between influenza subtypes
Disease dynamics and the interpretation of epidemiological data
Within-host dynamics and their significance for influenza epidemiology and evolution
The implications of historical pandemics for other diseases, particularly SARS

The working group also included a grab-bag section to discuss related issues such as variation in viral protein structure and relations between viral genotype and antigenic phenotype. Through short presentations and discussions that include all participants, the working group aimed to improve understanding of influenza viruses on the basis of mathematical modeling combined with molecular, epidemiological, and immunological data.

Working Group on Genetics and Evolution of Pathogens, November 24 - 25, 2003

Organizer:
Zhilan Feng, Purdue University

The main purpose of this meeting was to bring together international leaders and active researchers working in areas related to the modeling, simulation and analysis of the evolution of pathogens, to explore more effective approaches to dealing with challenges in understanding the evolution of infectious agents when epidemiology is coupled with ecology and/or genetics. The topics included mathematical/statistical/computational approaches to modeling transmission dynamics of multiple strains; impact of control strategies on the evolution of an infectious agent (e.g., virulence, drug-resistance); multiple time-scales in coupled systems; population genetics/population biology of both parasites and hosts. The focused diseases included malaria, tuberculosis, influenza, and other infectious diseases for which more is known about their evolution.

Working Group: Privacy / Confidentiality of Health Data, December 10 - 12, 2003

Organizers:
Rakesh Agrawal, IBM Almaden
Larry Cox, CDC
Joe Fred Gonzalez, CDC
Harry Guess, University of North Carolina

Privacy concerns are a major stumbling block to public health surveillance, in particular bioterrorism surveillance and epidemiological research. Moreover, the Health Insurance Portability and Accountability Act (HIPAA) of 2002 imposes very strict standards for rendering health information not individually identifiable. How to use large health care databases to detect medical or terrorist risks and improve health care quality while maintaining privacy and confidentiality of the data is a serious challenge. This working group explores computational techniques for ensuring that the identity of an individual contained in a released data set cannot be identified. The challenge is to produce anonymous data that is specific enough to be useful for research and analysis. It considers ways to remove direct identifiers (social security number, name, address, telephone number), and ways to aggregate, substitute, and remove information from data sets. Also of interest are questions having to do with using electronic data matching to link data elements from various sources/data sets in order to identify individuals, while maintaining privacy of others. The group investigates methods for privacy protection in field-structured data and ways to extend existing methods to large data sets, as well as systems to render textual data sufficiently anonymous. Finally, the group explores formal frameworks for disclosure control and formal protection models. Sixteen talks were presented in this working group meeting, including:

Overview of Statistical Disclosure Limitation, Lawrence H. Cox, Associate Director, ORM, NCHS, CDC
Legal and Regulatory Framework in the United States and the European Union, Oliver Johnson, Merck and Co., Inc.
The Health Insurance Portability and Accountability Act (HIPAA) and its Implications on Epidemiological Research Using Large Databases, K. Arnold Chan, Harvard University

Health Care Databases under HIPAA: Statistical Approaches to De-identification of Protected Health Information, Judith Beach, Quintiles Transnational
Protecting the Privacy of Healthcare Data While Preserving the Utility of Geographic Location Information for Epidemiologic Research, Daniel Barth-Jones, Center for Healthcare Effectiveness Research and Department of Medicine, Wayne State University, School of Medicine
Privacy Technologies and Challenges in their Deployment, Tomas Sander, HP Labs
Software Demonstration of the use of Hippocratic Database Technology in Supporting a Health Care Provider. Tyrone Grandison, IBM
Cryptographic Techniques for Confidentiality of Aggregate Statistics on Health Data, Giovanni DiCrescenzo, Telecordia
Tutorial on Data Mining, David Madigan, Rutgers University
Using Data Mining Techniques to Harvest Information from Clinical Trials, Richard D. De Veaux, Williams College, Williamstown, MA.
Experimental Results on Privacy-Preserving Statistics Computation, Rebecca Wright, Stevens Institute of Technology
Semantic Web Services for Privacy/Confidentiality of Health Care Data, Nabil Adam, Rutgers University
Privacy/Confidentiality Issues in Collecting Agricultural Data, Gary Smith, University of Pennsylvania
Private Analysis of Data Sets, Benny Pinkas, HP Labs, NJ
Overview of Masking Schemes for Microdata, Jay J. Kim, ORM, NCHS, CDC
Statistical Disclosure Limitation in Tabular Data and Related Mathematical and Computational Problems, Lawrence H. Cox, ORM, NCHS, CDC

The working group developed a variety of ideas at this meeting that will lead to future investigations. A key set of challenges arises for teams involving cryptographers and epidemiologists. A meeting to explore these issues is currently being planned. A second major challenge falls in the area of data de-identification and the role of combinatorial optimization in this field. The working group plans a meeting at which statisticians, epidemiologists, and combinatorial optimizers all discuss the issues and lay out a research agenda. Additional challenges lie in identifying specific guidelines for statisticians in certifying HIPAA compliance. The working group will be organizing a tutorial on this topic.

Challenges at the interface between cryptography and epidemiology/health data analysis are given below. Future meetings will produce similar lists for the interface between data de-identification and combinatorial optimization and for the interface between HIPAA compliance and statistics.

1. Different Functionalities and Specific Challenges for Cryptography.
(a) Does transferring data between a hospital and testing lab or other problems of transferring health data require any different cryptographic tools than we need for financial transactions?
(b) We should distinguish between problems of transferring data and problems of computing with data, especially distributed data. See 2 for challenges in this direction.
(c) How do we improve the performance of cryptographic schemes (secure multiparty computation) to make them affordable for practical applications?

(d) How do we prove compliance, cryptographically, with a stated privacy policy?
2. Privacy-preserving Data Mining and Privacy-preserving Data Sharing.
(a) Identify specific functionalities needed for health data applications.
(b) Make secure multi-party computation more efficient for large databases (a generic challenge).
(c) Extend secure multi-party computation to clustering. Since clustering is hard, we might have to settle for approximate solutions. More generally, can we extend secure multi-party approximation?
(d) Is it possible to modify secure multi-party computation protocols so one doesn't have to access all data elements?
(e) What are the issues involved in privacy-preserving data sharing in general and secure multiparty computation in particular if we want to take into consideration what the output itself might leak about the data?
3. Tracking Disclosed Information (a topic related to secure software and secure computing environments as well as cryptography)
(a) Can we "send" with disclosed information some restrictions on its use, e.g., future disclosure?
(b) Can we "send" with disclosed information restrictions on the length of time it can be saved/used?
(c) Can we do this tracking if there are later changes in disclosure limitations?
4. Can we develop good auditing technologies?
This question applies well beyond cryptography. In health data, it is concerned with distinguishing between a transaction (e.g., looking at a patient record) that is legitimate and one that is not. A well-known method involves tracking authorizations. However, are there smart methods to audit large data sets of transactions to find illegitimate transactions?
5. "Customizable" Privacy
Software employed by different partners may differ in privacy protections/policy and processing. This presents cryptography with complex privacy management concerns and it would be important to develop privacy protocols that are readily "customizable" to different users. How do we achieve customized privacy that would satisfy/balance the privacy policies of all participants?
6. Dynamic Query Authorization and Forbidden Question Combinations
(a) It is an old topic to change query authorization based on previous queries so as to make it impossible to make forbidden inferences. But how do we do this in the encryption situation and with widely distributed data sets?
(b) A simpler challenge arises if we have specific questions and some combination of them that is forbidden in advance. Even here, there are cryptographic challenges if we hide the questions from the database owner.
7. Revealing Partial Information
It may not be known in advance which information will and will not be sensitive. Traditionally, cryptography does not allow information leakage unless it is explicitly defined as part of the input. Dynamically-changing disclosure limitations pose challenges for cryptography, e.g., in secure multiparty computation.
8. Cleaning Data and Maintaining Privacy
Data preparation and cleaning is a major part of real life statistics. Can this be done in a privacy enhanced way?

An extensive report on this activity is given in *Report on DIMACS Working Group on Privacy / Confidentiality of Health Data*, which can be found at
http://dimacs.rutgers.edu/Workshops/Health/priv-health-data-12-03.pdf

Working Group on Adverse Event/Disease Reporting, Surveillance, and Analysis, Second Meeting: February 19 - 20, 2004

Organizers:
Donald Hoover, Rutgers University, Statistics
David Madigan, Rutgers University, Statistics
Henry Rolka, CDC

Analytical methods for detecting unusual events and/or changepoints in univariate data streams have attracted considerable attention over the last several decades. In the context of syndromic surveillance, for example, methods based on control charts, scan statistics, Bayesian models, sequential probability ratio tests, etc., are common. Recently, however, practitioners face the new challenge of monitoring multidimensional data streams, often involving heterogenous data types and varying time scales. For example, a bioterrorism detection application might simultaneously monitor 911 call volume, syndromic codes from emergency room chief complaints, over-the-counter medications, and absenteeism.

This working group brought together researchers working on analytic methods in this area. The first morning featured panels and talks on existing activities in multivariate surveillance. The afternoon featured presentations on relevant analytical methods from other areas such as statistical process control, visualization, and temporal Bayesian networks. The second (half-day) attempted to summarize the state-of-the-art and propose a research agenda.

Working Group on Data Mining and Epidemiology, Second Meeting, March 18-19, 2004

Organizers:
James Abello, DIMACS
Graham Cormode, DIMACS
Kenton Morgan, University of Liverpool
David Ozonoff, Boston University

This meeting brought together researchers from a variety of areas: epidemiologists, statisticians, data miners and others. The aim was to foster greater communications between these disparate groups, to help data mining experts understand concrete epidemiological questions, and to help epidemiologists understand better the results possible from using data mining techniques. Several collaborations have already resulted from interactions at the meeting and after, and publications from talks given at the meeting are in progress. A report of this meeting, detailing some of the questions posed as challenges to the area, is in preparation.

Epidemiology is an observational science that concerns itself with finding and explaining patterns of health and disease in populations, usually of humans, but also populations of animals,

insects and plants. Data mining is an active area of research interested in finding algorithms for describing latent patterns in often very large data sets. This Working Group has the objective of fostering collaboration between these two disciplines. In March of 2004 it organized a two-day meeting at DIMACS to bring these two groups together in a format designed to initiate such collaborations.

Each day there were a small number of presentations from epidemiologists, statisticians and computer scientists. They gave a detailed description of a specific study analyzed by methods current in today's epidemiological practice. The studies were selected to represent a range of typical techniques such as exploratory factor analysis, multiple logistic regression, proportional hazards models, generalized additive models and other current methods. The epidemiologists concentrated on the epidemiological aspects of the study, with some relatively brief explanations of the main ideas by the statisticians. Computer scientists or discrete mathematicians commented on algorithmic issues and alternate approaches to the underlying questions. This led into a moderated audience-wide discussion of each study.

The goal was to identify some fruitful areas for applying new techniques from theoretical computer science, discrete mathematics, and statistics and to allow new collaborations among participants to be initiated. This working group brought together with epidemiologists experts interested in exploring data mining and algorithmic techniques potentially useful in epidemiology from a range of viewpoints. In order to facilitate a synthesis of perspectives, the working group was organized around half-day sessions, each with one or two talks followed by active discussion. The five primary topics of inquiry were:

Disease Dynamics, Epidemiological Studies and the Interpretation of Data
Genetic, Pharmaceutical and Environmental Epidemiology
Elements of Statistical Learning and Models
Regression Methods, Support Vector Machines and Cluster Analysis
Association Rules and Lattice/Graph Theoretical Methods

The working group meeting included a problem section and a report on the first Epidemiological DIMACS Challenge that began in September 2003. Through short presentations and discussions that included all participants, the workshop aimed to foster research cooperation among Epidemiologists, Computer Scientists and Mathematicians. The goal was to identify fundamental epidemiological problems that can benefit from efficient computational, statistical and mathematical models that can aid in the processing and understanding of combined epidemiological, and immunological data.

An important outcome of the meeting was the identification of data mining methods appropriate for the early indication of plant diseases. This has led to a collaboration between DIMACS and the National Plant Diagnostic Network headquartered at UC Davis.

Working Group on Methodologies for Comparing Vaccination Strategies, May 17-20, 2004

Organizers:
John Glasser, CDC

Herbert Hethcote, University of Iowa

This working group focused on the use of mathematical modeling and computer simulations to provide theoretical comparisons of vaccination and other intervention programs. Examples of the effective use of modeling were studied for vaccine-preventable diseases such as measles, rubella, varicella, pertussis, polio, Haemophilus influenza type b and pneumococcal, Hepatitis A and B, and Lyme disease.

When vaccination strategies compare vaccination of different age groups, age-structured mathematical models are necessary. Computer simulations of these highly complex systems with hundreds of demographic and epidemiological classes allow comparisons of vaccination strategies and rapid evaluation of potential intervention programs when field trials would be prohibitively expensive. Both expected and worst case scenarios must be considered. We brought working epidemiologists and policymakers together with experts in mathematical modeling and computer simulation in order to formulate and analyze models to deal with current problems and explore ways to simulate larger, more complex, more realistic models.

There was a good deal of discussion about influenza. The persistence of influenza depends on its ability to evolve so that new strains and subtypes of the virus appear and old ones reappear. This constant evolution means that vaccines need to be updated frequently and that resistance to drug therapies can easily arise. These issues make modeling difficult and require large, powerful computational methods to analyze the models. Some specific research issues to which the group applied its methods are the following.

How will varicella vaccination affect the incidence of shingles?
Is acellular pertussis vaccination of adolescents or adults more effective in reducing cases among infants than vaccination of children?
Because the oral polio vaccine virus can mutate back to the wild virus, as occurred recently in the Caribbean and the Philippines, can polio be eradicated using the oral polio vaccine alone?
How does the yearly composition of the influenza vaccine affect the drift of the influenza virus?

There was also discussion of vaccination strategies in the event of or in anticipation of bioterrorist attacks involving smallpox, anthrax, or influenza. Research issues here include:

What are the risks/advantages of different smallpox vaccination strategies using existing supplies of the vaccine?
How do the conclusions change if we dilute existing supplies so as to allow more people to be vaccinated but can only achieve a certain probability of protection or only a reduction of symptoms to treatable levels?
How do vaccination strategies depend upon our ability to identify a spot smallpox epidemic?
How do these conclusions change for diseases such as anthrax where transmission is not person-to-person?

Organizers:
Natalia Komarova, Rutgers University and Institute for Advanced Study, Princeton,
Arnold Levine, Institute for Advanced Study, Princeton, and Robert Wood Johnson School of Medicine and Dentistry

Many cancers are characterized by genomic instability. This often involves gross chromosomal abnormalities such as deletion and duplication of chromosomes or chromosome parts, chromosomal rearrangements and mitotic recombinations, generally termed as ``chromosomal instability''. Other types of genomic instability are characterized by an increased rate of small-scale genetic changes (such as microsatellite instability). The role of genomic instability (GI) for cancer progression is a very important, yet unresolved question.

Mechanisms of genomic instability are intimately related to many aspects of cell biology, such as cell cycle regulation, DNA damage and repair, cell aging and telomere function. On the other hand, the role of GI in cancer initiation and progression can be studied from the point of view of somatic evolution of cells, where mathematical modeling can play an important part. The goal of this workshop is to bring together people who study the phenomenon of GI from different perspectives, for extended discussions and learning. In particular we hope to start a dialog between experimental scientists and mathematical modelers and identify areas for cross-pollination between the two communities.

Theoretical approaches in the area of GI and cancer progression are still in the budding stage. However we believe that quantitative methods can provide an important tool in deciphering the mystery of GI. This is similar in spirit to studying infectious diseases, where mathematicians and biologists working hand by hand have generated many useful insights. We expect that a similar phenomenon will happen in the area of cancer research. This workshop is the first attempt to create a forum for experimental biologists and modelers of genomic instability, which will hopefully encourage collaboration and improve our understanding of cancer.

DIMACS/BIOMAPS Seminar in Quantitative Biology and Epidemiology:

Graham Cormode and Patrick De Leenheer, post-doctoral research fellows at DIMACS, in association with members of the BioMaps research group, organized the joint DIMACS/BIOMAPS Seminar Series on Quantitative Biology and Epidemiology during the 2003-2004 academic year. In total, there were almost forty talks given in the first year of this joint venture. The audience came from a wide variety of disciplines, including Biology, Mathematics, Medicine, Physics, Chemistry, Computer Science and more, including regular participants from Rutgers, The University of Medicine and Dentistry of New Jersey (UMDNJ), Robert-Wood Johnson Medical School (RWJMS), Princeton and beyond. The program of this series, including titles and speakers is given here. Abstract information can be found at:

http://dimacs.rutgers.edu/SpecialYears/2002_Epid/seminars.html

Wednesday, September 17, 2003

Speaker: **Marcelo Magnasco**, Rockefeller University
Title: The dynamical basis of auditory acuity

Wednesday, September 24, 2003
Speaker: **Monica Driscoll**, Rutgers University
Title: Elaborating cellular and molecular mechansims of aging in C. elegans: lessons and themes from simple old animals

Wednesday, October 1, 2003
Speaker: **Joshua Weitz**, Princeton University
Title: Scaling laws in organismal biology: Navigating the morass

Wednesday, October 15, 2003
Speaker: **Boris Kholodenko**, Thomas Jefferson University
Title: Molecular and mechanistic analyses of cellular networks: Can we navigate through molecular jungles

Wednesday, October 22, 2003
Speaker: **Stanislav Shvartsman**, Princeton University
Title: Modeling and manipulating EGFR-mediated cell communication in development

Wednesday, October 29, 2003
Speaker: **Sergei Grigoryev**, Penn State University College of Medicine
Title: Higher-order chromatin structure and its implications for gene regulation

Wednesday, November 5, 2003
Speaker: **Ira Schwartz**, Naval Research Laboratory
Title: Non-equilibrium stochastic outbreaks: From prediction to control

Wednesday, November 12, 2003
Speaker: **Jonathan Dushoff**, Princeton University
Title: Detecting selection from patterns of codon usage

Wednesday, November 19, 2003
Speaker: **Burkhardt Rost**, Columbia University
Title: Evolution teaches predicting protein function?

Tuesday, November 25, 2003
Speaker: **Ulf Dieckmann**, IIASA, Austria
Title: Spatial complexity in ecology and evolution

Wednesday, November 26, 2003
Speaker: **Horst Thieme**, Arizona State University
Title: Competitive dynamics in a model for river blindness with cross-immunity

Wednesday, December 3, 2003

Speaker: **Natalia Komarova**, Rutgers University
Title: Modeling DNA transcription dynamics

Wednesday, December 10, 2003
Speaker: **Reka Albert**, Penn State University
Title: Topology and robustness of the segment polarity gene network

Wednesday, December 17, 2003
Speaker: **Philip Holmes**, Princeton University
Title: Piecewise-holonomic mechanics, hybrid dynamical systems, and escaping cockroaches

Wednesday, January 21, 2004
Speaker: **Laura Finzi**, University of Milan
Title: Topological modifications of DNA in transcriptional regulation

Wednesday, January 28, 2004
Speaker: **Yuhai Tu**, IBM Research
Title: Receptor cross talk and signal amplification: some recent results on signal transduction in bacterial chemotaxis

Thursday, February 5, 2004
Speaker: **Santiago Schnell**, Oxford University
Title: Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws

Tuesday, February 10, 2004
Speaker: **Peter Thomas**, Salk Institute
Title: Inside the mind of the amoeba: Simulation and analysis of biochemical signal transduction channels

Thursday, February 12, 2004
Speaker: **Stanislav Shvartsman**, Princeton University
Title: Modeling and manipulation of EGFR signaling in Drosophila development

Wednesday, February 18, 2004
Speaker: **Robijn Bruinsma**, UCLA
Title: The physics of cell adhesion and mechano-sensing

Wednesday, February 25, 2004
Speaker: **Rahul Siddharthan**, Rockefeller University
Title: PhyloGibbs: A Gibbs sampler for motif detection incorporating phylogenetic information

Wednesday, March 3, 2004
Speaker: **Gerard Manning**, Rutgers University
Title: Counterion condensation on polyelectrolytes: Theory and applications

Wednesday, March 10, 2004
Speaker: **Charles Peskin**, NYU
Title: Random walks of some biomolecular motors

Wednesday, March 17, 2004
Speaker: **Donald Hoover**, Rutgers University
Title: Estimating the impact of *P. carinii* prophylaxis on HIV disease

Wednesday, March 24, 2004
Speaker: **Robert Miura**, NJIT
Title: Modelling spreading cortical depression

Wednesday, March 31, 2004
Speaker: **Galit Lahav**, Harvard University
Title: Dynamics of the p53-mdm2 feedback loop in living individual cells, and the design-principles of biological feedback

Wednesday, April 7, 2004
Speaker: **Peter Wolynes**, UC, San Diego
Title: Issues in stochastic cell biology

Wednesday, April 14, 2004
Speaker: **Gustavo Stolovitzky**, IBM, T.J. Watson Research Center
Title: Noise model and statistical significance in MPSS transcription profiling

Tuesday, April 20, 2004
Speaker: **Gerald Manning**, Rutgers University
Title: Counterion condensation and DNA self-assembly

Wednesday, April 21, 2004
Speaker: **Atul Narang**, University of Florida
Title: Gradient sensing by eucaryotic cells: Signaling mechanisms and spatiotemporal dynamics

Wednesday, April 28, 2004
Speaker: **Hal Smith**, Arizona State University
Title: Gene transfer in biofilms: a mathematical model

Wednesday, May 5, 2004
Speaker: **Mike Tyers**, Samuel Lunenfeld Research Institute, Canada
Title: How big is big enough? Cell size and commitment to division in budding yeast

Wednesday, May 12, 2004
Speaker: **Vladimir Pavlovic**, Rutgers University
Title: Mathematical models for comparative genomics---successes and challenges

Wednesday, May 19, 2004

Speaker: **Herbert Hethcote**, University of Iowa
Title: Comparison of five new pertussis vaccination strategies

Wednesday, May 26, 2004
Speaker: **Steven H. Kleinstein**, Princeton University
Title: Estimating hypermutation rates during immune responses

FINDINGS – Data Mining and Epidemiology

This meeting brought together researchers from a variety of areas: epidemiologists, statisticians, data miners and others. The aim was to foster greater communications between these disparate groups, to help data mining experts understand concrete epidemiological questions, and to help epidemiologists understand better the results possible from using data mining techniques. Several collaborations have already resulted from interactions at the meeting and after, and publications from talks given at the meeting are in progress. A report of this meeting, detailing some of the questions posed as challenges to the area, is in preparation.

Epidemiology is an observational science that concerns itself with finding and explaining patterns of health and disease in populations, usually of humans, but also populations of animals, insects and plants. Data mining is an active area of research interested in finding algorithms for describing latent patterns in often very large data sets. This Working Group has the objective of fostering collaboration between these two disciplines. In March of 2004 it organized a two-day meeting at DIMACS to bring these two groups together in a format designed to initiate such collaborations.

Each day there were a small number of presentations from epidemiologists, statisticians and computer scientists. They gave a detailed description of a specific study analyzed by methods current in today's epidemiological practice. The studies were selected to represent a range of typical techniques such as exploratory factor analysis, multiple logistic regression, proportional hazards models, generalized additive models and other current methods. The epidemiologists concentrated on the epidemiological aspects of the study, with some relatively brief explanations of the main ideas by the statisticians. Computer scientists or discrete mathematicians commented on algorithmic issues and alternate approaches to the underlying questions. This led into a moderated audience-wide discussion of each study.

The goal was to identify some fruitful areas for applying new techniques from theoretical computer science, discrete mathematics, and statistics and to allow new collaborations among participants to be initiated. This working group brought together with epidemiologists experts interested in exploring data mining and algorithmic techniques potentially useful in epidemiology from a range of viewpoints. In order to facilitate a synthesis of perspectives, the working group was organized around half-day sessions, each with one or two talks followed by active discussion. The five primary topics of inquiry were:

Disease Dynamics, Epidemiological Studies and the Interpretation of Data

Genetic, Pharmaceutical and Environmental Epidemiology
Elements of Statistical Learning and Models
Regression Methods, Support Vector Machines and Cluster Analysis
Association Rules and Lattice/Graph Theoretical Methods

The working group meeting included a problem section and a report on the first Epidemiological DIMACS Challenge that began in September 2003. Through short presentations and discussions that included all participants, the workshop aimed to foster research cooperation among Epidemiologists, Computer Scientists and Mathematicians. The goal was to identify fundamental epidemiological problems that can benefit from efficient computational, statistical and mathematical models that can aid in the processing and understanding of combined epidemiological, and immunological data.

An important outcome of the meeting was the identification of data mining methods appropriate for the early indication of plant diseases. This has led to a collaboration between DIMACS and the National Plant Diagnostic Network headquartered at UC Davis.

Working Group on Methodologies for Comparing Vaccination Strategies, May 17-20, 2004

Organizers:
John Glasser, CDC
Herbert Hethcote, University of Iowa

This working group focused on the use of mathematical modeling and computer simulations to provide theoretical comparisons of vaccination and other intervention programs. Examples of the effective use of modeling were studied for vaccine-preventable diseases such as measles, rubella, varicella, pertussis, polio, Haemophilus influenza type b and pneumococcal, Hepatitis A and B, and Lyme disease.

When vaccination strategies compare vaccination of different age groups, age-structured mathematical models are necessary. Computer simulations of these highly complex systems with hundreds of demographic and epidemiological classes allow comparisons of vaccination strategies and rapid evaluation of potential intervention programs when field trials would be prohibitively expensive. Both expected and worst case scenarios must be considered. We brought working epidemiologists and policymakers together with experts in mathematical modeling and computer simulation in order to formulate and analyze models to deal with current problems and explore ways to simulate larger, more complex, more realistic models.

There was a good deal of discussion about influenza. The persistence of influenza depends on its ability to evolve so that new strains and subtypes of the virus appear and old ones reappear. This constant evolution means that vaccines need to be updated frequently and that resistance to drug therapies can easily arise. These issues make modeling difficult and require large, powerful computational methods to analyze the models. Some specific research issues to which the group applied its methods are the following.

How will varicella vaccination affect the incidence of shingles?

Is acellular pertussis vaccination of adolescents or adults more effective in reducing cases among infants than vaccination of children?
Because the oral polio vaccine virus can mutate back to the wild virus, as occurred recently in the Caribbean and the Philippines, can polio be eradicated using the oral polio vaccine alone?
How does the yearly composition of the influenza vaccine affect the drift of the influenza virus?

There was also discussion of vaccination strategies in the event of or in anticipation of bioterrorist attacks involving smallpox, anthrax, or influenza. Research issues here include:

What are the risks/advantages of different smallpox vaccination strategies using existing supplies of the vaccine?
How do the conclusions change if we dilute existing supplies so as to allow more people to be vaccinated but can only achieve a certain probability of protection or only a reduction of symptoms to treatable levels?
How do vaccination strategies depend upon our ability to identify a spot smallpox epidemic?
How do these conclusions change for diseases such as anthrax where transmission is not person-to-person?

DIMACS/IAS Workshop on Genomic Instability in Cancer: Biological and Mathematical Approaches, June 8 - 9, 2004

Organizers:
Natalia Komarova, Rutgers University and Institute for Advanced Study, Princeton,
Arnold Levine, Institute for Advanced Study, Princeton, and Robert Wood Johnson School of Medicine and Dentistry

Many cancers are characterized by genomic instability. This often involves gross chromosomal abnormalities such as deletion and duplication of chromosomes or chromosome parts, chromosomal rearrangements and mitotic recombinations, generally termed as ``chromosomal instability''. Other types of genomic instability are characterized by an increased rate of small-scale genetic changes (such as microsatellite instability). The role of genomic instability (GI) for cancer progression is a very important, yet unresolved question.

Mechanisms of genomic instability are intimately related to many aspects of cell biology, such as cell cycle regulation, DNA damage and repair, cell aging and telomere function. On the other hand, the role of GI in cancer initiation and progression can be studied from the point of view of somatic evolution of cells, where mathematical modeling can play an important part. The goal of this workshop is to bring together people who study the phenomenon of GI from different perspectives, for extended discussions and learning. In particular we hope to start a dialog between experimental scientists and mathematical modelers and identify areas for cross-pollination between the two communities.

Theoretical approaches in the area of GI and cancer progression are still in the budding stage. However we believe that quantitative methods can provide an important tool in deciphering the mystery of GI. This is similar in spirit to studying infectious diseases, where mathematicians and

biologists working hand by hand have generated many useful insights. We expect that a similar phenomenon will happen in the area of cancer research. This workshop is the first attempt to create a forum for experimental biologists and modelers of genomic instability, which will hopefully encourage collaboration and improve our understanding of cancer.

DIMACS/BIOMAPS Seminar in Quantitative Biology and Epidemiology:

Graham Cormode and Patrick De Leenheer, post-doctoral research fellows at DIMACS, in association with members of the BioMaps research group, organized the joint DIMACS/BIOMAPS Seminar Series on Quantitative Biology and Epidemiology during the 2003-2004 academic year. In total, there were almost forty talks given in the first year of this joint venture. The audience came from a wide variety of disciplines, including Biology, Mathematics, Medicine, Physics, Chemistry, Computer Science and more, including regular participants from Rutgers, The University of Medicine and Dentistry of New Jersey (UMDNJ), Robert-Wood Johnson Medical School (RWJMS), Princeton and beyond. The program of this series, including titles and speakers is given here. Abstract information can be found at:

http://dimacs.rutgers.edu/SpecialYears/2002_Epid/seminars.html

Wednesday, September 17, 2003
Speaker: **Marcelo Magnasco**, Rockefeller University
Title: The dynamical basis of auditory acuity

Wednesday, September 24, 2003
Speaker: **Monica Driscoll**, Rutgers University
Title: Elaborating cellular and molecular mechansims of aging in C. elegans: lessons and themes from simple old animals

Wednesday, October 1, 2003
Speaker: **Joshua Weitz**, Princeton University
Title: Scaling laws in organismal biology: Navigating the morass

Wednesday, October 15, 2003
Speaker: **Boris Kholodenko**, Thomas Jefferson University
Title: Molecular and mechanistic analyses of cellular networks: Can we navigate through molecular jungles

Wednesday, October 22, 2003
Speaker: **Stanislav Shvartsman**, Princeton University
Title: Modeling and manipulating EGFR-mediated cell communication in development

Wednesday, October 29, 2003
Speaker: **Sergei Grigoryev**, Penn State University College of Medicine
Title: Higher-order chromatin structure and its implications for gene regulation

Wednesday, November 5, 2003
Speaker: **Ira Schwartz**, Naval Research Laboratory
Title: Non-equilibrium stochastic outbreaks: From prediction to control

Wednesday, November 12, 2003
Speaker: **Jonathan Dushoff**, Princeton University
Title: Detecting selection from patterns of codon usage

Wednesday, November 19, 2003
Speaker: **Burkhardt Rost**, Columbia University
Title: Evolution teaches predicting protein function?

Tuesday, November 25, 2003
Speaker: **Ulf Dieckmann**, IIASA, Austria
Title: Spatial complexity in ecology and evolution

Wednesday, November 26, 2003
Speaker: **Horst Thieme**, Arizona State University
Title: Competitive dynamics in a model for river blindness with cross-immunity

Wednesday, December 3, 2003
Speaker: **Natalia Komarova**, Rutgers University
Title: Modeling DNA transcription dynamics

Wednesday, December 10, 2003
Speaker: **Reka Albert**, Penn State University
Title: Topology and robustness of the segment polarity gene network

Wednesday, December 17, 2003
Speaker: **Philip Holmes**, Princeton University
Title: Piecewise-holonomic mechanics, hybrid dynamical systems, and escaping cockroaches

Wednesday, January 21, 2004
Speaker: **Laura Finzi**, University of Milan
Title: Topological modifications of DNA in transcriptional regulation

Wednesday, January 28, 2004
Speaker: **Yuhai Tu**, IBM Research
Title: Receptor cross talk and signal amplification: some recent results on signal transduction in bacterial chemotaxis

Thursday, February 5, 2004
Speaker: **Santiago Schnell**, Oxford University
Title: Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws

Tuesday, February 10, 2004
Speaker: **Peter Thomas**, Salk Institute
Title: Inside the mind of the amoeba: Simulation and analysis of biochemical signal transduction channels

Thursday, February 12, 2004
Speaker: **Stanislav Shvartsman**, Princeton University
Title: Modeling and manipulation of EGFR signaling in Drosophila development

Wednesday, February 18, 2004
Speaker: **Robijn Bruinsma**, UCLA
Title: The physics of cell adhesion and mechano-sensing

Wednesday, February 25, 2004
Speaker: **Rahul Siddharthan**, Rockefeller University
Title: PhyloGibbs: A Gibbs sampler for motif detection incorporating phylogenetic information

Wednesday, March 3, 2004
Speaker: **Gerard Manning**, Rutgers University
Title: Counterion condensation on polyelectrolytes: Theory and applications

Wednesday, March 10, 2004
Speaker: **Charles Peskin**, NYU
Title: Random walks of some biomolecular motors

Wednesday, March 17, 2004
Speaker: **Donald Hoover**, Rutgers University
Title: Estimating the impact of *P. carinii* prophylaxis on HIV disease

Wednesday, March 24, 2004
Speaker: **Robert Miura**, NJIT
Title: Modelling spreading cortical depression

Wednesday, March 31, 2004
Speaker: **Galit Lahav**, Harvard University
Title: Dynamics of the p53-mdm2 feedback loop in living individual cells, and the design-principles of biological feedback

Wednesday, April 7, 2004
Speaker: **Peter Wolynes**, UC, San Diego
Title: Issues in stochastic cell biology

Wednesday, April 14, 2004
Speaker: **Gustavo Stolovitzky**, IBM, T.J. Watson Research Center
Title: Noise model and statistical significance in MPSS transcription profiling

Tuesday, April 20, 2004
Speaker: **Gerald Manning**, Rutgers University
Title: Counterion condensation and DNA self-assembly

Wednesday, April 21, 2004
Speaker: **Atul Narang**, University of Florida
Title: Gradient sensing by eucaryotic cells: Signaling mechanisms and spatiotemporal dynamics

Wednesday, April 28, 2004
Speaker: **Hal Smith**, Arizona State University
Title: Gene transfer in biofilms: a mathematical model

Wednesday, May 5, 2004
Speaker: **Mike Tyers**, Samuel Lunenfeld Research Institute, Canada
Title: How big is big enough? Cell size and commitment to division in budding yeast

Wednesday, May 12, 2004
Speaker: **Vladimir Pavlovic**, Rutgers University
Title: Mathematical models for comparative genomics---successes and challenges

Wednesday, May 19, 2004
Speaker: **Herbert Hethcote**, University of Iowa
Title: Comparison of five new pertussis vaccination strategies

Wednesday, May 26, 2004
Speaker: **Steven H. Kleinstein**, Princeton University
Title: Estimating hypermutation rates during immune responses

Findings:

FINDINGS - Statistical Measures for Syndromic Surveillance

Marianne Frisén, Statistical Research Unit, Göteborg University, Sweden has developed important new collaborations and new research ideas from her participation in the special focus. She got new insight into the need for proper evaluation measures. In syndromic surveillance there is a need for continual observation of one or more time series, with the goal of detecting an important change in the underlying process as soon as possible after it has occurred. Statistical methods are necessary to separate important changes from stochastic variation. The statistical methods suitable for this differ from the standard hypothesis testing methods. Also the measures for evaluation differ. Frisén has developed statistical measures for evaluation of prospective surveillance. Most of the commonly used methods are optimal in some respect. Different criteria of optimality are used in different subcultures of statistical surveillance. Frisén demonstrated the shortcomings of some criteria of optimality. The most commonly used optimality criterion is based on the average run length, ARL. Another criterion is based on a utility function. A third criterion is that of minimax. Evaluation of methods for syndromic surveillance in practice is very important. It is necessary to know the basic properties of a system before it is implemented,

including its statistical properties.

FINDINGS – Epidemiological Modeling

Karl Hadeler, University of Tűbingen, with his collaborators, has successfully completed three projects that were stimulated by his participating in this special focus.

A) In epidemic modeling and in designing control measures one can either look at the onset of an outbreak, i.e. at the basic reproduction number of the infected cases in a susceptible population, or at the maximal prevalence of an epidemic or at the prevalence in an endemic situation. Although these quantities are unrelated for arbitrary transmission laws, there are simple functional relations for the classical transmission kinetics that Hadeler and Lourdes Esteva used to estimate parameters from data.

B) By tradition, there are two ways of modeling the spread of infectious diseases in a spatially distributed population, by contact distributions or by diffusion equations. In connection with possible deliberate release of infectious agents, both approaches have been discussed, and the question of the relation between them and the appropriate scaling of parameters has been raised. To answer these questions, Hadeler has taken a unified approach in the form of a model for three types of population: susceptible individuals, sedentary infected individuals (motion restricted to local area) and migrating infected individuals (wide range motion). He showed that by taking appropriate limits in parameter space one arrives either at diffusion equations or at contact distributions whereby the latter can be seen as describing rapid excursions of highly infectious long-distance travelers.

C) Starting from the epidemics scenario, Hadeler and Mark Lewis have extended the idea of distinguishing sedentary and migrating phases to ecological problems. Hadeler and Thomas Hillen then extended the idea and to quiescent phases in dynamic systems in general.

FINDINGS – Sensor Networks

Graham Cormode, Martin Farach-Colton, Rohan Fernandes, Miguel Mosteiro and Muthu Muthukrishnan have worked on a variety of mathematical optimization problems relating to the placement and organization of sensor networks with particular application to tracking the spread and transmission of disease and other epidemic-like phenomena. They have gotten results bounding the time of communication of events in these networks, and shown how to compute maximum likelihood estimators for readings with errors between multiple sensors. (Farach-Colton and Muthukrishnan are senior faculty in the project, Cormode a project postdoc, Fernandes a project graduate student and Mosteiro a graduate student in Farach-Colton's group.)

Sensor networks, that is, networks of very inexpensive sensors distributed at random, are likely to become an important tool for epidemiological data gathering, for example when the sensors detect bioactive agents. The cheaper the sensors, the easier they will be to deploy. But cheap sensors have many limitations -- low transmission power, no global positioning systems, low memory. Thus, they don't know where they are when they get thrown out of an airplane, and they can only talk to a few of their nearby neighbors.

Cormode, Farach-Colton, Fernandes, Mosteiro and Muthukrishnan, have studied how to find highly connected low-degree subnetworks of sensor networks in sensors randomly distributed in 2-space. They have preliminary results that show that such networks must exist, but their algorithm for constructing them relies on sensors knowing where they are.

The second thrust of their work has been in algorithms for having sensors compute their position from knowledge of the approximate distance to their close neighbors. They have shown that sensor networks, whose distance measurements form distance threshold graphs (that is, you know the distances to nearby neighbors but have no information of distances to distant neighbors) have interesting rigidity properties in two dimensions, which means that the problem of finding the position of each sensor given noisy distance measurements is highly constrained. They are in the process of speeding up their algorithm for finding the position of each sensor.

FINDINGS – New Approach for Comparative Poisson Trials

Comparative Poisson trials of prophylactic interventions, such as vaccines, can be lengthy and costly. Donald Hoover (senior faculty project member) evaluated two easily implemented approaches to reduce numbers of disease cases and person years of follow up for comparative Poisson trials with fixed numbers of cases i) altering the portion allocated to treatment, and ii) curtailed stopping if numbers of cases in the treatment or control group indicate that the null hypothesis has already been rejected or will not be rejected.

FINDINGS – Epidemiology and Graph Theory

Alain Franc, Institut National de la Recherche Agronomique (INRA), Département Ecologie des Forêts, Prairies et Milieux Aquatiques, CDA & UMR Biogeco, Bordeaux, found that his participation in the special focus led to a new involvement in theoretical epidemiology, with an emphasis on graph theory. Metapopulation dynamics and epidemiology of some diseases on sessile hosts (like pathogen fungi / tree interactions) can be modeled by contact processes on a graph. There is recent work on the structure of a graph, with classical notions like degree, diameter, small world property, etc., but less work on dynamics on graphs. Franc modeled time discrete contact process on graphs with known degree distribution, with local mortality and spread along the edges with constant diffusion coefficient. He showed that the knowledge of the degree distribution only leads to a mean field approximation of the contact process. The fraction of infected hosts (or occupied patches) is modeled as a function of the degree of the vertex. As a consequence of mean field approximation, the stationary distribution at equilibrium belongs to a universal family of curves, which is independent of the degree distribution. He gave an analytic description of this family. Each curve in the family is identified by one parameter only, which can be the mean infection rate (or occupancy rate) for the whole graph, obtained as a solution of an implicit equation containing the degree distribution. Finally, he found an exact equation relating the density of occupancy and the probability for types of edges, as a first step towards pair approximation on graphs.

NEW COLLABORATION – Spatial Modeling of Foot-and-Mouth Disease

The Spatio-temporal modeling meeting has resulted in a new, funded collaboration between Rowland Kao, University of Oxford, and D.T. Haydon (formerly University of Guelph, now Glasgow) on epidemic tree analysis and spatial modeling of foot-and-mouth disease. The grant is for £48,800 over one year, funded by DEFRA (Department of the Environment, Food and Rural Affairs in the UK).

Initially, control of foot and mouth disease (FMD) in the 2001 epidemic followed traditional methods: restriction of animal movements, the rapid detection, removal of livestock, and decontamination on infected premises (IPs), and the rapid detection and removal of all livestock on premises believed to have been in potentially infectious contact with known IPs (dangerous contacts or DCs). When it became clear that these measures as implemented were proving insufficient to control the epidemic, modifications to the policy were made. In particular, efforts to rapidly remove all livestock on premises proximate to IPs were prioritized over the cull of DCs, by reason of a statistically determined risk of having been or becoming infected – this policy became known as the contiguous premises (CP) cull. Soon after the policy changes were implemented, the epidemic began to decline; however, the policy changes became a highly contentious issue. In particular, there were strongly held arguments that diverting resources from the cull of DCs, compromised an important part of the control efforts and resulted in needless loss of livestock on CPs at no real risk of infection.

Kao and Haydon propose to analyze the relative impact of the DC and CP culls for the various regional epidemics following the imposition of national movement restrictions on Feb. 23, 2001. In order to evaluate the true value of removing a DC, the number and distribution of potentially infectious links between premises must be considered. This can be represented as a map running forward through time of the connections from all IPs, henceforth called the "contact tree". While impossible to recover in full, it should be possible to infer critical information about the contact tree, in particular, the very broad question of whether transmission of infection is a result of many potentially infectious contacts with relatively low probability of infection (a dense contact tree), or relatively few potentially infectious contacts with relatively high probability of infection (a sparse contact tree). The principal data set that will be used is the inferred contact structure between IPs only (i.e. the "epidemic tree"), which is the subset of all potentially infectious links resulting in infection on the target premises. In combination with demographic information taken from a variety of data sources, the epidemic tree will be used to infer the sparseness of the contact tree. The approach used will be a combination of computer simulation and statistical analysis, the latter based on established techniques from the similar study of phylogenetic trees in population genetics. Recommendations will be made, first about the relative value of the DC and CP culling policies, and also about the type of data that could be collected both before and during a disease outbreak to help characterize the contact tree and its role in disease transmission.

INTERNATIONAL COLLABORATION – Profiling Health Care Providers

Guido Consonni, a visitor at DIMACS**,** has worked on profiling health care providers with senior project faculty member David Madigan, Statistics Department, Rutgers University. This problem is important in order to compare performances of health care providers. One issue that is of special concern is adjusting for case mix, which varies across health providers. Another issue

pertains to uncertainty in the evaluation process. They have used a Bayesian approach and have set up a hierarchical model. In this way they are able to borrow strength across providers, which is especially useful if only limited data are available on a specific provider. Concentrating on an outcome that is dichotomous, they set up a logistic regression model and defined two adjusted functional indexes of health performance, which they named Provider and Population, each representing the probability of the desired outcome as a function of patients' covariates. In this way they are able to better differentiate among Providers. This also provides a basis for policy evaluation. They expect to apply this methodology to a data set available though Madigan relating to a set of hospitals in Pennsylvania.

INTERNATIONAL COLLABORATION – Discrete Graphical Models

Guido Consonni, a visitor at DIMACS**,** is working on discrete graphical models with Alberto Roverato, University of Modena and Reggio Emilia, Italy, also visiting DIMACS.

Graphical models and Bayesian Networks are an important area of modern research. In particular they are used to encode the joint distribution of many variables and as such have been used in a variety of contexts (e.g. probabilistic expert systems). In particular they have focused on discrete Bayesian networks where each variable is categorical. The structure underlying a graphical model is an undirected graph from which one can read the conditional independence properties of the joint distribution. They aimed at providing a unified and general framework for the description, analysis and interpretation of such discrete graphical models. To this end, they assumed for simplicity that all variables are binary and that the joint distribution exhibits no conditional independencies (equivalently the graph is complete). Having identified a specific set of canonical statistics, and corresponding canonical parameters, they developed a general set-based notation to explicate such quantities in the joint exponential family density and develop relations, based on the Moebius inversion formula, among three distinct model parameterizations (mean, standard and canonical). They computed the variance function and two specific Jacobians that are useful to recover it. They believe that this preliminary work will prove to be essential for later developments in the field with reference to decomposable graphical models: in particular computation of the information matrix and prior distributions for the parameters of graphical models (and properties thereof).

This research has had a strong impact on Consonni's research. He is currently working on finding prior distributions for contingency tables with a general conditional independence structure that can be represented by a decomposable model.

FINDINGS – Syndromic Surveillance and Public Health

Michael A. Stoto, RAND Center for Domestic and International Health Security, is beginning a project with the Department of Health in the District of Columbia looking at their syndromic surveillance program. The February DIMACS meeting on Adverse Event/Disease Reporting, Surveillance, and Analysis was very useful to him in terms of getting ideas for the multivariate analysis of these data. To aid the early detection of bioterror events, public health officials and researchers have developed a new method called "syndromic surveillance," which involves collecting and analyzing statistical data on health trends–such as symptoms reported by people

seeking care in emergency rooms or other health care settings. By focusing on symptoms rather than confirmed diagnoses, syndromic surveillance aims to detect bioterror events earlier than would be possible with traditional disease surveillance systems. Although syndromic surveillance is new and still largely untested, many public health agencies are investing substantial sums to develop and implement these systems. Despite the potential, however, their analysis shows that the benefits of any syndromic surveillance system will depend on how effectively it is integrated into public health practice. Moreover all syndromic surveillance systems face inherent trade-offs (among sensitivity, timeliness, and the number of false positives) that limit their effectiveness, and it must be concluded that their benefits have not yet been clearly established. Syndromic surveillance, however, may turn out to have greater benefits for other areas of public health. He and his collaborators have submitted an abstract to the American Public Health Association's meeting in DC, Fall, 2004.

FINDINGS – Chemostat Models

Patrick De Leenheer, post-doctoral fellow at DIMACS, has initiated several collaborations during his stay at DIMACS. All of them involved his post-doc advisor, Eduardo Sontag from Rutgers. The first result was obtained from joint work with senior project organizer Simon Levin from Princeton and Christopher Klausmeier from Georgia Tech. They considered chemostat models (a chemostat is a biological reactor in which one or more species compete for one or more nutrients) and studied the dynamics of a single species, limited by two nutrients, assuming that nutrient uptake and growth are decoupled. For a broad class of uptake and growth functions they showed that a nontrivial equilibrium might exist. Moreover, if it exists it is unique and globally stable, generalizing a previous result by Legović and Cruzado. The resulting publication is listed in the Papers section of this report.

A second result followed from collaboration with David Angeli from the University of Firenze, Italy. They considered again a chemostat model, but this time an arbitrary number of species was assumed to compete for a single nutrient. Also, here they idealized the conversion of nutrients into new biomass, in contrast to the previously described work. A distinct feature of this model however -one that sets it apart from the traditional chemostat models- is that crowding effects are taken into consideration. The model can be rewritten as a negative feedback interconnection of two systems that are monotone (as input-output systems). Moreover, these subsystems behave nicely when subject to constant inputs. This allows the use of a particular small-gain theorem that has recently been developed for feedback interconnections of monotone systems. Both global stability and coexistence results have been obtained (the latter holds if crowding effects are large enough). Their coexistence result should be compared to the classical chemostat theory where coexistence does not occur. The resulting publication is listed in the Papers section of this report.

De Leenheer, Angeli, and Sontag were invited to write a tutorial paper for the 16th Mathematical Theory of Networks and Systems conference, to be held in Leuven, Belgium, Summer, 2004. This publication is listed in the Papers section of this report. They reviewed some of the main features of monotone systems, highlighting some application areas (mainly in biology, chemistry). A convergence result for solutions of monotone systems was given with a very short and self-contained proof. Ordinarily such proofs are rather lengthy and require knowledge of a

number of fundamental results from the theory of monotone systems. As an immediate application they considered the behavior of a general class of chemical reaction networks with monotone but otherwise arbitrary reaction kinetics. The main result is that every solution converges to a steady state. In addition they showed that the inclusion of diffusion effects in the models does not alter this conclusion.

FINDINGS – Persistence of Influenza

The research of Simon Levin, Jonathan Dushoff, and Joshua Plotkin on influenza epidemiology was facilitated by this special focus. The evolution of the etiological agents of disease presents one of the greatest challenges for their control, and makes essential complementing standard epidemiological investigations with broader approaches that allow for evolutionary change. Given the stunning genetic diversity that is possible for many such agents, such as the influenza virus, it is impossible to represent all of the diversity manifest at the level of amino acid sequences. Levin, Dushoff, and Plotkin show that drift-variant influenza strains naturally cluster into groups that are associated with functionally important epitopic regions. Dominant clusters typically replace each other every 2–5 years, and this feature is fundamental to the development of vaccination strategies. They furthermore show that stochastic fluctuations can greatly magnify small interference effects among strains, or even among subtypes, leading for example to competitive exclusion in situations where such effects would be unexpected based on the usual deterministic models. They suggest that this effect might be involved in the explanations of some persistent empirical anomalies.

FINDINGS – Antibiotic Resistance

Senior project organizer Simon Levin, Jonathan Dushoff (DIMACS visitor), Dave Smith (DIMACS visitor), Eli Perencevich and Gerry Harris obtained results on antibiotic-resistant bacteria. Infections with antibiotic-resistant bacteria (ARB) in hospitalized patients are becoming increasingly frequent despite extensive infection-control efforts. Infections with ARB are most common in the intensive care units of tertiary-care hospitals, but the underlying cause of the increases may be a steady increase in the number of asymptomatic carriers entering hospitals. Carriers may shed ARB for years but remain undetected, transmitting ARB to others as they move among hospitals, long-term care facilities, and the community. Levin, Dushoff, Smith, Perencevich, and Harris applied structured population models to explore the dynamics of ARB, addressing the following questions: (i) What is the relationship between the proportion of carriers admitted to a hospital, transmission, and the risk of infection with ARB? (ii) How do frequently hospitalized patients contribute to epidemics of ARB? (iii) How do transmission in the community, long-term care facilities, and hospitals interact to determine the proportion of the population that is carrying ARB? They offered an explanation for why ARB epidemics have fast and slow phases and why resistance may continue to increase despite infection-control efforts. To successfully manage ARB at tertiary-care hospitals, regional coordination of infection control may be necessary, including tracking asymptomatic carriers through health-care systems.

FINDINGS – Tuberculosis Genetics

Selective pressures on proteins are usually measured by comparing nucleotide sequences. Jonathan Dushoff (DIMACS visitor), Joshua Plotkin (DIMACS visitor), and Hunter Fraser introduced a method to detect selection on the basis of a single genome sequence. They catalogued the relative strength of selection on each gene in the entire genomes of *Mycobacterium tuberculosis* and *Plasmodium falciparum*. Their analysis confirms that most antigens are under strong selection for amino-acid substitutions, particularly the PE/PPE family of putative surface proteins in *M. tuberculosis* and the EMP1 family of cytoadhering surface proteins in *P. falciparum*. They also identified many uncharacterized proteins that are under strong selection in each pathogen. They provided a genome-wide analysis of natural selection acting on different stages of an organism's life cycle: genes expressed in the ring stage of *P. falciparum* are under stronger positive selection than those expressed in other stages of the parasite's life cycle. Their method of estimating selective pressures requires far fewer data than comparative sequence analysis, and it measures selection across an entire genome; the method can readily be applied to a large range of sequenced organisms.

NEW COLLABORATIONS – Models of the Spread of Tuberculosis and West Nile

In August 2002, two computer scientists, Tom Jacob and Armin Mikler, came to DIMACS for the tutorial "Epidemiology for Mathematical Scientists." They found epidemiology to be a very interesting application possibility for them to move into a new area, and went back to the University of North Texas determined to start serious research in this area. Today, they have a group of three faculty members (the two of them plus Joseph Oppong, a medical geographer), five doctoral students, six master's students and four undergraduates working with epidemiologists and biologists on models of the spread of tuberculosis and West Nile disease. They are also developing a simulator for large communicable disease studies. Currently, this work is being done on a 16-node gigabit ethernet, but they hope to be able to expand it to 64 nodes in the near future. Jacob returned to DIMACS in December 2003 as a full-fledged participant in a research "working group" on health data privacy in which the other major participants included leading researchers from the Centers for Disease Control and Prevention.

COLLABORATION – Data Mining to Find Patterns in Cancer Data

James Abello (DIMACS visitor), Ilya Muchnik, Dona Schneider, Alex Pogel (DIMACS visitor), David Ozonoff (senior project faculty member) and Frank Wong (DIMACS visitor) are in the initial stages of a collaboration to explore the application of state-of-art Data Mining techniques to SEER Cancer Data. The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute is an authoritative source of information on cancer “incidence” (newly diagnosed cases) and “survival” (proportion of patients alive at some point subsequent to the diagnosis) in the United States. Updated annually, SEER currently collects and publishes cancer incidence and survival data from 11 population-based cancer registries and three supplemental registries covering approximately 14% of the US population. For example, in the SEER 12 Registry Database, cases diagnosed from 1992 through 2000 are available, one record for each of 1,559,500 tumors. SEER data have been widely utilized to identify and understand the reasons for geographic and population differences in cancer patterns, to investigate the influence of environmental factors on cancer incidence and survival, and to study cancer treatment outcomes. Standard epidemiologic techniques for exploring SEER data include comparing a variety of rates

and ratios and using regression methods. More recently, the use of geographic information systems (GIS) has been used to evaluate clustering. Data Mining is the mechanized process of identifying or discovering useful structure in data. The resultant structure (and associated confidence level) serves as triggers to stimulate further special studies with additional data and information. It is important to demonstrate the usefulness of data mining techniques to epidemiologists by exploring a data set typically analyzed with standard epidemiologic methods. By demonstrating the utility of data mining with such a data set, they may effect a paradigm shift, enabling epidemiologists to expand their *modus operandi* to include examining underlying data structure. If they are successful in providing new insights, it would offer a novel approach that could be made routine for the effective use of SEER and possibly other health data sets. Expansion of knowledge related to health data sets may even lead to better public health and health care policy.

FINDINGS – Saturation Effects in SIS Epidemics in Heterogeneous Populations

Project graduate student Jaewook Joo and Professor Joel L. Lebowitz have studied the behavior of susceptible-infected-susceptible (SIS) epidemics. They investigated saturation effects in SIS models of the spread of epidemics in heterogeneous populations. The structure of interactions in the population is represented by networks with a given connectivity distribution. Considering cases where the transmission of infection between nodes depends on their connectivity, they introduced a saturation function that reduces the infection transmission rate across an edge going from a node with high connectivity. A mean field approximation with the neglect of degree-degree correlation then leads to a finite threshold value. They investigated via computer simulation the contact process on a heterogeneous regular lattice and compare the results with those obtained from mean field theory with and without neglect of degree-degree correlations.

Joo and Lebowitz also investigated the time-evolution and steady states of the stochastic susceptible-infected-recovered-susceptible (SIRS) epidemic model on one- and two- dimensional lattices. They compared the behavior of this system, obtained from computer simulations, with those obtained from the mean-field approximation (MFA) and pair-approximation (PA). The former~(latter) approximates higher order moments in terms of first~(second) order ones. They found that the PA gives consistently better results than the MFA. In one dimension the improvement is even qualitative. See the Paper section for their publications.

FINDINGS – Lattice Theory Applied to Epidemiology

As described in the annual report for year one, Alex Pogel became involved in the DIMACS' Special Focus on Computational and Mathematical Epidemiology through the efforts of David Ozonoff, at the DIMACS Summer School Tutorial titled "Epidemiology for Mathematical Scientists", held in August 2002. This led to a continuing collaboration with Ozonoff. In this research, they have introduced two new concepts in epidemiology, the generalized and closed set contingency tables, based on mathematical foundations in lattice theory. As the name suggests, this is a generalization of the conventional contingency table or cross-classification. Their further development of this idea is leading to some new research directions in epidemiology and new methods of data analysis. They have submitted two external funding proposals (with more to come).

FINDINGS – Bigraph Decomposition and Concept Lattices

In May 2003, after Alex Pogel (DIMACS visitor) made a presention to the Working Group on Data Mining and Epidemiology, an impromptu discussion with James Abello (DIMACS visitor) led to a research effort involving various bigraph decomposition methods (as Abello had applied at AT&T Labs) and their connection with the concept lattice. Their submission of a small part of the research effort (level-aware breadth first search and concept lattices) has been accepted for publication by the Journal of Universal Computer Science. Another paper, with greater scope than the first, is in development and should be completed near the end of the summer, 2004, and supporting software is being developed at this time. They are using the SEER Cancer registry database as a test case, as they develop data analysis methods that combine a variety of bigraph decomposition methods -- the iterated peeling construction that reduces a graph to its core, cutsets of bipartite graphs, and level-aware breadth first search (with its connection with the concept lattice loosened, so that it can interact with the other two methods just mentioned) -- and the concept lattice view of the data. The basic idea is to use the decomposition methods for overview, and the lattice view for zooming. The main work is to understand the handoff, i.e. how to take the information provided by a particular decomposition, and use it in computing and viewing a concept lattice. Overall, the hope is that a variety of structural views of the data should help guide the user in analysis. Advances in this direction will probably be followed by work on how to also use the lattice for overview.

FINDINGS - New Factors in Surviving Prostate Cancer

Modern computer science methods, particularly, machine learning predictor design open a unique opportunity to discover epidemiological factors which play critical roles in mortality and survival time distributions. It is well known how hard it is to estimate the distributions with a high confidence. Particularly, it is hard when the final goal is to estimate the influences of some health environment modifications on changes of the distributions. The traditional epidemiological approach uses classical tests of statistical hypothesis that two distributions (before and after the above mentioned modifications) are equal. Unfortunately, available data is so limited that these tests cannot work. The method proposed in the research of Ilya Muchnik (DIMACS visitor) avoids this limitation. To compare badly estimated distributions, he considers the corresponding random variable (usually, boolean) as a function of many epidemiological and medical arguments. An estimate of the function becomes a prediction model. The key idea is that such a function estimate has a high confidence even when the available data is very limited.

Using this method on data of about 50,000 patients with prostate cancer who died from the disease (data was received from SEER database: http://seer.cancer.gov/), Muchnik found that race and marital status showed twice the influence on survival time as age. The significant role of race is a well-known epidemiological fact. At the same time the significant role of marital status (and, more that its significance is greater than that of age) is a new observation, which needs to be investigated from different epidemiological aspects. In other words, the observation is a new hypothesis that has to be tested.

Extending the same analysis on all main types of cancer diseases (approximately ten types) for

which SEER data is available may bring several similar interesting novel observations-hypotheses. Muchnik plans to do that study. He hopes that results of this analysis will be useful for many practitioners who regularly work with SEER data. He plans to support the practitioners by a complete such analysis on SEER data, and, more,
to make available the developing software. The intention is to make the software user friendly so that users can work with the SEER data independently.

FINDINGS – Data Mining in the Chicken Industry

Ilya Muchnik (DIMACS visitor) and Kenton Morgan (DIMACS visitor), funded in part by this grant and in part by a new grant they have received, are applying data mining techniques to do an epidemiological study of the chicken industry in United Kingdom. The objects for the study are 600 chicken farms cross the UK. (out of the total of 960). A single farm is described by about 150 variables, most of which are nominal, some are ordinal and numerical. The data has a large number of missing values. The main problem is to discover which variables have maximum influence/effect on the distribution of basic chicken disease syndromes, such as wet litter and coccidiosis. Muchnik and Morgan plan to try to solve the problem using the novel semi-supervised clustering method, which was developed at DIMACS. For this particular project they have adapted the method by taking into account non-standard "heterogeneous" space of data (nominal, ordinal and numerical variables together). The critical new idea in their semi-supervised clustering method and its "ideal" correspondence to the project problem is a particular construction at the foundation of the method, the linkage function, which estimates the "significance" of a subspace of variables on the "separability" of a given object class from other classes. Using such linkage function one could simultaneously find (1) a group of variables defining a space in which space the considered class is the farthest from objects which are not in the class, and, (2) a group of objects from the complementary class which are "mostly close" in a particular sense to the object of the considered class, a "neighborhood" of the class. In the case of the chicken farms, the class is a set of farms with diseased chickens and the group of variables is exactly the set of characteristics of the farms by which the farms are different from other farms which don't have an animal with the considered disease. They see two possible interpretation for the neighborhood of the class: either it is a set of farms that have an diseased animal, but the farmers don't know it yet, or they are farms with "high risk" of becoming "diseased farms." In both cases, this information is not only useful from a practical point of view, but also can help find an answer to the fundamental question: does the extracted group of variables define a "real cause" of considered disease? All the software tools have been developed for this project and the data is in the process of being prepared.

FINDINGS – Using Wireless Sensors to Monitor People *In Loco*

Su Chen, Amit Gaur, S. Muthukrishnan (senior project faculty member), and David Rosenbluth collaborated on using wireless sensors to collect data about people's physiological parameters (pressure, ecg, ekg, accelerations, location, background noise etc.) under natural settings for a prolonged period of time. Such data will be invaluable for a host of epidemiological studies about both people's behavioral as well as physiological patterns, and will shed light on the role of data generation and analysis in epidemiology. Sensors, both large and small, have traditionally been used for measurement and monitoring in a number of basic applications (sensing smoke,

fire, or biohazards, or security etc) using fixed networking infrastructure. But networking them using wireless media gives more dynamism and power: they can be used in remote, autonomous monitoring as well as for mobile environments. Muthukrishnan and his research group used on-person wireless sensors for monitoring physiological, behavioral and health-related data. In particular, they focused on *in loco* monitoring, that is, persistent, unobtrusive monitoring of human factors in *natural situations*. This has many applications in health monitoring, social networking, etc. Their work presents valuable insight into the infrastructure requirements and the key challenges of in loco monitoring and applications that work across the entire lifecycle of data (typically applications deal with one or two aspects in the lifecycle of data): the need to coordinate resources to gather and process multiple data streams; the need to cross reference data streams and associate patterns in behavioral data streams with those in physiological data streams; the need to decompose data streams into application relevant units; maintaining persistence in data collection; dealing with data quality problems; data compression and streaming; etc.

FINDINGS - Discrete Graphical Models

Alberto Roverato, a DIMACS visitor, worked with Valentina Leucari on discrete graphical models. Graphical models and Bayesian Networks are an important area of modern research. In particular they are used to encode the joint distribution of many variables and as such have been used in a variety of contexts (e.g. probabilistic expert systems). In particular, Roverato and Leucari focused on discrete Bayesian networks where each variable is categorical. The structure underlying a graphical model is an undirected graph from which one can read the conditional independence properties of the joint distribution. They aimed at providing a unified and general framework for the description, analysis and interpretation of such discrete graphical models. To this end, they assumed for simplicity that all variables are binary and that the joint distribution exhibits no conditional independencies (equivalently the graph is complete). Having identified a specific set of canonical statistics, and corresponding canonical parameters, they developed a general set-based notation to explicate such quantities in the joint exponential family density and develop relations, based on the Moebius inversion formula, among three distinct model parameterizations (mean, standard and canonical). They computed the variance function and two specific Jacobians that are useful to recover it. They believe that this preliminary work will prove to be essential for later developments in the field with reference to decomposable graphical models: in particular computation of the information matrix and prior distributions for the parameters of graphical models (and properties thereof).

FINDINGS - Association Patterns of Gene Expression Data

Alberto Roverato (DIMACS visitor) is working with David Madigan (senior project faculty member) and Rebecka Jornsten on association patterns of gene expression data. In the analysis of gene expression data, every variable measures the expression value of one single gene in a tissue. Typically, thousands of genes are observed on a small number of tissues and, since the number of observations for each gene is much smaller then the number of genes, the problem of modeling the relations among variables cannot be addressed with the traditional model selection techniques. This is known as the "large p small n" paradigm. Roverato, Madigan, and Jornsten have considered the problem in which p variables have been observed but only q variables at a

time can be handled, where q is much less than p. The main aim of their work is the identification of the conditional independence structure of the p variables by means of an undirected graphical model. During Roverato's visit at DIMACS, they addressed the theoretical aspects of the problem: firstly, they studied the conditions under which it is meaningful to try to recover the conditional independence structure of the p-dimensional joint distribution through the analysis of all the q-dimensional marginal distributions. Secondly, they developed a set of necessary and sufficient conditions leading to the characterization of a class of undirected graphical models for which the q-dimensional marginal distributions allow the faithful reconstruction of the conditional independence structure of the p variables. The next steps of the project involve the implementation of the procedure in a computer program and the validation of the procedure on some gene expression datasets.

FINDINGS - The "Voter Model" with Confidence

Stephen Hartke, graduate student supported by the project, got interested in the use of graphs to model both the spread of opinions and the spread of diseases through social networks. Graph-theoretical models in these areas can be deterministic or stochastic. He studied the voter model, which represents the opinions of individuals or their disease states by 0's and 1's and involves a stochastic process for modifying an individual's opinion or disease state based on the opinions or disease states of his neighbors in the graph. Among the major issues in such models are whether in the long run some sort of consensus situation will occur in the sense that all individuals end up with the same opinion or disease state. This is the case in the classic voter model. What is missing from this basic voter model is the role of the confidence an individual has in his opinion or, equivalently, the resistance an individual has to being infected by a pathogen. Bringing in confidence levels leads to a much more complex mathematical problem. Hartke formulated an intricate "voter model with confidence levels" and obtained extremely nice results about it. He showed that consensus is once again attained and that it is still possible to determine the probability of different common states in the long run. Not only that, but the results are much more interesting because they relate the outcome probabilities to the structure of the underlying graph whereas in the classic voter model, they depend only on the degrees of the vertices. This and the following results of Hartke are part of his Ph.D. thesis in Mathematics at Rutgers under the direction of Fred Roberts, which he will defend in July 2004.

FINDINGS - Competition Graphs and Phylogeny Graphs

Project graduate student Stephen Hartke worked on competition graphs and phylogeny graphs. Competition graphs arise from ecological problems but also have applications in communication, coding, modeling of economic systems, etc. The related phylogeny graphs arise from problems of phylogenetic tree reconstruction and have many analogous properties. The problems of recognizing competition graphs and phylogeny graphs reduce to the problems of computing parameters called the competition number and phylogeny number. These problems are NP-complete, but one of the long-time challenges in the field is to find algorithms that will compute these parameters using an "elimination procedure" that considers one vertex at a time. Suh Kim and Fred Roberts found an elimination algorithm that computed the competition number for "kite-free" graphs. Hartke found a much simpler proof than ours to show that the algorithm works and modified the algorithm to compute the phylogeny number. Kim and Roberts

conjectured that their algorithm would work for all graphs, but Hartke constructed a clever example to show that this conjecture was wrong.

FINDINGS - Competition between Spread of Disease and Attempts to Control It with Limited Resources

Project graduate student Stephen Hartke has been investigating the spread of disease through networks from the point of view of a "competition" between the spread of disease and attempts to control it with limited resources. Specifically, he has been looking at spread of disease from vertices to neighbors in graphs when a public health authority can vaccinate a limited number of people each time period and the disease then spreads from infected people to uninfected and unvaccinated neighbors in the next time period. The goal is to minimize the total number of people who become infected. Hartke's work on this problem has involved the situation where one vertex is infected to begin with and one vertex can be vaccinated each time period and he has shown that in this case, it is NP-complete to find an optimal vaccination strategy. For the special case of trees, Hartke has shown that the greedy algorithm always protects at least half as many people as the optimal algorithm and that this result is tight in the sense that there are graphs where the number of people protected by the greedy algorithm is arbitrarily close to half the optimal.

FINDINGS - Masking Microdata

Joe Fred Gonzalez, Jr. and Lawrence H. Cox, co-organizers for the Working Group on Privacy and Confidentiality of Health Data, one of the invited speakers, Jay J. Kim, National Center for Health Statistics (NCHS), and one of the participants, Myron Katzoff, NCHS, have initiated research and obtained results in a couple of areas dealing with masking microdata. They have studied the effects of rounding continuous data using specific rules. Data such as incomes are frequently rounded. Rounding may be done to protect the confidentiality of records in a file or to enhance readability of the data, or by the notion that the digits subject to rounding are inconsequential. The rounding may not have any effect on the bias of an estimator, but may have a large impact on variance. Integers can be expressed as $x = qB + r$, where q is the quotient, B is the base, and r is the remainder. B is a constant, but q and r are random variables. They use four rules for rounding r to observe the effects of rounding on bias and variance. They assume a uniform distribution on r, but no specific distributional assumption is made on q. When $q = 0$, they show that the variance after rounding is three times the variance before rounding. As the variance of q gets larger, the effect of rounding on the variance decreases. They have computed the disclosure risk in terms of the posterior probability $P(x|qB)$.

Another problem this group has studied is the effects of grouping data on first and second distribution moments. Data such as income are often grouped and released as interval data, considered to be one of the best ways of summarizing data that has disclosure risk implications as well. Class marks (midpoints) of intervals are then used to calculate the mean and variance of the grouped data. In most situations, using midpoints for every observation in the interval smoothes the data, thereby reducing the variance. It can be shown, as in analysis of variance, that, using midpoints, we lose the within-interval variance component if within-interval data have a uniform distribution. However, if distributions within some intervals are peaked or

skewed, use of the midpoints of the interval data can result in higher variance estimates than would be obtained with the raw data. Moreover, for those data, the mean of the grouped data based on the use of midpoints is biased. If class (conditional) means are used for calculating overall mean and variance, the mean of the raw data can be recaptured and the variance will be lower. They have obtained some initial results from their study of the impact of accepted practices for approximating moments with summarized data.

As a result of their work, Gonzalez, Cox, Kim, and Katzoff will present two papers at the American Statistical Association Joint Statistical Meetings (JSM 2004), *Statistics as a Unified Discipline,* August 8-12, 2004 in Toronto, Canada. The papers are "Effects of Rounding Continuous Data Using Specific Rules" and "Effects of Grouping Continuous Data on First and Second Distribution Moments." These papers will be published in the ASA Survey Research Methods Section proceedings.

FINDINGS – Vaccination Strategies

Chris Bauch, McMaster University, attended the June 2002 DIMACS International Conference on Computation and Mathematical Epidemiology, and saw a talk on mathematical modeling of smallpox, and a separate talk on using game theory to model the interaction between the US government and terrorists. This stimulated a paper on applying game theory to assess voluntary vaccination policy for smallpox, which appeared in *Proceedings of the National Academy of Sciences,* September 2, 2003. Bauch, Galvani, and Earn showed that if individuals decide whether to vaccinate according to self-interest, the level of herd immunity achieved might differ from what is best for the population as a whole. They used a synthesis of game theory and epidemic modeling to formalize the conflict between self-interest and group interest and showed that voluntary vaccination in unlikely to reach the group-optimal level. This shortfall could result in a substantial increase in expected mortality after a bioterrorist attack.

FINDINGS - Data Mining Tool in Disease Diagnosis

New scientific problems, arising from the human genome project, are challenging the classical means of using statistics. Yet quantified knowledge in the form of rules and rule strengths based on real relationships in data, as opposed to expert opinion, is urgently required for researcher and physician decision support. The problem is that with many parameters, the space to be analyzed is highly dimensional. That is, the combinations of data to examine are subject to a ombinatorial explosion. These combinatorial considerations are particularly problematic for identifying those associations called “Unicorn Events” that occur very infrequently. Barry Robson, IBM, has developed a novel approach to the solution of the problem using the Zeta function. Another problem investigated by Robson is that of combining qualitative and quantitative data from a variety of clinical observations to form conclusions that can be assembled in a form better suited for decision support. This technique has applications to the study of real, extensive patient record data as well as molecular design data. He has developed methods of rationalizing and condensing the rules implied by statistical associations and variances between data. Advanced computation was used to design the Mad Cow Disease diagnostic of which Robson was co-inventor, widely used in Europe and now accepted by the US Department of Agriculture April, 2004.

Products:

PAPERS

Abello, J., Pogel, A., Miller, L., "Graph partitions and formal concept lattices," to appear in *Journal of Universal Computer Science*.

Abello, J., Pogel, A., "Bigraph decompositions and concept lattices," tentative title, in preparation.

Balthrop, J., Forrest, S., Newman, M. E. J., and Williamson, M. M., "Technological networks and the spread of computer viruses," *Science* 304, 527-529 (2004).

Bauch, C. T., Galvani, A. P., and Earn, D. J. D., "Group interest versus self-interest in smallpox vaccination policy," *Proceedings of the National Academy of Sciences*, Vol 100, No. 18, 10564-10567 (2003).

Boechler, M., "A framework for capability-based evaluation of chem/bio defense resources," *Proceedings Military Operations Research Society 71st Symposium* (2003).

Buckeridge, D. and Frisén, M., "The workflow problem," in preparation.

Chen, S., Gaur, A. Muthukrishnan, S., and Rosenbluth, D., "Wireless *in loco* sensor data collection and applications," in preparation.

Cormode, G. and Muthukrishnan, S. "Estimating dominance norms of multiple data streams," *Proceedings of European Symposium on Algorithms*, 148-160 (2003).

Cox, L., Gonzalez, Jr., J. F., and Katzoff, M., "Effects of rounding continuous data using specific rules," to appear in *Proceedings of the ASA Joint Statistical Meetings Survey Research Methods Section*.

Cox, L., Gonzalez, Jr., J. F., and Katzoff, M., "Effects of grouping continuous data on first and second distribution moments," to appear in *Proceedings of the ASA Joint Statistical Meetings Survey Research Methods Section*.

Cross, P. C., Lloyd-Smith, J. O., Bowers, J., Hay, C. T., Hofmeyr, M., and Getz, W. M., "Integrating association data and disease dynamics: an illustration using African buffalo in Kruger National Park," submitted to *Annales Zoologici Fennici*.

De Leenheer, P., Levin, S.A., Sontag, E.D, and Klausmeier, C.A., "Global stability in a chemostat with multiple nutrients," submitted to *Journal of Mathematical Biology*.

De Leenheer, P., Levin, S.A., Sontag, E.D, and Klausmeier, C.A., "Global stability in a chemostat with multiple nutrients," DIMACS Technical Report 2003-40, 2003.

De Leenheer, P., Angeli, D., and Sontag, E.D., "Crowding effects promote coexistence in the chemostat," submitted to the *Journal of Mathematical Analysis and Applications*.

De Leenheer, P., Angeli, D., and Sontag, E.D., "Crowding effects promote coexistence in the chemostat," DIMACS Technical Report 2003-44, 2003.

De Leenheer, P., Angeli, D., and Sontag, E.D., "A tutorial on monotone systems - with an application to chemical reaction networks," to appear in *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), July 5-9, Leuven, Belgium.*

De Leenheer, P., Angeli, D., and Sontag, E.D., "A tutorial on monotone systems - with an application to chemical reaction networks," DIMACS Technical Report 2004-15, 2004.

De Leenheer, P., Angeli, D., and Sontag, E.D., "Monotone chemical reaction networks," to be submitted to *Journal of Mathematical Chemistry*.

De Leenheer, P., Angeli, D., and Sontag, E.D., "Monotone chemical reaction networks," DIMACS Technical Report 2004-16, 2004.

De Leenheer, P., and Sontag, E.D., "A note on the monotonicity of matrix Riccati equations," to be submitted to the *IEEE Transactions on Automatic Control*.

Dushoff, J., Levin, S. A., and Plotkin, J. B., "Evolution and persistence of influenza {A} and other diseases," Math. Biosci. 188: 17-28 (2004).

Dushoff, J., Harris, A. D., Levin, S. A., Perencevich, E. N., and Smith, D. L., "Persistent colonization and the spread of antibiotic resistance in nosocomial pathogens: Resistance is a regional problem," PNAS 101: 3709-3714 (2004).

Dushoff, J., Fraser, H. B., and Plotkin, J. B., "Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*," *Nature* 428: 942-945 (2004).

Franc, Alain, "Metapopulation dynamics as a contact process on a graph," *Ecological Complexity*, Volume 1, Issue 1, 49-63 (2004).

Frisén, M. and Sonesson, C., "Optimal surveillance," to appear in *Spatial Surveillance for Public Health*. Eds, Lawson, A. and Kleinman, K.

Frisén M., "Statistical measures for evaluation of methods for syndromic surveillance," to appear in *Conference Proceedings 2003 National Syndromic Surveillance, Morbidity and Mortality Weekly Report Supplement*.

Hadeler, K.P., "The role of migration and contact distributions in epidemic spread," in C. Castillo-Chavez, H.T. Banks (eds), *Biomathematical Modeling Applications in Homeland Security, Frontiers in Applied Mathematics*, SIAM, 203-214 (2003).

Hadeler, K.P., and Hillen, T., "Hyperbolic systems and transport equations in mathematical biology," chapter in *Analysis and Numerics for Conservation Laws (AnumE) Report 2004*, Springer, to appear.

Hadeler, K.P., and Hillen, T., "Reaction transport equations with coupled dynamics and with sedentary states," in preparation.

Hartke, S., "Graph-theoretic Models of Spread and Competition," Ph.D. Thesis, Department of Mathematics, Rutgers University, July 2004.

Hoover, D.R., Donnay A, Mitchell, et.al., "Reproducibility of immunological tests used to assess multiple chemical sensitivity syndrome," *Clin & Diag Lab Immunol* 10(6):1029-1036 (2003).

Hoover, D. R., "Subject allocation and curtailment for fixed event comparative Poisson trials," *Statistics in Medicine*, to appear.

Joo, J., and Lebowitz, J. L., "Behavior of SIS epidemics on heterogeneous networks with saturation," DIMACS Technical Report 2004-14, 2004.

Joo, J., and Lebowitz, J. L., "Behavior of susceptible-infected-susceptible epidemics on heterogeneous networks with saturation," to appear in *Phys. Rev. E*.

Joo, J., and Lebowitz, J. L. "Pair approximation of the stochastic susceptible-infected-susceptible epidemic model on the hypercubic lattice," submitted to *Phys. Rev. E*.

Lin J., Andreasen, V., Casagrandi, R., and Levin, S. A., "Traveling wave solutions in a model of influenza A drift," *Journal of Theoretical Biology* 222: 437-445 (2003).

Pilyugin, S. S., Reeves, G. T., and Narang, A., "Stability of mixed microbial cultures: Connecting theory and experiments, Part 1. Unstructured model," submitted to *Mathematical Biosciences.*

Pilyugin, S. S., Reeves, G. T., and Narang, A., "Stability of mixed microbial cultures: Connecting theory and experiments, Part 1. Structured model," submitted to *Mathematical Biosciences.*

Roberts, F.S., "Challenges for discrete mathematics and theoretical computer science in the defense against bioterrorism," in C. Castillo-Chavez and H.T. Banks (eds.), *Mathematical and Modeling Approaches in Homeland Security*, SIAM Frontiers in Applied Mathematics Series (2003), 1-34.

Robson, B., "Clinical and pharmacogenomic data mining: 1. Generalized theory of expected information and application to the development tools," *Journal of Proteome Research* (2), 283-302 (2003).

Robson, B., "Clinical and pharmacogenomic data mining: 2. A simple method for the combination of information from associations and multivariances to facilitate analysis, decision, and design in clinical research and practice," *Journal of Proteome Research,* to appear.

Roverato, A., "A Unified approach to the characterization of equivalence classes of DAGs, chain graphs with no flags and chain graphs," *Scandinavian Journal of Statistic,* to appear.

Solari, H. G., and Natiello, M. A., "Stochastic population dynamics: The Poisson approximation," *Physical Review E*, 67, 031918 (2003).

Stoto, M. A., Schonlau, M., and Mariano, L. T., "Syndromic surveillance: Is it worth the effort?" *Chance*, Vol. 17, No. 1, 19–24 (2004).

BOOKS

Dieckmann, U., Feng, Z., and Levin, S. (eds.), *Disease Evolution: Models, Concepts, and Data Analyses*, American Mathematical Society, in preparation.

TALKS

Mike Boechler, "A framework for capability-based evaluation of chem/bio defense resources," 71st Symposium of the Military Operations Research Society, Quantico, Va., June 10-12, 2003.

Su Chen, Amit Gaur, S. Muthukrishnan, and David Rosenbluth, "Wireless *in loco* sensor data collection and applications," presented at the Emerging Applications of Wireless Infrastructure at WWW04 conference, New York, May 18, 2004.

Guido Consonni, University of Pavia (Italy), gave a talk on "Hierarchical mixture priors for the analysis of gene expression data in cancer studies" at the DIMACS Epidemiology Seminar, December 8, 2003

Marianne Frisén, "Optimal surveillance of health events." Invited talk at the International Biometric Society Eastern North American Region conference Pittsburgh, PA, March 30, 2004.

Joe Fred Gonzalez, Lawrence Cox, Jay J. Kim, and Myron Katzoff, "Effects of rounding continuous data using specific rules," American Statistical Association Joint Statistical Meetings (JSM 2004), *Statistics as a Unified Discipline*, Toronto, Canada, August 8-12, 2004.

Joe Fred Gonzalez, Lawrence Cox, Jay J. Kim, and Myron Katzoff, "Effects of grouping continuous data on first and second distribution moments," American Statistical Association Joint Statistical Meetings (JSM 2004), *Statistics as a Unified Discipline*, Toronto, Canada, August 8-12, 2004.

Stephen Hartke, “The Voter Model with Confidence Levels,” presented at the Workshop on Extremal Combinatorics, Alfred Renyi Institute of Mathematics, Budapest, Hungary, on April 16, 2003.

Stephen Hartke, “The Voter Model with Confidence Levels,” presented in an AMS Session on Dynamical Systems at the Joint Mathematics Meetings of the AMS and the MAA in Baltimore, Maryland, on January 16, 2003.

Jaewook Joo, “Spreading of disease in population with hetereogeneous mixing,” Department of Biomedical Engineering, Boston University, Boston, MA, April 2004.

Jaewook Joo, “Spreading of disease in population with hetereogeneous mixing,” Department of Physics, Pennsylvania State University, State College, PA, March 2004.

Jaewook Joo, “Behavior of SIS epidemics on heterogeneous networks with saturation,” 90th Statistical Mechanics Conference, Rutgers University, New Brunswick, NJ, December 2003.

Sergei S. Pilyugin, “Dynamics of mixed microbial growth: Stability of a structured model” DIMACS Seminar Series, April 27, 2004.

Sergei S. Pilyugin, “Dynamics of mixed microbial growth: Stability of a structured model” Mathematical Biosciences Institute, OSU, May 3, 2004.

Fred Roberts, “Mathematics and bioterrorism: Graph-theoretical models of spread and control of disease.” DIMACS Connect Institute July 26, 2002.

Fred Roberts, “Mathematics and bioterrorism: Graph-theoretical models of spread and control of disease.” West Point Discrete Math Meeting October 2002.

Fred Roberts, “Challenges for discrete mathematics and theoretical computer science in the defense against bioterrorism.” SIAM Meeting, July 11, 2002.

Fred Roberts, “Homeland security: What can mathematics do?” Precalculus Conference, Center for Mathematics, Science, and Computer Education, Rutgers University, March 19, 2004.

Michael A. Stoto, “Syndromic surveillance: Is it worth the effort?” Conference on Statistical Issues in Counterterrorism, May 29-30, 2003, The Keck Center of the National Academies, New York.

Michael A. Stoto, “Syndromic surveillance: Is it worth the effort?” International Biometric Society Eastern North American Region Spring Meeting, Pittsburgh, PA, March 29, 2004.

Michael A. Stoto, “Syndromic surveillance: Is it worth the effort?” University of Pittsburgh Center for Research on Health Care, Pittsburgh, PA, March 30, 2004.

Michael A. Stoto, "Syndromic surveillance: Is it worth the effort?" University of New Mexico, 2004 WNAR/IMS Meeting, June 27-30, 2004.

Peter Sandman, "Risk Communication Applications to Intelligence ... and to Controversies over Intelligence," Invited talk at the Defense Intelligence Agency, Bolling Air Force Base, Washington, DC, August 19, 2003. (It was also teleconferenced to other Defense Intelligence Agency sites.)

Dona Schneider, "Descriptive epidemiology for data miners," March 19, 2004 at the DIMACS Working Group on Data Mining and Epidemiology

Websites

http://dimacs.rutgers.edu/Workshops/AdverseEvent/;
http://dimacs.rutgers.edu/Workshops/Analogies/;
http://dimacs.rutgers.edu/Workshops/Modeling/;
http://dimacs.rutgers.edu/Workshops/WGDMEpid/;
http://dimacs.rutgers.edu/SpecialYears/2002_Epid/

Reports

Report on DIMACS Working Group on Privacy / Confidentiality of Health Data
Report Authors: Hiran Subramaniam and Zhiqiang Yang, Department of Computer Science, Stevens Institute of Technology
Date of Report: December 20, 2003
http://dimacs.rutgers.edu/Workshops/Health/priv-health-data-12-03.pdf

DIMACS Working Group Report on Data Mining and Epidemiology, March 18-19, 2004
Report Authors: James Abello and Graham Cormode, DIMACS
In preparation.
http://dimacs.rutgers.edu/Workshops/WGDataMining/material/ (password protected)

Contributions:

CONTRIBUTIONS TO THE DISCIPLINE

The "discipline" is by definition a combination of disciplines. Many of the results described in the Activities and Findings section of this report illustrate this combination of disciplines, in particular the application of methods of computer science and mathematics to problems of epidemiology. Some examples are Frisen's and Stoto's work on syndromic surveillance to detect bioterrist events, Kao and Haydon's work on controlling the spread of foot-and-mouth disease, Consonni and Madigan's work on comparing the performances of health care providers, Muchnik's work on surviving prostate cancer, Muchnik and Morgan's efforts to understand the causes of diseases on chicken farms, Chen, Gaur, Muthukrishnan, and Rosenbluth's wireless sensors to gather health data from people "on the go," Levin, Dushoff, Smith, Perencevish, and Harris' work on the transmission of antibiotic resistant bacteria, and Jakob and Mikler's model

of the spread of the West Nile virus. Several participants and groups of participants, including Hadeler, Franc, Hartke, and Joo and Lebowitz, have created other models of epidemics. These are all described in more detail earlier in this report.

Introducing people to this combination of disciplines has been a key goal and a key accomplishment of this project. In addition, the following is a selection of comments we received from participants in our project activities.

1. "The workshop Computational and Mathematical Epidemiology had a sensible impact on my research projects, …. A short list:

> 1. The workshop was very relevant in deciding whether I should move into Population Dynamics or not. The meeting was important in gauging the receptivity of the community to my interests and ideas. The environment of the meeting persuaded me that it was worth the effort to join this community.
>
> 2. Discussion held at the meeting with J. Aron and D. Burke as well as other participants were decisive to open a new research line: eco-epidemiology of Aedes Aegypti and Dengue. We are actively pursuing the research and early results are very encouraging
>
> 3. I felt encouraged to complete our work on a Poisson approximation to stochastic population dynamics. The main results are published in Physical Review….

The work has triggered some unusual events. For example, the method presented has been promptly incorporated to the studies of Molecular Networks relevant to genetic technology (see the conference by J. Peccoud -Pioneer, Dupont- at http://www.crhc.uiuc.edu/Multi/) and the (unusual for a physicist or applied mathematician) invitation to join the Medical Science Monitor as a reviewer."
Hernán G Solari, University of Buenos Aires

2. Atul Narang, Department of Chemical Engineering, University of Florida has developed a working relationship with Eduardo Sontag, Department of Mathematics, Rutgers University. Narang has been working on the dynamics of bioreactors. When he approaches control-theoretic questions for bioreactors, he expects to be collaborating closely with Sontag. Two of Sontag's students, German Enciso and Magdalena Chaves, have already benefited from their interaction with Narang.

3. "Last May I attended a DIMACS working group on Data Mining and Epidemiology. At the meeting I heard some thought provoking presentations, especially one from Dr. David Madigan about methodological issues in comparing hospital performance. Dr. Madigan presented some ideas that I plan to revisit in the future. I also sat in on a subgroup that was planning to sponsor a data analysis competition using one of the public health data sets (I believe it was the SEER cancer data). I thought this was a great idea. Since I work in biostatistics and epidemiology I'm very interested in the activities of this working group and look forward to future meetings."
Brian James, Pfizer

4. “I attended a seminar on adverse events and surveillance organized by DIMACS Adverse Event/Disease Surveillance, Reporting and Analysis Working Group. The seminar was very helpful. It gave me a better understanding of what others know about surveillance. We apply surveillance (risk management) methods on a daily basis.” Juhaeri Juhaeri, Senior Manager, Global Epidemiology, Aventis Pharmaceuticals Inc.

5. “My participation in the DIMACS Working Group on Adverse Event/Disease Reporting, Surveillance, and Analysis II (Second Meeting: February 19 - 20, 2004) has in fact led to new collaborations. I´m now preparing a paper with Dr. David Buckeridge (Stanford University) about the workflow problem, to be submitted in the next few months. The meeting was a great opportunity for research contacts, and I expect to participate in future workshops.” Luiz Duczmal, Statistics Department, Universidade Federal de Minas Gerais, Brazil

6. “We are able to borrow ideas from infectious disease tracking statistics to model temporal trends in air pollution and asthma. Discussions with your speakers during the break gave us chances to see software demos (Terra Seer) and collect ideas for "Clustering detection" in chronic diseases. We have had a chance to follow up in obtaining and using SatScan for cluster detection on statewide asthma data.” Stella Manchun Tsai, Research Scientist, New Jersey Department of Health and Senior Services

7. “I did have a generally positive experience with the DIMACS conference last year (on signal detection), and it did lead to followup discussions with CDC and with FDA. A little hard to identify specific research that's clearly traceable to the meeting, but it did raise our awareness of some of the syndromic surveillance issues that has been included in some of our proposals (e.g., our SBIR Phase II proposal to DOD related to signal detection in longitudinal data (military patient records)). Chan Russell, Lincoln Technologies

8. “As a participating epidemiologist, I have

1. Collaborated with a team of mathematicians who wished to explore large health-related data sets. We examined the SEER data and found (after some concerted effort) that one of the three approaches we selected did not work, a second approach turned into a doctoral thesis topic, and a third is still evolving.
2. I now serve as an outside thesis committee member for a computer science doctoral student. I find that very exciting. Ilya Muchnik is the advisor for the student thesis.
3. I gave a talk on "Descriptive Epidemiology for Data Miners" in March 19, 2004 at DIMACS to about 20 data miners. That effort sparked discussion about differences in our vocabularies and approaches to problems. As a result of that presentation, I have been asked by James Abello to write a book chapter on descriptive epidemiology for data miners. What a wonderful prospect!” Dona Schneider, Rutgers University

9. “The one meeting I attended at DIMACS was very informative, interesting, and well run. The meeting helped foster an incipient collaboration with Matt Williamson who also attended the meeting.” Stephanie Forrest, University of New Mexico

10. Dunrie A. Greiling, TerraSeer, attended the DIMACS tutorial June, 2003. It resulted in a collaboration between her organization and the American Association of Poison Control Centers,

through the contact she made with Bill Watson at the meeting. She submitted a grant to NIH to fund a project with Watson and the AAPCC as consultants/a test organization to deploy the software they proposed to build. It wasn't given a fundable score, but they will rewrite it and resubmit.

11. "My participation in the DIMACS conference on modeling the social side of outbreaks/attacks led to an invitation to talk to the DIA (which I accepted) and to participate in a CIA-sponsored examination of SARS (which I declined because of a schedule conflict)." Peter Sandman, Psandman.com

12. "We are beginning a project with the Department of Health in the District of Columbia looking at their syndromic surveillance program. The January DIMACS meeting was very useful to me in terms of getting ideas for the multivariate analysis of these data. Although it is early we have submitted an abstract to the American Public Health Association's meeting in DC in the fall…." Mike Stoto, RAND Center for Domestic and International Health Security

13. "I have found the special focus meetings that I have attended very valuable in terms of stimulating research ideas and the establishment of scientific connections with other important researchers in this area. I have also incorporated knowledge gained at one of the meetings (DIMACS Working Group on Modeling Social Responses to Bio-terrorism Involving Infectious Agents) into a subsequent paper presented at last year's annual meeting of the Military Operations Research Society:
http://www.mors.org/publications/abstracts/71morss/wg2_abs.htm

In addition, at that DIMACS meeting and a related predecessor (DIMACS Working Group Meeting: Mathematical Sciences Methods for the Study of Deliberate Releases of Biological Agents and their Consequences) I was fortunate to make the acquaintance of Dr. Ellis McKenzie who subsequently contracted our group to model a set of bioterrorism scenarios. That work is ongoing and we expect that it will result in at least one publication for submission within the next 6-9 months.

From the limited exposure that I have had to the DIMACS group, I must say that I am, nonetheless, very impressed by the caliber of the presentations and discussions. I think that NSF would be very hard pressed to get more 'bang for the buck', so to speak, than they get with this group, and I very much look forward to attending future DIMACS sessions." Michael Boechler, Product Group Manager, Bioterrorism Preparedness and Response Innovative Emergency Management, Inc.

CONTRIBUTIONS TO OTHER DISCIPLINES OF SCIENCE AND ENGINEERING

Since the "discipline" is inherently multidisciplinary, there is no separate entry in this section.

CONTRIBUTIONS TO EDUCATION AND HUMAN RESOURCES DEVELOPMENT

Many graduate students, undergraduates, and several postdocs participate in the program. Three Rutgers students, Rohan Fernandez (Computer Science), Stephen Hartke (Mathematics), and Jaewook Joo (Mathematics and Physics) have been heavily involved. Other local graduate students and many non-local students were also involved as visitors and workshop/working

group attendees. The project has two primary postdocs, Graham Cormode and Patrick De Leenheer, as well as participation by postdoc visitor Michael Capalbo and many visiting postdocs. More senior people were also heavily influenced by the project, being exposed to new directions of research and changing their fields as a result. The impact on the careers of the students and postdocs is illustrated by a few examples.

1. "As a participant in the Computational and Mathematical Epidemiology program I had the opportunity to meet and to start a collaborative work with several specialists in the biomedical field from the Cancer Institute of New Jersey, the Institute for Advanced Study, the Broad Institute, NIH and IBM. Also I attended several tutorials focused on computational biology and I appreciate them as very useful for my activity. The major topic of research I am involved in is cancer genomics. So far I obtained in collaboration with specialists from the Institute for Advanced Study and from the Cancer Institute of New Jersey, some results for a time-series microarray data analysis of the stress responsive gene pathways (manuscript submitted). Currently I am working with specialists from the Institute for Advanced Study and from IBM in the detection of combinatorial biomarkers for prostate cancer and I intend to present the results of this study in a forthcoming national conference on mass spectrometry." Gabriela Alexe, graduate student, RUTCOR, Rutgers University

2. "I am a doctoral student at UC Berkeley. I participated in the opening meeting of the special focus on epidemiology in summer 2002, and in the working group on spatiotemporal and network modeling of disease in April 2003. I found both meetings very stimulating, and made many excellent professional contacts… information and perspectives I gained have certainly influenced my subsequent research. In particular, my group recently submitted a manuscript incorporating network models of disease dynamics, and my experience in the working group had a strong influence on this work. Another project that I'm currently working on is also benefiting from the perspectives gained in the working group. I'm grateful to DIMACS for facilitating these very useful meetings." James Lloyd-Smith, Biophysics, UC Berkeley

3. "The workshop stimulated me to think about Statistical Quality Control (SQC) tools for monitoring incidence data. Not much has been written about this prior to 9/11, except an article that I found in Technometrics by S.B. Vardeman. The SQC tools do require adjustment for seasonality and trends, and conventional time series methods often will not apply, so the workshop has stimulated me to think about such methods, particularly in connection with the situation in my home state (Colorado) with respect to West Nile Virus. I hope to prepare a paper based on my DIMACS presentation and will acknowledge DIMACS accordingly. The workshop also stimulated me to work with a colleague on the theory of the spatial scan statistic for identifying disease clusters. We are preparing a short note on this and will keep you informed. I also based one homework problem and one final exam problem on the problem of time trends and spatial cluster detection of disease incidence in Mathematical Statistics I." Karen Kafadar, University of Colorado

4. "Ongoing work with Stephanie Forrest at UNM as a result of the meeting [I attended] has been quite fruitful. One of her students, Justin Balthrop [whom I met at the meeting] came to work with me for a 6 month internship, and he will be writing a paper for a conference this year. A collaboration with Prof. Forrest, Justin Balthrop and Mark Newman has resulted in a paper in

Science: 'Technological networks and the spread of computer viruses'…" Matt Williamson, Hewlett-Packard Labs, Bristol, England
See the Papers section of this report.

5. As reported in the Section on Findings, Su Chen, Amit Gaur, Muthu Muthukrishnan, David Rosenbluth used on-person wireless sensors for monitoring physiological, behavioral and health-related data. In particular, they focused on *in loco* monitoring, that is, persistent, unobtrusive monitoring of human factors in *natural situations*. "Based on my work last summer [on wireless sensor networks], I taught a course in Spring 04 at Rutgers on expanding this functionality." Muthu Muthukrishnan, Rutgers University

In addition, the following graduate students have undertaken small research projects under support of the project. Their work is described under their names in the list of project participants.

Liming Wang, Mathematics, winter 03/04
Yihua Wu, Computer Science, winter 03/04
Smriti Bhaga, Computer Science, summer 2004
German Enciso, Mathematics, summer 2004
Samir Goel, Computer Science, summer 2004
Jaewon Kang, Computer Science, summer 2004
Liming Wang, Mathematics, summer 2004

Visitors:

David Angeli, University of Firenze, 3/20/2004-3/28/2004
Guido Consonni, Universitá di Pavia, 11/21/2003-12/20/2003
Atul Narang, University of Florida, 4/20/2004-4/21/2004
Sergei Sergeevitch Pilyugin, University of Florida, 4/26/2004-4/28/2004
Alberto Roverato, Universitá di Economia Politica, 11/21/2003-12/20/2003
Frank van Ham, Technische Universiteit Eindhoven, 3/6/2004-3/28/2004