DIMACS Technical Report 2002-40

Challenges for Discrete Mathematics and Theoretical Computer Science in the Defense Against Bioterrorism

Fred S. Roberts
DIMACS and Department of Mathematics
Rutgers University
96 Frelinghuysen Road
Piscataway, NJ 08854

Draft of September 19, 2002

DIMACS is a partnership of Rutgers University, Princeton University, AT\&T Labs-Research, Bell Labs, NEC Research Institute and Telcordia Technologies (formerly Bellcore).

DIMACS was founded as an NSF Science and Technology Center, and also receives support from the New Jersey Commission on Science and Technology.

Abstract

Dealing with bioterrorism requires detailed planning of preventive measures and responses. Both planning and response require precise reasoning and extensive analysis of the type that mathematical scientists are very good at. Discrete mathematics and theoretical computer science, broadly defined, seem very relevant. We describe challenges for discrete mathematics and theoretical computer science raised by the need to plan for and defend against bioterrorist attacks.

Challenges for Discrete Mathematics and Theoretical Computer Science in the Defense Against Bioterrorism

Fred S. Roberts
DIMACS and Department of Mathematics
Rutgers University
96 Frelinghuysen Road
Piscataway, NJ 08854

Draft of September 19, 2002

Abstract

Dealing with bioterrorism requires detailed planning of preventive measures and responses. Both planning and response require precise reasoning and extensive analysis of the type that mathematical scientists are very good at. Discrete mathematics and theoretical computer science, broadly defined, seem very relevant. We describe challenges for discrete mathematics and theoretical computer science raised by the need to plan for and defend against bioterrorist attacks.

Acknowledgements: Research supported by NSF grant EIA-0205116 to Rutgers University. Parts of this paper were borrowed from grant proposals and many people contributed pieces to the proposals. In particular, I would like to thank Steve Greenfield, Stuart Haber, Eddie Holmes, Donald Hoover, Frank Hwang, Mel Janowitz, Brenda Latka, Simon Levin, David Madigan, Ilya Muchnik, S. Muthukrishnan, Rafail Ostrovsky, David Ozonoff, Benny Pinkas, Tomas Sander, Mike Steele, Dan Wartenberg, Tom Webster, and Rebecca Wright, for ideas used in this paper. My apologies to those whose names I may have inadvertently omitted. Some parts of this paper owe a great deal to the DIMACS Working Group Meeting on Mathematical Sciences Methods for the Study of Deliberate Releases of Biological Agents and their Consequences (http://dimacs.rutgers.edu/Workshops/WGDeliberate/), and in particular the report that came out of that meeting: Castillo-Chavez and Roberts [2002]. I would like to thank all of the participants in that meeting for their insights, in particular the discussion group on challenges for computer science, and especially my co-author, Carlos Castillo-Chavez.

1. Introduction

The recent acts of bioterrorism using anthrax have raised great concern about the use of biological agents as weapons of mass destruction. This has led to major national debate and discussion, for example about the possible impact of viruses such as smallpox as a biological weapon and about the possibility that bioterrorists would take advantage of genetic manipulation of highly variable viruses such as influenza. The great concern about the deliberate introduction of diseases by bioterrorists has raised new challenges for mathematical scientists. Dealing with bioterrorism requires detailed planning of preventive measures and responses: vaccination, vaccine dilution, vaccine stockpiling, quarantine of individuals, buildings, or regions, and systematic surveillance of emergency rooms, food supply, and water supply, to name just a few examples. Both planning and response require precise reasoning and extensive analysis of the type that mathematical scientists are very good at. Discrete mathematics and theoretical computer science, broadly defined, seem very relevant. In this paper, we describe challenges for discrete mathematics and theoretical computer science raised by the need to plan for and defend against bioterrorist attacks. To begin, however, we talk briefly about a variety of mathematical sciences approaches to the defense against bioterrorism.

Understanding infectious systems consisting of parasites, vectors, and hosts requires being able to reason about highly complex biological systems, sometimes with hundreds of demographic and epidemiological variables. Obtaining such understanding is crucial in the defense against bioterrorism. Components of host-pathogen systems are sufficiently numerous and their interactions sufficiently complex that intuition alone is insufficient to fully understand the dynamics of such interactions. Yet, experimentation or field trials are often prohibitively expensive or unethical and do not always lead to understanding at a fundamental level. Therefore, mathematical modeling becomes an important experimental and analytical tool. Mathematical models have become important tools in analyzing the spread and control of infectious diseases, especially when combined with the use of powerful, modern computer methods for analyzing and/or simulating the models. (See for instance Nowak and May, 2000.) Mathematical approaches can also be used to compare alternative intervention programs, to prepare responses, as guides for training exercises and scenario development, to help with risk assessment, as an aid in forensic analysis, and to predict future trends. All of these issues are central to the extensive planning required to deal with bioterrorist attacks.

2. Methods of Mathematical and Computational Epidemiology

Many of the research challenges facing us in the defense against bioterrorism apply more generally to the defense against disease. Thus, we shall speak rather generally here about methods of "mathematical and computational epidemiology." Mathematical models of infectious diseases go back to Daniel Bernoulli's mathematical analysis of smallpox in 1760 and have been developed extensively since the early 1900s. Hundreds of mathematical models have been published since, exploring the effects of bacterial, parasitic, and viral pathogens on human populations. The results have highlighted and formalized such concepts as the core population in sexually transmitted diseases (Hethcote and Yorke, 1984) and made explicit other concepts such as herd immunity for vaccination policies (Anderson,

1982). Key pathogens that have been studied are: Malaria (Anderson and May, 1991), *Neisseria gonorrheae* (Hethcote and Yorke, 1984); *M. tuberculosis* (Blower, et al., 1995, Blower, Small, and Hopewell, 1996, Castillo-Chavez and Feng, 1997, 1998, Murray, 1980); *HIV* (Perelson, et al., 1996, Anderson, Ascher, and Shepard, 1998); and *T. palladum* (Bailey, 1975). Mathematical modeling has provided new insights on important issues such as drug-resistance, rate of spread of infection, epidemic trends, and the effects of treatment and vaccination. (For instance, for modeling work on vaccination strategies, see Garnett and Waddell, 2000, Grenfell and Bolker, 1998, Hethcote, 1999, Rohani, Earn, and Grenfell, 2000, Schuette and Hethcote, 1999.) Yet, for many infectious diseases, we are far from understanding the mechanisms of disease dynamics. The modeling process can lend insight and clarification to data and theories. Mathematical models, with the aid of computer simulations, are useful theoretical and experimental tools for building and testing theories about complex biological systems involving disease, assessing quantitative conjectures, determining sensitivities to changes in parameter values, estimating key parameters from data, and preparing intervention strategies against newly emerging diseases or deliberate biological attacks. The size of modern epidemiological problems and the large data sets that arise call out for the use of powerful computational methods for studying these large models. As Levin, et al. (1997) point out, "imaginative and efficient computational approaches are essential in dealing with the overwhelming complexity of [such] biological systems." New mathematical and computational methods are needed to deal with the dynamics of multiple interacting strains of viruses through the construction and simulation of dynamic models, the problems of spatial spread of disease through pattern analysis and simulation, and the early detection of emerging diseases or bioterrorist acts through rapidly-responding surveillance systems. An important challenge for the mathematical sciences is to investigate and seek to stimulate the development of such new methods and to use the new methods in a variety of new applications.

Various methods from the mathematical sciences have been used in epidemiology, It is important to try to involve a larger number of mathematical scientists in partnership with epidemiologists by exposing the mathematical sciences community to the models and methods that have already been applied to epidemiology, and exploring new variants on these models and methods for dealing with new and complex issues as yet unexplored, under the guidance of leading mathematical and biological scientists already involved in the enterprise. Statistical methods have long been used in mainstream epidemiology largely for the purpose of evaluating the role of chance and confounding induced associations (Szklo and Nieto, 2000). Epidemiologists seek to ferret out sources of systematic error ("bias and confounding") in the observations, evaluate the role of uncontrollable error (using statistical methods) in producing the results, and interpret the results using correlative information from the medical and biological sciences.

A smaller but venerable tradition within epidemiology has considered the spread of infectious disease as a dynamical system and applied difference equations and differential equations to that end (Brauer and Castillo-Chavez, 2001, Diekmann and Heesterbeek, 2000, Hethcote, 2000). Dynamical systems, both continuous and discrete, are relevant to many epidemiological models (classical examples are in host-parasitoid models such as Nicholson-Bailey or SIR or SIRS). Examples of relevant work on host-parasitoid and host-pathogen models are the models of urinary tract infections by Gordon and Riley (1992), mycoparasite-immune dynamics by Antia, Levin, and May (1994) and Antia, Koella, and Perrott (1996),

antimicrobial chemotherapy by Lipsitch and Levin (1997,1998), and the recent work on tuberculosis by Wigginton and Kirschner (2001). Critical phenomena models (see, for example, Stauffer and Aharony, 1992) can be used to model the sudden phase transition that occurs when a disease becomes epidemic. Threshold phenomena are well-known in dynamical systems models of infectious disease (Brauer and Castillo-Chavez, 2001). These newer phase-transition models can be important tools to understand how sporadic disease occurrence can erupt into sudden epidemics (Eastern Equine Encephalitis and West Nile Virus are two important examples where understanding the phenomenon is critically needed). Yet, some dynamic models of spread of viruses in non-biological systems (Faloutsos, Faloutsos, and Faloutsos, 1999, Albert, Jeong, Barabasi, 1999) do not exhibit threshold effects, and it will be important to understand why and what insight this can give us about spread of disease. In spite of this large body of work, little systematic effort has been made to apply today's powerful computational methods to these dynamical systems models and relatively few mathematical scientists have been involved in the process. It is an important challenge to the mathematical sciences to change this situation.

Probabilistic methods, in particular stochastic processes, have also played an important role in infectious disease modeling. (See, e.g., Mollison, 1995, Durrett, 1999.) Random walk models have been used to study a variety of diseases (see, e.g., Hadeler, 1997). Percolation theory has been used to model spread of contagious phenomena like forest fires and epidemics (Stauffer and Aharony, 1992) and related research on interacting particle systems and probabilistic cellular automata (see, e.g., Mollison, 1995, Rand, Keeling, and Wilson, 1995, Levin and Durrett, 1997) has also become useful in epidemic modeling. Markov chain Monte Carlo methods are promising tools used to fit data to models of numerous diseases (see, e.g., Gilks, Richardson, Spiegelhalter, 1996, O'Neill, et al., 2000). Moment closure methods have been useful in modeling STD's, tuberculosis, and other diseases (see, e.g., Keeling, 2000). Computational methods for simulating stochastic processes in complex spatial environments or on large networks have started to enable us to simulate more and more complex biological interactions. However, here again, few mathematical scientists have been involved in efforts to bring the power of modern computational methods to bear and it is a challenge to try to remedy this situation.

3. Discrete Mathematics and Theoretical Computer Science

A variety of other potentially useful approaches to epidemiological issues have not yet attracted the attention of many in the mathematical sciences community and some relevant methods of computer science and mathematics are not widely known among epidemiologists and biologists. Specifically, many fields of science, and in particular molecular biology, have made extensive use of the methods of discrete mathematics (***DM***), broadly defined, especially those that exploit the power of modern computational tools. These are guided by the algorithms, models, and concepts of theoretical computer science (***TCS***) that make these tools more available than they have been in the past. Yet, these methods remain largely unused in the study of the transmission dynamics of infectious diseases.

One major development in epidemiology that makes the tools of DM and TCS especially relevant is the use of Geographic Information Systems (GIS). These systems allow analytic approaches to spatial information not used previously (see, e.g., Curtis, 1999, Glass, et al., 1995, Johnson, McDade, and Griffith, 1995, Kitron, et al, 1994, Nuckols, Ellington, and Faidi, 1996, Talbot, Forand, and Haley, 1998,

Vine, Degnan, and Hanchette, 1997, Wartenberg, 1994, Wartenberg, et al., 1997, Yoon, 1995). Another related development is the availability of large and disparate computerized databases on subjects related to disease. This calls for use of modern methods of data mining. Data mining methods of cluster analysis, visualization, and learning, grounded in TCS and statistics, are key to detection/surveillance of bioterrorist attacks and are extremely relevant to spatial/temporal patterning, the recognition that a disease has reached epidemic stage, and the construction of exposure categories. (See, e.g., Carpenter, 2001, Heisterkamp, Doornbos, and Nagelkerke, 2000.)

DM and TCS are also relevant to the increasing emphasis in infectious disease modeling of an evolutionary point of view. To fully understand issues such as immune responses of hosts; coevolution of hosts, parasites, and vectors; drug response; and antibiotic resistance; biologists are increasingly taking approaches that model the impact of mutation, selection, population structure, selective breeding, and genetic drift on the evolution of infectious organisms and their various hosts. (See, e.g., Nowak and May, 2000.) Moreover, recent advances in genomics provide useful tools that could be used to fight bioterrorism. For example, DNA sequencing is routinely used to characterize pathogens' strain phylogenies (see, e.g., Murray, 2002), a critical step in the identification of potential sources of supply of an agent. Epidemiologists are only beginning to become aware of some of the mathematical sciences tools available to analyze these complex problems, e.g., new methods of classification and phylogenetic tree reconstruction grounded in concepts and algorithms of TCS and developed through the explosion in "computational biology." Many recent methods of phylogenetic tree reconstruction resulted from partnerships between mathematical and computer scientists and biological scientists through the DIMACS Special Focus on Mathematical Support for Molecular Biology (See "Mathematical Support for Molecular Biology – Achievements – August 1994 to August 2000," http://dimacs.rutgers.edu/About/specialfocusreports.html.) Yet, a great deal more needs to be done. Many of the phylogenetic techniques were developed for more traditional, well-behaved evolutionary problems (see, e.g., Bush, et al., 1999a,b). However, the traditional model of a binary tree with a small number of species as the leaves is insufficient to capture the "quasispecies" nature of many viruses and the very high substitution rates of retroviruses. Collaboration between mathematical scientists and epidemiologists involving the development and use of phylogenetic methods in epidemiology is likely to take both fields of epidemiology and mathematical science in new and fruitful directions.

Mathematical models that capture the process of one protein interfering with the folding of another seem relevant to diseases such as bovine spongiform encephalopathy ("mad cow" disease or BSE), Alzheimer's Disease, Huntington's Disease, and amyloidosis (Anderson, et al., 1996, Levin, et al., 1997, Nowak, et al., 1998). These models might shed light on important epidemiological questions involving crossing species barriers and dose-response relationships. Models of protein folding are often based on global minimization of energy functions (Pardalos, Shalloway, and Xue, 1996) but increasingly on methods blending DM and TCS (Berger and Singh, 1997, Singh, Berger, and Kim, 1999, Singh, et al., 1998).

The study of spread processes on graphs and networks has a long history in the mathematical analysis of social networks and opinion formation and in the developing theory of distributed computing. These processes, whether deterministic

or stochastic, have had some application in the modeling of diseases, but there is a significant body of mathematical and modeling work that seems very relevant to epidemiology that needs exploration. (Examples of relevant work are Poljak and Sura, 1983, Falmagne, 1997, Peleg, 1998, Mustafa and Pekec, 2001, and Dreyer, 2000.)

In the remainder of this paper, we go into more detail about some of the challenges for DM and TCS, with special emphasis on bioterrorism. Many of the challenges are of interest in and of themselves and new results/methods stimulated by these challenges should be of use in a wide variety of fields.

4. Detailed Challenges for DM/TCS

DM deals with arrangements, designs, codes, patterns, schedules, and assignments. These are all very relevant to aspects of the defense against bioterrorism, as we shall point out. TCS deals with the theory of computer algorithms. During the first 40 years of the computer age, TCS, aided by powerful mathematical methods, in particular DM, but also logic and probability, had a direct impact on technology, by developing models, data structures, algorithms, and lower bounds that are now at the core of computing. DM and TCS have found extensive use in many areas of science and public policy, with computational molecular biology being a specific and important case in point. These tools, which seem very relevant to epidemiological problems and in particular to the defense against bioterrorism, are not very well known among those working on public health problems.

4.1. Detection/Surveillance

A key to the prevention of bioterrorist attacks and to the successful response to them if they cannot be prevented is early detection that an attack is planned or has taken place. A basic requirement is the development of efficient and effective surveillance systems. Many of the most important challenges for DM and TCS arise from this requirement. Surveillance has long been a major function of the public health system. Concern about terrorist-instigated disease outbreaks has renewed interest in surveillance as one of the basic lines of defense.

Disease or event or symptom reporting and surveillance systems represent a primary epidemiological data source for the study of/alert to adverse reactions to medication, emerging diseases, or bioterrorist attacks. These systems synthesize data from millions of reports. There are many challenges for mathematical scientists to investigate current major issues confronting adverse event/disease reporting, surveillance, and analysis. A key issue is speed. Quick detection of a new disease outbreak is critical in determining the ultimate impact of the outbreak. In the US, two major reporting/surveillance systems are AERS and VAERS. Neither is specifically relevant to bioterrorism defense, but issues relating to these databases are. AERS, the Adverse Event Reporting System (http://www.fda.gov/cder/aers/) is a database of drug adverse reactions reported by health professionals and others. AERS is administered by the Food and Drug Administration (FDA). The system contains adverse reactions detected and reported after marketing of a drug for a specified time period. AERS contains over two million cases. VAERS, the Vaccine Adverse Event Reporting System (http://www.vaers.org), is a Cooperative Program for Vaccine Safety of the FDA and the Centers for Disease Control and Prevention

(CDC). VAERS is a post-marketing safety surveillance program, collecting information about adverse events that occur after the administration of US licensed vaccines. Reports are provided by all concerned individuals: patients, parents, health care providers, pharmacists, and vaccine manufacturers. The VAERS database is publicly available and contains over 100,000 reports. It is important to develop new databases that are more specifically geared to bioterrorism defense. The appropriateness of different databases is a research issue of some significance. Other databases being developed and explored in connection with bioterrorism defense include emergency department reports, prescription drug information, and HMO data. Other potential sources of surveillance data under discussion include calls to doctors' offices and nurses' hotlines, hits to certain websites, etc. Even data from mobile or wearable sensors might be useful.

Analyses of surveillance data must confront several difficulties. These include adverse event recognition, underreporting, biases, estimation of population exposure, report quality, and, most importantly, no denominator or control group. In many cases it is difficult to discern whether or not a reported symptom or adverse reaction was from a terrorist event or instead was a consequence of some underlying condition (see, for example, Koch-Weser, Sellers, and Zacest, 1977, Rawlins, 1994, and Strom and Tugwell, 1990). Similar considerations would apply if we were trying to understand whether symptoms came from a natural or terrorist-caused disease. Several methodological issues relating to these reporting mechanisms and others emphasizing disease/symptom reporting form the basis for a research agenda. These include application of computational and statistical methods for early detection of emerging trends; modification of algorithms of streaming data analysis designed to set off early warning alarms; application of data mining methods; development of causal inferential methods in the absence of controls; study of ways to eliminate bias; design of verification methodology. This last issue is especially pressing since large-scale medical record databases that now exist in certain sub-populations (e.g., HMOs, military) can provide a basis both for assessing the quality of surveillance data and for validating analyses. Other issues arise from taking reports supplied in "natural language" and translating them into a form amenable for analysis using formal or semi-formal methods.

New computational methods are needed to deal with a variety of large, complex data sets of various kinds arising in epidemiology, not just the disease and adverse event reporting systems. Notifiable infectious diseases provide a huge testbed of data for surveillance, for planning and evaluation of intervention programs, and for hypothesis generation in etiologic studies (Levin, et al., 1997). Data sets connecting environmental factors and disease can be used in much the same way (Paulu and Ozonoff, 1998, Vieira, et al., 2001). While there is a long history of use of such data sets in epidemiology, much work will be involved in applying or modifying known methods to bioterrorist applications. Some examples of large data sets arising in epidemiological studies involve fat-free body mass (Chagnon, et al., 2000), linkage scans (Province and Single, 2000), the health of Gulf War veterans (Proctor, et al., 2000) and alcoholism (Saccone, et al., 2000). A long list of links to large health-related data sets can be found at the website http://www.ehdp.com/vitalnet/datasets.htm . Such data, while often massive in quantity, is uneven in quality and completeness and heterogeneous in nature, and is typical of data we might expect in bioterrorist surveillance. Work will be required to build on relevant data mining studies in the epidemiological literature (see e.g., Brosssette, et al., 1999, Brossette, et al., 1998, DuMouchel, 1999, Forgionne,

Gangopadhyay, and Adya, 2000, Holmes, Durbin, and Winston, 2000, Openshaw, Turton, and MacGill, 1999, Pendharkar, et al., 1999, Richards, et al., 2001), in particular the development of new algorithmic methods for data mining in epidemiology involving clustering methods. Automatic environmental monitoring and risk evaluation for cancer provides a sample motivation even if cancer is not considered a bioterrorist threat. The issues are similar. In the US, there is a rich history of cancer mapping, highlighted by the release of the first US Cancer mortality atlas in 1975, the recent development of the US Cause of Death Atlas, and the National Cancer Institute's data set of about 10 million US cancer cases. Additional data comes from questionnaires based on individual patient and resident information; public registries with cancer incidences aggregated by county; population-based cancer registries aggregated by city and town; birth and death registries; environmental data such as sample databases of water conditions and air quality records; census data such as geographic databases with accurate locations of population; and remotely-sensed data providing information on land use patterns or air pollution distribution. All of these databases have different temporal and spatial assumptions (for example, different frequencies of collection, different spatial resolution (by state, by county, by zip-code, by square kilometer), etc. In short, the data sets are large, complex, heterogeneous, and contain many errors.

4.1.1. Streaming Data Analysis

Much modern data analysis involves data that streams by so quickly or in such quantity that one has only one shot at it. A similar issue arises when the data is concerned with problems to which a rapid response is essential. This situation is certainly true of much of the data that the intelligence community looks at in trying to assess or detect new terrorist threats. The technical term the TCS community has come to use for the analysis of this type of data is ***streaming data analysis***. (See, for example, Cortes, et al., 2000, Ganti, Gehrke, and Ramakrishnan, 1999 and Ullman, 1999.) Streaming data analysis is widely used to detect trends and sound alarms in telecommunications and finance. For instance, AT&T uses this to detect fraudulent use of credit cards or impending billing defaults and a group at Columbia University has developed methods for detecting fraudulent behavior in financial systems. (For references on intrusion detection and billing faults, see Balasubramaniyan, et al., 1998, Brunken, Mager, and Putzke, 1989, Chan, et al., 1999, Hume, Daniels, and MacLellen, 1999, or Stolfo, et al., 2000. For references on applications to financial systems, see Dhar, 1998, or Podding and Huber, 1999.) Streaming data analysis uses algorithms based in TCS, but these algorithms have not as yet been applied to disease or bioterrorist event detection.

Among the research issues that face mathematical scientists in the attempt to develop methods of streaming data analysis and adapt them to bioterrorist event detection are the need to modify methods of data collection, transmission, processing, and visualization. The use of decision trees, vector-space methods, and Bayesian and neural nets should be explored. One needs to understand how the results of monitoring systems are best reported and visualized and to understand the extent to which they can lead to fast and safe automated responses. We need to understand also how relevant queries are best expressed, giving the user sufficient power while implicitly restraining him/her from incurring unwanted computational overhead.

4.1.2. Cluster Analysis

Cluster analysis offers the promise of pattern extraction from complex data sets involving different temporal and spatial assumptions, arising from different sources, etc. Clustering problems arise in widely varied applications involving molecular databases, astrophysical studies, geographic information systems, software measurement, world-wide ecological monitoring, medicine, and analysis of telecommunications and credit card data for fraud. Methods of clustering developed in these areas (see, for example, Arratia and Lander, 1990, Baritchi, Cook and Holder, 2000, Church, 1997, Godehart, 1990, Karger, 1997, Neil, 1992) need to be modified for applications in epidemiology and in particular in bioterrorist event surveillance and detection. Clustering methods aim at finding within a given set of entities subsets called clusters that are both homogeneous and well-separated and these concepts are central in applications of clustering to many practical problems of detection, decisionmaking, or pattern recognition.

Application of traditional clustering algorithms to bioterrorism defense in particular and epidemiological problems in general is hindered by the extreme heterogeneity of the data; we need to develop new approaches to deal with such heterogeneities. Among the concepts and algorithms that need modification are those used widely in the clustering literature such as maximum sum of splits, minimum sum of diameters, etc. (See DuMerle, et al., 2000, Guenoche, Hansen, and Jaumard, 1991, Hansen, Frank, and Jaumard, 1989, and Hansen and Jaumard, 1987.) Promising algorithmic methodologies for clustering heterogeneous data are in the papers Kempner, Mirkin, and Muchnik [1997], Kuznetsov and Muchnik [1982], and Muchnik and Shvartser [1990]. We also need to build on traditional statistical approaches to heterogeneous epidemiological data (see, e.g., Cox and Piegorsch, 1996a,b, Dominici, et al., 2000, and Patil, 1991).

4.1.3. Visualization of Data

Huge data sets are sometimes best understood by visualizing them. Visualization provides ways to reduce complex quantitative or qualitative information to an easily understood form, such as a map or a color-coded alert. Sheer data sizes require new visualization regimes, which require suitable external memory data structures to reorganize tabular data in secondary storage so that access, usage, and analysis are facilitated (see, e.g., Abello and Vitter, 1999). The mathematical sciences community is faced with the challenge that developing visualization algorithms becomes harder when data arises from various sources and each source contains only partial information, as is the case, for example, with the cancer monitoring data mentioned above and, more generally, with all kinds of epidemiological data.

Visualization focuses on creating insightful visual representations of complex data. Creating the visual representation usually is done in several phases in a visualization pipeline. (See for example Bajaj, C., Pascucci, V., and Schikore, 1997 and Munzner, 1998.) First, the raw data is transformed into a higher-level metaphor (e.g., a graph). Then, the metaphor is transformed into geometry by means of some embedding operation. Finally, the geometry is rendered into an image. Throughout the visualization pipeline emphasis is given to the clarity and aesthetics of these transformations. Quantitative approaches such as minimizing edge crossings (in graphs), area, etc., do not scale when there are more data elements than there are display pixels. New paradigms emphasize perspective views, hierarchical decompositions, overlays, advanced user-interfaces, and real-time rendering

techniques. There has been much work defining models and metaphors but relatively little on formal algorithmic definitions and implementations. Research is required on core algorithmic issues as well as their use in discovery, and also aspects of modeling.

4.1.4. Data Cleaning

In detecting diseases or bioterrorist events, a major problem is that the data is very "dirty." The data is dirty due to manual entry, lack of uniform standards for content and formats, data duplication, and measurement errors. The TCS community has been developing data cleaning tools (see, e.g., Galhardas, et al., 2000a, 2000b, Lambert, et al., 1999) and the website www.research.att.com/~tamr/dataquality.html). Among the methods developed so far are methods known as duplicate removal, merge purge, and automated detection. Much more work is needed to develop such methods and adapt them to the detection and surveillance problems important for bioterrorism defense.

4.1.5. Dealing with Reports in "Natural Language"

Still another set of research issues arises from the use of natural language in surveillance and reporting systems: Devise effective methods for translating natural language input into formats suitable for statistical analysis (prior work on machine natural language processing and information retrieval is relevant); develop computationally efficient methods to provide automated responses consisting of follow-up questions; develop semi-automatic systems to generate queries based on dynamically changing data, indicating developing epidemiological trends. Relevant to these questions is work in VanLehn and Niu [to appear] on interpreting natural language reports based on probabilistic models of context, work in Langlotz, Shortliffe, and Fagan [1986], McConachy, Korb, and Zuckerman [1998], and Zuckerman, McConachy, and Korb [1998] on communicating uncertain information and summarizing rough trends, and work in Horvitz and Paek [1999, 2001] and Walker [2000] on decision-theoretic methods for asking followup questions in natural language processing. Earlier work on electronic surveillance reporting from public health reference laboratories is also very relevant here (e.g., Bean and Martin, 2001, Bean, Martin and Bradford, 1992, Hutwagner, et al., 1997, and Mahon, et al., 1995).

4.1.6. Cryptography and Security

A key difficulty in biosurveillance, as well as in intelligence work in general and also in epidemiological research in general, is the need to protect the privacy of individuals. There is a pressing need to devise effective methods for protecting privacy of individuals about whom data is provided to biosurveillance teams, data from emergency department visits, doctor visits, prescriptions, and possibly biosensors in public places. There is also a pressing need to develop ways to share information between databases of different agencies while protecting privacy. For example, how can we make a simultaneous query to two datasets without compromising information in those datasets? (For example, is individual xx included in both sets?) Issues here include insuring accuracy and reliability of responses, authentication of queries, policies for access control and authorization.

As an example, consider the situation where different agencies need to collaborate in order to determine security-relevant outcomes based on their disjoint data; however, they wish to do so without requiring any agency to reveal all its data to the others. The ability to collaborate without revealing information could be instrumental in fostering inter-agency collaboration. Technically speaking, ***privacy-preserving computation*** enables several parties that have private inputs to compute a joint function of their inputs while ensuring that the computation process itself does not reveal any additional information except for the final output. For example, in the case of computing the intersection of several datasets, no party should learn any information about the other parties' data, except for the fact that the elements that appear in the intersection appear in every data set.

The algorithms community has investigated related problems, such as private information retrieval and secure multi-party computation. Secure multi-party computation (see e.g., Canetti and Genaro, 1996, Cramer, Damgård, and Maurer, 2000, Fitzi, et al., 2001), is a very general and well-studied paradigm that allows two or more parties to evaluate a function of their inputs in such a way that no more is revealed to a party or a set of parties about other parties' inputs than what is implied by their own inputs and outputs, even if some of the parties act maliciously. However, this approach has so far not been generally efficient enough to be used in practice. There has been some amount of work in the databases community on privacy aspects of data access. However, these techniques have been designed for a few specific problems -- data mining and auditing boolean values -- and as such do not address many of the central problems such as finding common elements, finding large join values, merging, etc. New methods are needed for dealing with the set of important data access problems with multiple data sets, and the challenge in performing them in a privacy-preserving manner.

Realistic problems require the development of efficient methods for comparing data sets when the entries could be in different, often incompatible formats and there could be errors in the data. The problem can be described as follows: Suppose we are given two datasets (or two sets of responses from two different datasets) where some of the entries in each response in one dataset correspond to the same entity (for example, the same name, or the same address) in the second dataset and our task is to match an entry in one dataset to most likely data entries in the second datset. The difficulty arises since the names could be slightly misspelled or addresses could be in a different format. (Otherwise a trivial sorting procedure of both datasets would work.) Nevertheless, we wish to find, in an efficient way, all entries that appear to have a lot in common. The naive solution takes each entry in the first dataset and compares it to every entry in the second dataset, which leads to quadratic running time. A reasonable objective is to be able to achieve this task with far greater efficiency. Some preliminary results, which use ideas from dimension reduction, clustering and searching strategy, suggest that we can do significantly better than the trivial quadratic solution, even for higher-dimensional problems. To go beyond this one needs to study how the solution could be further extended to maintain privacy.

4.2. Social Networks

Diseases are often spread through social contact. Contact information is often key to controlling an epidemic, man-made or otherwise. There is already a considerable literature in epidemiology devoted to social networks (see, e.g., Anderson and May,

1991, Blanchard, Bolz, and Kruger, 1990, Kretzschmar and Morris, 1996, Morris and Dean, 1994, Watts, 1999, Watts and Strogatz, 1998), since diseases are often spread through social contact. A major focus is the spread of AIDS and other sexually transmitted diseases, but there is interest in a wide variety of other diseases as well, and in particular some of these, such as smallpox, that are considered serious bioterrorist threats.

4.2.1. Simple Threshold Models of Spread of Disease through a Network

Methods of DM have had a long history in the study of social networks (Degenne and Forse, 1999) since those networks have a natural graph structure.

In the simplest discrete model of spread of disease through a network, we look at dynamically changing networks with change taking place at discrete times. Here, vertices (individuals) are infected or non-infected. An individual becomes infected at time t+1 if sufficiently many of his/her neighbors are infected at time t. We call this the ***threshold model***. Saturation models in economics are analogous as are models of spread of opinions through social networks – we return to this analogy below.

There are complications and variants on the simplest model. We can consider infection only with a certain probability. We can think of individuals as having degrees of immunity and infection taking place only if sufficiently many neighbors are infected and degree of immunity is sufficiently low. We can add a category for recovered individuals or levels of infection.

In these models, a variety of research issues remain. What sets have the property that their infection guarantees the spread of the disease to x% of the vertices? What vertices need to be "vaccinated" to make sure a disease does not spread to more than x% of the vertices? How do the answers to these questions depend upon the network structure? How do they depend upon the choice of threshold? Some of these questions have been studied. See Dreyer [2000] for a survey.

4.2.2. Analogies to Models of Spread of Opinion

An old subject in sociology, psychology, and economics involves dynamic models for how individuals change their opinions over time until, hopefully, some consensus is reached. We think of individuals as vertices of a graph or network, a social network. Edges (directed or undirected) represent influence. The edges might have real numbers or weights on them, representing relative influence. Models of the process range from Markov models (Degroot, 1974, French, 1956, Roberts, 1976) to dynamic models on graphs closely related to neural nets (Goles and Martinez, 1992, Poljak and Sura, 1983, Dreyer, 2000).

In the simplest models, opinions are +1 or -1, each vertex has an opinion at any given time, and opinions change at discrete times. Dreyer (2000) studied the ***k-threshold model*** in which the value (+1 or -1) at a vertex changes at discrete times if at least k of its neighbors have the opposite value. He also studied the ***monotone*** variant of this where change can only go from value -1 to value +1. As noted in the previous subsection, this model is useful in studying the spread of infectious disease: an individual gets infected if at least k of its neighbors are infected, and once infected an individual always remains infected. A similar threshold model

arises in the study of approval voting, where each voter approves of all candidates whose "rating" lies above his threshold. This model is developed in or hinted at in Brams and Fishburn (1983) and Merrill (1988). It should be worthwhile to investigate for *k*-threshold processes probabilistic models where infection takes place only with a certain probability; these are similar to those used for approval voting in Doignon and Regenwetter (1997), Falmagne and Regenwetter (1996), and Regenwetter and Grofman (1998). Along these lines, it should be worthwhile to investigate relationships between *k*-threshold models and the models of stochastic token theory developed in Doignon and Falmagne (1997), Falmagne (1996, 1997), Falmagne and Doignon (1997), Falmagne, Regenwetter, and Grofman (1997), and Regenwetter, Falmagne, and Grofman (1999). Still more complicated models arise when individuals have not only opinions but degrees of confidence in their opinions. In epidemic terms, confidence can be interpreted as degree of immunity and then we are again in the situation discussed in the previous subsection. One model has opinions change if enough neighbors hold the opposite opinion with sufficiently high confidence. If an opinion changes, the confidence in the new opinion goes down and otherwise it goes up, in discrete increments. Similar ideas underly models of Hoffman (1995), Latane (1996), Latane and Fink (1996), Latane and Nowak (1994), and Merrill (1999). Finally, other interesting variants of these approaches would involve adding a "recovered" category, levels of infection, etc., again as noted in the previous subsection.

A variant of the threshold process is the ***majority process*** in which an individual changes opinion (between +1 and -1) at discrete times if a majority of its neighbors have the opposite opinion. While this process does not seem relevant to epidemiology in the case that opinions can fluctuate back and forth between +1 and –1, it could be relevant if opinions can only change from –1 to +1. Thus, it is worth reviewing the relevance to epidemiology of research on majority processes by Mustafa and Pekec (2001), Pekec (1995), Peleg (1998), Flocchini, et al. (1998, 1999) and Luccio (1998). Variants of majority processes have been widely studied under the names "automata networks" and "synchronous neural networks." Majority processes have been applied to the study of social impact (Poljak and Sura, 1983, Roberts, 1976, Roberts, 1979a) through social networks, as have the synchronous neural net models (Latane and Fink, 1996). Automata networks have been widely studied as models in biology, chemistry, and physics and as models of abstract computation. (See, e.g., Bienenstock, Fogelman-Soulie, and Weisbuch, 1986.)

4.2.3. Analogous Models in Distributed Computing

It is also worth reviewing the relevance to epidemics of threshold and majority processes arising in distributed computing. Here, the goal is to eliminate damage caused by failed processors (vertices) or at least to restrict their influence, and this is accomplished for example by maintaining replicated copies of crucial data and, when a fault occurs, letting a processor change "state" if a majority of its neighbors are in a different state. (See Attiyah and Welch, 1998, Bracha, 1987, Dwork, et al., 1998, Lamport, Shostak, and Pease, 1982, Linial, et al., 1993, Peleg, 1996). Other applications of similar ideas in distributed computing arise in distributed database management, quorum systems, and fault-local mending.

4.3. Evolution

The problem of reconstructing phylogenetic trees for viruses and other pathogens is a central one in mathematical modeling of infectious disease systems and has the potential to help us understand and predict the origin, evolution, and likely future development of such pathogens. Thus, models of evolution might shed light on new strains of infectious agents used by bioterrorists. Phylogeny is now a central tool for studies into the origin and diversity of viruses such as HIV (see, e.g., Pybus, Rambaut, and Harvey, 2000, Rambaut, et al., 2001) and dengue fever virus (Rambaut, 2000). These and other investigations have provided new insights, such as identifying the possible pattern of transfer of HIV-type viruses between primate species. Phylogenetic techniques have also proved useful in mapping the evolution of different strains of the human influenza A virus (Bush, et al., 1999a, 1999b), with the goal of predicting which strain is most likely to cause future epidemics, with applications to vaccine development.

Phylogenetic analysis might help in identification of the source of an infectious agent. For example, as noted earlier, DNA sequencing is routinely used to characterize pathogens' strain phylogenies (see, e.g., Murray, 2002), a critical step in the identification of potential sources of supply of an agent that might have been used in a bioterrorist attack. Phylogenetic models can help to identify the subtype of an organism and narrow the search for its origin. Moreover, because an organism has been cultured between times it is passed from lab to lab, phylogenetic analysis can help to identify which lab an organism came from. A major research challenge is to build on phylogenetic methods developed by computational biologists to explore ways in which such methods can be applied and developed to shed new light on the evolution of viruses and other pathogens. The phylogeny problem has benefited greatly from the algorithmic approaches of DM/TCS. Of note are information-theoretic bounds on tree reconstruction methods (Erdos, et al. 1999, Steel and Szekely 1999, Steel and Szekely 2001), optimal tree refinement methods (Bonet, et al. 1998), the disk-covering method (Huson, Nettles, and Warnow, 1999), and maximum parsimony heuristics, nearest-neighbor-joining method, and hybrid methods combining these two approaches (Huson, et al, 1999). Additional topics of note are ways to compare two phylogenetic trees using distance-based approaches, to find the linear-cost subtree-transfer distance between two trees (DasGupta, et al., 1999) or to find the consensus phylogeny among those obtained using alternative reconstruction methods (see, e.g., Janowitz, et al., 2002, LaPointe and Cucumel, 2002, Wilkinson, 1996, Wilkinson and Thorley, 2002).

Many of the phylogenetic techniques in use today were originally developed to investigate more traditional and well-behaved evolutionary problems, where historical relationships are typically represented by a binary tree with a small number of species appearing as the leaves (tip vertices). In epidemiology the picture is more complex and this observation underlies the future research challenges. Even if there is a single underlying tree, it may typically have thousands of vertices, and many of these may be of high degree. Furthermore, data may be available not just for the species at the leaves of the tree, but for species distributed at vertices throughout the tree, particularly when the evolution of a virus is studied by serial sampling in patients. This is true for retroviruses which have a very high substitution rate, and whose molecular evolution may be up to 10^6 times more rapid than eukaryotic or prokaryotic genes (Drummond and Rodrigo, 2000, Rodrigo, et al., 1999, Shankarappa, et al., 1999). New methods for dealing with these complications should be investigated.

To complicate the picture further, it may well be more appropriate to represent the evolution of a virus by a collection of trees, or by a digraph (or network) to recognize the "quasispecies" nature of viruses, such as in the application of split decomposition by Dopazo, Dress, and von Haeseler [1993]. Relating population genetics considerations (currently handled by the "coalescent" model) to phylogeny considerations is also potentially useful (Rodrigo and Felsenstein, 1999). However, even here, theory has yet to be developed. For instance, the fact that retroviral evolution occurs within a host means that viral sequences sampled from different hosts must take account of the different dynamics of between-host transmission histories and within-host viral genealogies (Rodrigo, 1999). This has consequences for the inference of epidemiological parameters based on viral sequences obtained from several hosts, and these should be investigated. Finally, if one wishes to test particular epidemiological hypotheses it would be helpful to have techniques that avoid having to fix attention on one particular tree. This suggests devising fast methods that would average the quantities of interest over all likely trees, weighted by how well they describe the data – a challenge for modern computational tools.

4.4. Decision Making/Policy Analysis

DM/TCS have a close historical connection with mathematical modeling for decision making and policy making. In a decision making context, mathematical models can help us to understand fundamental processes, compare alternative policies and interventions, provide a guide for scenario development, guide risk assessment, aid forensic analysis, and predict future trends. Models can be used to determine policy options and to estimate the consequences of various actions. Ultimately, models are a tool used by decision makers and can be useful for training and educating decision makers in advance of bioterrorist attacks.

4.4.1. Consensus

Many decisions have to be made by groups. DM/TCS is fundamental to the theory of group decision making/consensus. Consensus theory (Arrow, 1951, Barthelemy, 1989. Moulin, 1988, Saari, 2001, Sen, 1970) is based on fundamental ideas in the theory of "voting" and "social choice." The key problem is to combine expert judgments (e.g., choices or rankings of alternatives) to make policy.

Consensus methods are coming to be widely used in biology in the field beginning to be called ***bioconsensus***. Typically, several alternatives (such as alternative phylogenetic trees or alternative molecular sequences) are produced using different methods or models, and then one needs to find a consensus solution. Day and McMorris [1992a] surveyed the use of consensus methods in molecular biology and Day [1995] and Day and McMorris [1992b] gave many references; there are literally hundreds of papers in the field of "alignment and consensus." Kannan [1995] surveyed consensus methods for phylogenetic tree reconstruction. Also of interest are the forthcoming books by Day and McMorris [to appear] and Janowitz, et al. [to appear]. Typical bioconsensus problems involve finding a common pattern in a library of molecular sequences (Galas, Eggert, and Waterman, 1985, Waterman. 1989, Waterman, Galas and Arratia, 1984, Mirkin and Roberts, 1993) and finding a consensus phylogeny given alternative phylogenies.

A special challenge for bioconsensus research in the bioterrorism arena, and more generally in epidemiology, is the fact that instead of many "decision makers" (voters) and few "candidates" as is traditional in consensus theory, we often have few decision makers and many candidates, where the candidates are many different parameters to modify. (For some relevant work, see Dwork, et al., 2000.)

There is a developing algorithmic point of view in consensus theory (see for example Bartholdi, Tovey, and Trick, 1989a,b, 1992): The goal is to find fast algorithms for finding the consensus policy. Surprisingly enough, there are cases where finding the "best alternative" according to some rule is computationally intractable. The algorithmic point of view will be important in the bioterrorism context where finding consensus alternatives quickly and efficiently could be significant. Much work is needed to develop good algorithms, efficient approximations, and tractable heuristic methods. (See Dwork, et al., 2000 and Marcotorchino and Michaud, 1981 for some work along these lines.)

Combining partial information to reach decisions involves consensus methods. Yet, most known results about computational complexity of consensus functions require assumptions about perfect information of individual preferences, and few similar results are known when there is only partial, distributed, or unreliable information or limits to computation. (For some relevant results, however, see Bar-Noy, et al., 1995, Berman and Garay, 1993, Papadimitriou and Yannakakis, 1991.)

4.4.2. Decision Science/Operations Research

Decision science is an old subject fundamentally based on mathematical methods. Modern decision science formalizes utilities and costs/benefits as well as uncertainty and risk. These are ideas that are very critical in decision making about policies to counter bioterrorist attacks. (See for example Kaufman, Meltzer, and Schmid, 1997.) DM/TCS are very important tools in formalizing and solving optimization problems of the type decision scientists are concerned with: Maximizing utility, minimizing pain, … In a bioterrorism context, minimizing number of casualties and minimizing economic cost of an attack are considerations. Yet, some of the modern ideas of decision science are not well known among epidemiologists and public health professionals. For example, mathematical and computational methods dealing with partially ordered algebraic systems form the foundations of the modern theory of measurement (Krantz, et al, 1971, Luce, et al., 1990, Roberts, 1979b, Suppes, et al., 1989). They can be used to determine what conclusions using scales of measurement are "meaningful" (Roberts, 1994, 1999). Measurement theory (a term which has a different connotation in epidemiology) does not seem to be known to practicing epidemiologists. Of particular significance for epidemiology is the following type of question: What kinds of statistical tests are legitimately applied to data where only order matters? (See Marcus-Roberts and Roberts [1987] and Roberts [1994].) This issue is somewhat recognized by epidemiologists, but its order-theoretic subtleties are usually not. A challenge for research is to find ways to modify existing methods of measurement theory to be applicable to epidemiological problems. For the most part, however, challenges for research in decision science as applied to bioterrorism defense are not so much for the development of new methods but for the application of existing tools, appropriately modified, to new applications.

More generally than decision making, in developing appropriate tools for defense against bioterrorist attacks, we will need to solve the types of optimization problems that are of important in the field of operations research. Indeed, OR was developed during World War II in the context of finding optimal solutions to a wide variety of planning and logistics problems in conflict situations. Relevant OR tools include queueing theory, which might be used to design optimal vaccination strategies, sizes of stockpiles of vaccines, allocation of medications, and analysis of bottlenecks in treatment facilities. Inventory theory can help with managing and controlling and distributing supplies of medications/vaccines. Location theory, another important field in OR, could be used to locate vaccination facilities. Scheduling theory is relevant to the planning of a variety of interventions and assignment theory to quarantine planning and assignment of manpower to treatment facilities. Operations researchers are familiar with the development of stochastic versions of deterministic models and the inclusion of stochastic elements will be important in planning for bioterrorism defense. As with decision science, challenges for research are not so much for the development of new methods but for the application of existing tools to new applications.

4.4.3. Game Theory

Game theory has been widely used in military decision making applications and, as such, is directly relevant to the bioterrorism context. Game theory has been used to model conflicts at various levels. Most widely known in game theory are the two-player games, often presented in matrix form. Here, we consider the rows to correspond to strategies available to one player (the attacker) and the columns to correspond to strategies available to a second player (the defender). The entry in a given row and column corresponds to the outcome if the players use the corresponding strategies. The problem in the bioterrorist context is that many of the entries in the matrix are difficult to calculate. Moreover, we might wish to figure out a probability distribution over strategies used by our opponent, and how we do this is difficult. Finding an optimal strategy as the defender, a key problem of two-person game theory, might be very difficult in this context, and research is needed both in determining the entries of appropriate matrices and in developing new methods for dealing with two-person games under such considerable uncertainty or wide ranges of possible values for the matrix entries.

A less-well-known branch of game theory involves multi-person games. The theory of multi-person games has a wide variety of applications in situations of conflict and cooperation and bargaining (Fudenberg and Tirole, 1991, Myerson, 1991). Of particular interest is the use of multi-person game theory to determine appropriate allocation of scarce resources to different players or components of a comprehensive policy. Game-theoretic solutions such as the Shapley value and the nucleolus have long been used to allocate costs to different users in shared projects. Examples of such applications in the past include allocating runway fees to different users of an airport, highway fees to different size trucks, costs to different universities sharing library facilities, fair allocation of telephone calling charges among users, and distribution of costs in multicast routing/multicast packet transmission. (See Feigenbaum, Papadimitriou, and Shenker, to appear, Loehman and Whinston, 1974, Lucas, 1981, Moulin, 1999, Okada, Hashimoto, and Young, 1982, Sharkey, 1995, Shenker and Moulin, 2001, Shubik, 1962, Straffin and Heaney, 1981, Thomas, 1980, Young, 1985.) Application of this theory to bioterrorism defense will require new concepts and ideas.

Game-theoretic and decision-theoretic methods that arise in the study of congestion problems on the Internet involve models of games in which players share a common resource with negative congestion effects. Such situations arise when users send information over a network, share a database, access a web server, or dial in to a service provider. Challenging mathematical issues arise in designing scheduling paradigms so that even when users act selfishly, the network achieves its performance goals (see Shenker, 1995). In a bioterrorism context, one might consider analogous ideas when congestion (such as at vaccine distribution sites) is an issue. Although this is somewhat speculative, it could be worthwhile thinking about game-theoretic methods in connection with vaccine distribution protocols.

Modern game theory has an algorithmic point of view not present in traditional game-theoretical analysis. There is increasingly an emphasis on finding efficient procedures for computing the winner in a game or the appropriate resource allocation corresponding to different game-theoretic solution concepts or simply computing values, optimal strategies, equilibria, and other solution concepts. These concepts can be easy, hard, or even impossible to compute (see Grigoriadis and Khachiyan, 1995, Koller and Megiddo, 1992, 1996). Work is needed to develop this algorithmic point of view in the bioterrorism context.

Cryptographic methods are needed to make sure that in the course of a game, a player's views, utilities, preferences, strategies are kept private. See Dodis, Halevi, and Rabin, [2000] and Linial [1994] for examples of work at the interface between cryptography and game theory. This is especially relevant in the conflict situations arising in bioterrorism defense when we wish to keep our defensive strategies private yet share them with a large number of individuals and agencies.

4.5. Order Theory

Many practical epidemiological problems involve the comparison of one or more quantities. Most often the quantities are rates or proportions leading to a measure of effect or association, but they may also involve distances, exposure categories, job titles, etc. Often the actual values in question are not important, only whether one value is smaller than or larger than a second, i.e., their order. It could be worthwhile to study how fundamental order-theoretic concepts of TCS and DM such as semiorders, interval orders, general partial orders, and lattices (Fishburn, 1985, Trotter, 1992) can be used to improve the results of epidemiological investigations. This topic is a bit more esoteric and may not be as immediately useful as some of the other topics discussed in this paper, but it is worth pursuing. It builds upon a large literature in DM/TCS dealing with order relations, computing them, approximating them, visualizing them, and assigning measures to them. However, it does not build upon a large body of work connecting these ideas to epidemiology, mostly upon the view of several active epidemiologists that these ideas are relevant.

For instance, one can give epidemiological concepts a careful definition in the language of partial orders and explore the use of visualization of order-theoretic concepts in epidemiologic studies. The latter will involve issues such as how best to visualize a poset through clever presentation of its Hasse diagram – an issue of great interest in the field of TCS known as graph drawing. One application of these ideas will arise in the problem of determining cutoffs or boundaries so as to categorize numerical attributes (as is done, for instance, in constructing age categories or

exposure categories in epidemiology). This can be thought of as the problem of assigning a real number to each attribute (e.g., exposure measurement or age) and then considering the set of all subjects whose assigned numbers exceed a threshold – a common construction in the theory of partially ordered sets.

As another example, point lattices may be regarded as a type of order theoretic lattice. The point lattice construction has found uses in epidemiology through visualizing the relationships of all possible contingency tables to various statistics, effect measures, and cut-off choices (see, e.g., Ozonoff and Webster, 1997) and has also been used in statistics (see Narayana, 1979). Challenges arising in extending these ideas include generalizing the concepts to higher-dimensional tables, where there are additional attributes (ordinal, numerical or nominal) besides case status and exposure; and applying lattice-theoretic approaches to measurement error. The point lattices formed by 2 x 2 contingency tables can be represented as n-element strings from a 2 letter alphabet $\{x,y\}$. Measurement errors can be thought of as being caused by a transposition from a substring xy to a substring yx. Similar transpositions have been studied from a more general viewpoint by lattice theorists (see Bennett and Birkhoff, 1994) as a special case of a Newman commutativity lattice. Many references can be found in Bennett and Birkhoff [1994], which also mentions connections with weak Bruhat orders of Coxeter groups. It should be worthwhile to examine how higher dimensional contingency tables relate to what are called multinomial lattices in Bennett and Birkhoff [1994] and study how combinatorial aspects of the Bruhat orders relate to probabilistic questions in epidemiology. Among other things, one can hope that these considerations will give us guidance on how to decide when observed data tables can be explained by chance alone.

As noted in Section 4.4.2, it will also be useful to consider the issue of what kinds of statistical tests are legitimately applied to data where only order matters (Marcus-Roberts and Roberts, 1987, Roberts, 1994); this issue is somewhat recognized by epidemiologists, but its order-theoretic subtleties are usually not. Mathematical and computational methods dealing with ordered algebraic systems form the foundations of the modern theory of measurement (Krantz, et al., 1971, Luce, et al., 1990, Roberts, 1979b, Suppes, et al., 1989) and can be used to analyze this and important related issues such as what conclusions using scales of measurement are "meaningful" (Roberts, 1994, 1999). It should be useful to analyze epidemiological studies from a measurement theory point of view.

4.6. Combinatorial Group Testing

Natural or human-induced epidemics might require us to test many samples from large populations at once. One important modern topic in DM/TCS that arose in epidemiology is the theory of group testing, which arose in connection with testing millions of World War II military draftees for syphilis (Dorfman, 1943). The idea is to avoid testing each individual and instead to divide them into groups and determine if some individual in the group is positive for the disease, updating the process by further subdividing groups that test positive. The subject is very relevant to schemes for large scale blood testing for viruses such as HIV (Gastwirth and Johnson, 1994, Tu, Litvak, and Pagano, 1994) and has clear relevance to massive testing that might be called for in the event of a bioterrorist attack. The modern theory of group testing (Du and Hwang, 2000) is heavily influenced by combinatorial methods, in particular by the methods of combinatorial designs and coding theory, and many modern

algorithmic methods of group testing, developed by theoretical computer scientists, are not yet widely known or used in epidemiology.

Group testing also arises in connection with the mapping of genomes (see, e.g., Balding and Torney, 1996,1997, Bruno, et al., 1995, Farach, et al., 1997, Ngo and Du, 2000). Here, we have a long list of molecular sequences, form a library of subsequences (clones), and test whether or not a particular sequence (a probe) appears in the library by testing to see which clones it appears in. Because clone libraries can be huge, we do this by pooling the clones into groups.

While the theory of combinatorial group testing has been extensively studied by a small group of researchers, it is not widely known. Moreover, to be readily applicable in an epidemiological context, a variety of new developments are needed. For instance, the "test kit" may have a size constraint, limiting the number of items that can be tested at once. Sometimes, an item can only go through a limited number of tests. In some applications, we might have certain items as neighbors or lined up in a linear fashion, and then only a set of consecutive items can be tested together. Most importantly, tests have errors and so the theory of group testing under errors needs further development.

5. Concluding Comments

So, would DM/TCS help with a deliberate outbreak of anthrax? What about a deliberate release of smallpox? Similar approaches, using mathematical models based in DM/TCS, have proven useful in many other fields, to make policy, plan operations, analyze risk, compare interventions, and identify the cause of observed events. Why shouldn't these approaches work in the defense against bioterrorism?

References:

Abello, J.M., and Vitter, J.S. (1999), *External Memory Algorithms*, DIMACS Series, Vol. 50, American Mathematical Society, Providence, RI.

Albert, R., Jeong, H., & Barabasi, A.L. (1999), "Diameter of the world wide web," *Nature,* **401**, 130-131.

Anderson, R.M. (1982), "Transmission dynamics and control of infectious disease agents," in R.M. Anderson and R.M. May (eds), *Population Biology of Infectious Diseases*, 149-176.

Anderson, R.W., Ascher, M.S., and Sheppard, H.W. (1998), "Direct *HIV* cytopathicity cannot account for *CD* 4 decline in *AIDS* in the presence of homeostasis: A worst-case dynamical analysis," *J. AIDS and Human Retrov.*, **17**, 245-252.

Anderson, R.M., Donnelly, C.A., Ferguson. N.M., Woolhouse, M.E., Watt, C.J., Udy, H.J., MaWhinney, S., Dunstan, S.P., Southwood, T.R., Wilesmith, J.W., Ryan, J.B., Hoinville, L.J., Hillerton, J.E., Austin, A.R., and Wells, G.A. (1996), "Transmission dynamics and epidemiology of BSE in British cattle," *Nature*, **382**: 779-788.

Anderson, R. M., and May, R. M. (1991), *Infectious Diseases of Humans*, Oxford University Press, UK.

Antia, R., Koella, J.C., and Perrott, P. (1996), "Models of the within-host dynamics of persistent mycobacterial infections," *Proc. R. Soc. Lond.*, **1**, 257-263.

Antia, R., Levin, B., and May, R.B. (1994), "Within-host population dynamics and the evolution and maintenance of microparasite virulence," *Amer. Naturalist*, **144**, 457-472.

Arratia, R., and Lander, E.S. (1990), ``The distribution of clusters in random graphs," *Adv. Appl. Math.*, **11**, 36-48.

Arrow, K.J. (1951), *Social Choice and Individual Values*, Wiley.

Attiya, H., and Welch, J. (1998), *Distributed Computing: Fundamentals, Simulations and Advanced Topics*, McGraw-Hill, England.

Bailey, N.T.J. (1975), *The Mathematical Theory of Infectious Diseases*, 2nd edition, Hafner, New York.

Balasubramaniyan, J.S., Garcia-Fernandez, J.O., Isacoff, D., Spafford, E.H., and Zamboni, D. (1998), "An architecture for intrusion detection using autonomous agents," *Proc. 14th IEEE Conf. on Computer Security Applications*, Dec 1998.

Brunken, J.N., Mager, R., and Putzke, R.A. (1989), "NEMOS - the network management system for the AT&T long distance network," *Proc. 1989 IEEE Int'l. Conf. on Communications*, vol. 3, pp. 1193 -1197.

Balding, D.J., and Torney, D.C. (1996), “Optimal pooling designs with error detection,” *J. Comb. Theory*, **A74**, 131-140.

Balding, D.J., and Torney, D.C. (1997), “The design of pooling experiments for screening a clone map,” *Fungal Genetics and Biology*, **21**, 302-307.

Bajaj, C., Pascucci, V., and Schikore, D. (1997), “The contour spectrum,” *Proc. IEEE Visualization*, pp. 167-173.

Baritchi, A., Cook, D.J., and Holder, L.B. (2000), ``Discovering structural patterns in telecommunications data,'' in *Proceedings of the Thirteenth Annual Florida AI Research Symposium.*

Bar-Noy, A., Deng, X., Garay, J.A., and Kameda, T. (1995), “Optimal amortized distributed consensus,” *Information and Computation*, **120**, 93-100.

Barthelemy, J.P. (1989), ``Social welfare and aggregation procedures: Combinatorial and algorithmic aspects,'' in F.S. Roberts (ed.), *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*, Vol. 17 of IMA Volumes in Mathematics and its Applications, Springer-Verlag, New York, pp. 39-73.

Bartholdi, J.J., Tovey, C.A., and Trick, M.A. (1989a), “The computational difficulty of manipulating an election,” *Social Choice and Welfare*, **6**, 227-241.

Bartholdi, J.J., Tovey, C.A., and Trick, M.A. (1989b), “Voting schemes for which it can be difficult to tell who won the election,” *Social Choice and Welfare*, **6**, 157-165.

Bartholdi, J.J., Tovey, C.A., and Trick, M.A. (1992), “How hard is it to control an election,” *Mathematical and Computer Modelling*, **16**, 27-40.

Bean, N.H., and Martin, M. (2001), “Implementing a network for electronic surveillance reporting from public health reference laboratories: An international perspective,” *Emerging Infectious Diseases*, **7**.

Bean, N.H., Martin, M., and Bradford, H. (1992), “PHLIS: An electronic system for reporting public health data from remote sites,” *Am. J. Public Health*, **82**, 1273-1276.

Bennett, M.K., and Birkhoff, G. (1994), “Two families of Newman Lattices,” *Algebra Universalis*, **32**, 115-144.

Berger, B., and Singh, M. (1997), "An iterative method for improved protein structural motif recogntion," *J. Comput. Biol.*, **4**, 261-273.

Berman, P., and Garay, J.A. (1993), “Cloture votes: $n/4$-resilient distributed consensus in $t+1$ rounds,” *Mathematical Systems Theory*, **26**, 3-20.

Bienenstock, E., Fogelman-Soulie, F., and Weisbuch, G. (1986) (eds.), *Disordered Systems and Biological Organization*, NATO ASI Series F: Computing and System Sciences, Springer Verlag, Berlin.

Blanchard, P., Bolz, G.F., and Kruger, T. (1990), *Mathematical Modelling on Random Graphs of the Spread of Sexually-transmitted Diseases with Emphasis on HIV Infection*, Springer, Berlin.

Blower, S.M., McLean, A.R., Porco, T.C., Small, P.M., Hopewell, P.C., Sanchez, M.A., and Moss, A.R. (1995), "The intrinsic transmission dynamics of tuberculosis epidemics," *Nature Medicine*, **1,** 815-821.

Blower, S.M., Small, P.M., and Hopewell, P.C. (1996), "Control strategies for tuberculosis epidemics: New models for old problems," *Science*, **273**, 497-500.

Bonet, M., Steel, M., Warnow, T., and Yooseph, S. (1998), "Faster algorithms for solving parsimony and compatibility," *The Journal of Computational Biology*, **5**, 409-422.

Bracha, G. (1987), "An $O(\log n)$ expected rounds randomized Byzantine generals protocol," *Journal of the ACM*, **34**, 910-920.

Brams, S.J., and Fishburn, P.C. (1983), *Approval Voting*, Birkhauser, Boston.

Brauer, F. and Castillo-Chavez, C. (2001), *Mathematical Models in Population Biology and Epidemiology*, Springer-Verlag.

Brossette, S.E., Jones, W.T., Sprague, A.P., Hardin, J.M., and Moser, S.A. (1999), "DMSS: A knowledge discovery system for epidemiologic surveillance," *The International Journal of Knowledge Discovery and Data Mining*.

Brossette, S.E., Sprague, A.P., Hardin, J.M., Waites, K.B., Jones, W.T., and Moser, S.A. (1998), "Association rules and data mining in hospital infection control and public health surveillance," *Journal of the American Medical Informatics Association*, **5**, 373-381.

Bruno, W.J., Balding, D.J., Knill, E.H., Bruce, D., Whittaker, C., Doggett, N., Stallings, R., and Torney, D.C. (1995), "Design of efficient pooling experiments," Genomics, **26**, 21-30.

Bush, R.M., Bender, C.A., Subbarao, K. Cox, N.J., and Fitch, W.M. (1999a), "Predicting the evolution of human influenza A," *Science,* **286**, 1921-1925.

Bush, R.M., Fitch, W.M., Bender, C.A., and Cox, N.J. (1999b), "Positive selection on the H3 Hemagglutinin gene of human influenza virus A," *Mol. Biol. Evol.*, **16,** 1457-1465.

Canetti, R., and Gennaro, R. (1996), "Incoercible multiparty computation," *37th IEEE Symposium on Foundations of Computer Science (FOCS' 96)*.

Carpenter, T.E. (2001), "Methods to investigate spatial and temporal clustering in veterinary epidemiology," *Prev. Vet. Med.*, **48**, 303-320.

Castillo-Chavez, C., and Feng, Z. (1997), "To treat or not to treat: The case of tuberculosis," *Journal of Mathematical Biology*, **35**, 629-659.

Castillo-Chavez, C., and Roberts, F.S. (2002), Report on DIMACS Working Group Meeting: Mathematical Sciences Methods for the Study of Deliberate Releases of Biological Agents and their Consequences, DIMACS Center, Rutgers University, Piscataway, NJ, http://dimacs/rutgers.edu/Workshops/WGDeliberate/FinalReport5-20-02.doc

Chagnon, Y.C., Borecki, I.B., Prusse, L., Roy, S., Lacaille, M., Chagnon, M., Ho-Kim, M.A., Rice, T., Province, M.A., Rao, D.C., and Bouchard, C. (2000), "Genome-wide search for genes related to the fat-free body mass in the Quebec family study," *Metabolism*, **49**, 203-207.

Chan, P., Fan, W., Prodromidis, A., and Stolfo, S. (1999), "Distributed data mining in credit card fraud detection," *IEEE Intelligent Systems*, **14**, 67-74.

Church, K. (1997), "Massive data sets and graph algorithms in telecommunications systems," session on Mathematical, Statistical, and Algorithmic Problems of Very Large Data Sets, American Mathematical Society Meeting, San Diego, CA.

Cortes, C., Fisher, K., Pregibon, D., Rogers, A., and Smith, F. (2000), "Hancock: A language for extracting signatures from data streams," KDD 2000.

Cox, L.H. and Piegorsch, W.W. (1996a), "Combining environmental information I: Environmental monitoring, measurement and assessment," *Environmetrics*, **7**, 299-308.

Cox, L.H. and Piegorsch, W.W. (1996b), "Combining environmental information II: Environmental epidemiology and toxicology," *Environmetrics*, **7**, 309-324.

Cramer, R., Damgård, I., and Maurer, U. (2000), "Efficient general secure multi-party computation from any linear secret-sharing scheme," *Proceedings of EUROCRYPT '00*, Springer Verlag LNCS, Brugge, Belgium.

Curtis, A. (1999), "Using a spatial filter and a geographic information system to improve rabies surveillance data," *Emerging Infectious Diseases*, **5**.

DasGupta, B., He, X., Jiang, T., Li, M., and Tromp, J. (1999), "On the linear-cost subtree-transfer distance between phylogenetic trees," *Algorithmica*, **25**, 176-195.

Day, W.H.E. (1995), ``The sequence analysis and comparison bibliography," http://edfu.lis.uiuc.edu/~class/sequence/

Day, W.H.E., and McMorris, F.R. (1992a), "Critical comparison of consensus methods for molecular sequences," *Nucleic Acids Research*, **20**, 1093-1099.

Day, W.H.E., and McMorris, F.R. (1992b), "Discovering consensus molecular sequences," in *Proceedings of 16. Jahrestagung, Gesellschaft fuer Klassifikation e.V.*, Dortmund, Germany, 1-3 April, 1992, a volume in the series Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, Berlin.

Day, W.H.E. and McMorris, F.R. (to appear), *Axiomatic Consensus Theory in Group Choice and Biomathematics,* SIAM, Philadelphia, PA, to appear.

Degenne, A., and Forse, M. (1999), *Introducing Social Networks*, Sage Publications, London, UK.

Degroot, M.H. (1974), "Reaching a consensus," *J. Amer. Stat. Assoc.*, **69**, 167-182.

Dhar, V. (1998), "Data mining in finance: Using counterfactuals to generate knowledge from organizational information systems," *Information Systems*, **23**, 423-437.

Diekmann, O. and Heesterbeek, J.A.P. (2000), *Mathematical Epidemiology of Infectious Diseases*, Wiley.

Dodis, Y., Halevi, S., and Rabin, T. (2000), "A cryptographic solution to a game theoretic problem," *Advances in Cryptology -- Crypto2000 Proceedings*, Lecture Notes in Computer Science, Vol. 80, pp. 112-130.

Doignon, J-P., and Falmagne, J-C. (1997), "Well graded families of relations," *Discrete Math.*, **173**, 35-44.

Doignon, J-P., and Regenwetter, M. (1997), "An approval-voting polytope for linear orders," *J. Math. Psychol.*, **41**, 171-188.

Dominici, F., Parmigiani, G., Wolpert, R. and Hasselblad, V. (2000), "Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs," *Journal of the American Statistical Association*, **94**, 16-28.

Dopazo, J., Dress, A., and von Haeseler, A. (1993), "Split decomposition: A technique to analyse viral evolution," *Proc. Natl. Acad. Sci. USA*, **90**, 10320-10324.

Dorfman, R. (1943), "The detection of defective members of a large population," *Annals of Math. Stat.*, **14**, 436-440.

Dreyer, P.A., (2000), *Applications and Variations of Domination in Graphs*, Ph.D. Thesis, Department of Mathematics, Rutgers University, New Brunswick, NJ.

Drummond, A., and Rodrigo, A.G. (2000), "Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA (sUPGMA), *Molecular Biology and Evolution*, **17**, 1807-1815.

Du, D-Z, and Hwang, F.K. (2000), *Combinatorial Group Testing and its Applications*, 2nd ed., World Scientific, Singapore.

Du Merle, O., Hansen, P., Jaumard, B., Mladenovic, N. (2000), "An interior point algorithm for minimum sum of squares clustering," *SIAM Journal of Scientific Computing*, **21**, 1485-1505.

DuMouchel, W. (1999), "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system, with discussion," *The American Statistician*, **53**, 177-202.

Durrett, R. (1999), "Stochastic spatial models," *SIAM Review*, **41**, 677-718

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2000), "Rank aggregation, spam resistance, and social choice," IBM Almaden, preprint, 2000.

Dwork, C., Peleg, D., Pippenger, N., and Upfal, E. (1988), "Fault tolerance in networks of bounded degree," *SIAM J. of Computing*, **17**, 975-988.

Erdos, P.L., Steel, M.A., Székely, L.A., and Warnow, T. J. (1999), "A few logs suffice to build (almost) all trees I," *Random Structures and Algorithms*, **14**, 153-184.

Falmagne, J.Cl. (1996), "A stochastic theory for the emergence and the evolution of preference relations," *Math. Soc. Sci.*, **31**, 63-84.

Falmagne, J.Cl. (1997), "Stochastic token theory," *J. Mathematical Psychol.*, **41**, 152-159.

Falmagne, J-C., and Doignon, J-P. (1997), "Stochastic evolution of rationality," *Theory and Decision*, **43**, 107-138.

Falmagne, J-C., and Regenwetter, M. (1996), "Random utility models for approval voting," *Journal of Mathematical Psychology*, **40**, 152-159.

Falmagne, J-C., Regenwetter, M., and Grofman, B. (1997), "A stochastic model for the evolution of preferences," in A.A.J. Marley (ed.*), Choice, Decision and Measurement: Essays in the Honor of R. Duncan Luce,* Erlbaum, Mahwah, NJ.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999), "On power-law relationships of the internet topology, *" Proc. SIGCOMM* '99, 251-262.

Farach, M., Kannan, S., Knill, E., and Muthukrishnan, S. (1997), "Group testing problems with sequences in experimental molecular biology," *Proc. of Sequences '97 (Conference on Compression and Complexity of Sequences*, Positano, Salerno, Italy).

Feigenbaum, J., Papadimitriou, C.H., and Shenker, S. (to appear), "Sharing the cost of multicast transmissions," *J. of Computer and System Sciences*. (For preliminary version, see *Proceedings 32nd ACM Symposium on Theory of Computing*, 2000.)

Fishburn, P.C. (1985), *Interval Orders and Interval Graphs*, Wiley, New York.

Fitzi, M., Garay, J.A., Maurer, U., and Ostrovsky, R. (2001), "Minimal complete primitives for secure multi-party computation," *Advances in Cryptology - CRYPTO '01*, Lecture Notes in Computer Science, Springer-Verlag.

Flocchini, P., Lodi, E., Luccio, F. , Pagli, L. , and Santoro, N. (1998), "Irreversible dynamos in tori," *Proceedings of Europar*, 554-562.

Flocchini, P., Lodi, E., Luccio, F. , Pagli, L. , and Santoro, N. (1999), "Monotone dynamos in tori," *Proceedings of the Sixth Colloquium on Structural Information and Communication Complexity*.

Forgionne, G.A., Gangopadhyay, A., Adya, M. (2000), "Cancer surveillance using data warehousing, data mining, and decision support systems," *Top. Health Inf. Manage.*, **21**, 21-34.

French, J.R.P., Jr. (1956), "A formal theory of social power," *Psychol. Rev.*, **63**, 181-194.

Fudenberg, D., and Tirole, J. (1991), *Game Theory*, MIT Press.

Galhardas, H., Florescu, D., Shasha, D., and Simon, E. (2000a), "An extensible framework for data cleaning," *Proceedings of the 16th Intern. Conf. on Data Engineering* (ICDE}, 312.

Galhardas, H., Florescu, D., Shasha, D., and Simon, E. (2000b), "AJAX: An extensible data cleaning tool," *SIGMOID Conference*, 590.

Galas, D.J., Eggert, M., and Waterman, M.S. (1985), ``Rigorous pattern recognition methods for DNA sequences: Analysis of promoter sequences from E. Coli," *J. Mol. Bio.*, **186**, 117-128.

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999), "Mining very large databases," *IEEE Computer*, **32**, 38-45.

Garnett, G.P., and Waddell, H.C. (2000), "Public health paradoxes and the epidemiological impact of an HPV vaccine," *J. Clin. Virol.*, **19**, 101-111.

Gastwirth, J., and Johnson, W. (1994), "Screening with cost-effective quality control, potential application to HIV and drug testing," *Journal of the American Statistical Association*, **89**, Theory and Methods, 972-993

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall.

Glass, G.E., Schwartz, B.S., Morgan, J.M. III, Johnson, D.T., Noy, P.M., and Israel, E. (1995), "Environmental risk factors for Lyme disease identification with geographic information systems," *Am. J. Public Health*, **85**, 944-948.

Godehart, E. (1990), *Graphs as Structural Models*, 2nd. ed., Friedr. Vieweg & Sohn, Braunschweig, Germany.

Goles, E., and Martinez, S. (1992*), Statistical Physics, Automata Networks, and Dynamical Systems,* Kluwer Academic Publishers, Dordrecht, Netherlands.

Gordon, D.M., and Riley, M.A. (1992), "A theoretical and experimental analysis of bacterial growth in the bladder," *Molecular Microbiology*, **6**, 555-562.

Grenfell, B.T., and Bolker, B.M. (1998), "Cities and villages: Infection hierarchies in a measles metapopulation," *Ecol. Let.*, **1**, 63-70.

Grigoriadis, M.D., and Khachiyan, L.G. (1995), “A sublinear-time randomized approximation algorithm for matrix games,” *Operations Research Letters*, **18**, 53-58.

Guenoche, A., Hansen, P., and Jaumard, B. (1991), ``Efficient algorithms for divisive hierarchical clustering with the diameter criterion," *J. of Classification*, **8**.

Hadeler, K.P. (1997), “Spatial epidemic spread by correlated random walk, with slow infectives,” in R.A. Jarvis et al. (eds.), *Ordinary and Partial Differential Equations, Proc. Conf. Dundee 1996*, Pitman Research Notes in Mathematics Ser. **370**, 18-32.

Hansen, P., Frank, O., and Jaumard, B. (1989), “Maximum sum of splits clustering,” *J. of Classification*, **6**, 177-193.

Hansen, P., and Jaumard, B. (1987), “Minimum sum of diameters clustering,” *J. of Classification*, **4**, 215-226.

Heisterkamp, S.H., Doornbos, G., and Nagelkerke, N.J. (2000), "Assessing health impact of environmental pollution sources using space-time models," *Stat. Med.*, **19**, 2569-2578.

Hethcote, H.W. (1999), "Simulations of pertussis epidemiology in the United States: Effects of adult booster doses," *Math. Biosci.*, **158**, 47-73.

Hethcote, H.W., and Yorke, J.A. (1984), *Gonorrhea Transmission Dynamics and Control*, Springer-Verlag, Berlin.

Hoffman, F. (1995), “Social impact models,” Southeastern Conference on Combinatorics, Graph Theory and Computing, Boca Raton, FL, 1995.

Holmes, J.H., Durbin, D.R., and Winston, F.K. (2000), "Discovery of predictive models in an injury surveillance database: An application of data mining in clinical research," *Proc AMIA Symp*, 359-363.

Horvitz, E. and Paek, T. (1999), "A computational architecture for conversation," *User Modeling Conference*, 201-210.

Horvitz, E. and Paek, T. (2001), "Harnessing models of users' goals to mediate clarification dialog in spoken language systems," *User Modeling Conference*.

Hume, A., Daniels, S., and MacLellen, A. (1999), “Gecko: Tracking a very large billing system,” *USENIX*, 1999.

Huson, D., Nettles, S., Rice, K., Warnow, T., and Yooseph, S. (1999), "The hybrid tree reconstruction method," *The Journal of Experimental Algorithmics*, **4**.

Huson, D., Nettles, S., and Warnow, T. (1999), "Disk-covering, a fast converging method for phylogenetic tree reconstruction," *Journal of Computational Biology*, **6**, 369-386.

Hutwagner, L.C., Maloney, E.K., Bean, N.H., Slutsker, L., and Martin, S.M. (1997), "Using laboratory-based surveillance data for prevention: An algorithm for detecting salmonella outbreaks," *Emerging Infectious Diseases*, **3**, 395-400.

Janowitz, M., LaPointe, F-J., McMorris, F.R., Mirkin, B., and Roberts, F.S. (eds.) (2002), *Bioconsensus*, DIMACS Series, American Mathematical Society, Providence, RI, to appear.

Johnson, D.L., McDade, K., and Griffith, D. (1995), "Seasonal variation in paediatric blood lead levels in Syracuse, NY, USA," *Environmental Geochemistry and Health*, **17**.

Kannan, S. (1995), "A survey of tree consensus criteria and methods," in S. Tavare (ed.), *Proceedings of Phylogeny Workshop*, DIMACS Tech. Rep. 95-48, DIMACS Center, Rutgers University, Piscataway, NJ, 1995.

Karger, D. (1997), "Information retrieval: Challenges in interactive –time manipulation of massive text collections," Session on Mathematical, Statistical, and Algorithmic Problems of Very Large Data Sets, American Mathematical Society Meeting, San Diego, CA.

Kaufman, A.F., Meltzer, M.I., and Schmid, G.P. (1997), "The economic impact of a bioterrorist attack: Are prevention and postattack intervention programs justifiable?" *Emerging Infectious Diseases*, **3**.

Keeling, M., (2000), "Multiplicative moments and measures of persistence in ecology," *J. Theo. Biol.*, **205**, 269-281.

Kempner, Y., Mirkin, B., and Muchnik, I. (1997), "Monotone linkage clustering and quasi-concave set functions," *Applied Mathematics Letters*, **4**, 19-24.

Kitron, U., Pener, H., Costin, C., Orshan, L., Greenberg, Z., and Shalom, U. (1994), "Geographic information system in malaria surveillance: Mosquito breeding and imported cases in Israel," *American Journal of Tropical Medicine*, **50**, 550-556.

Koch-Weser, J., Sellers, E.M., and Zacest, R. (1977), "The ambiguity of adverse drug reactions," *Eur. J. Clin. Pharmacol.*, **11**, 75-78.

Koller, D., and Megiddo, N. (1992), "The complexity of two-person zero-sum games in extensive form," *Games and Economic Behavior*, **4**, 528-552.

Koller, D., and Megiddo, N. (1996), "Finding mixed strategies with small supports in extensive form games," *International J. of Game Theory*, **25**, 73-92.

Krantz, D.H., Luce, R.D., Suppes, P., and Tversky, A. (1971), *Foundations of Measurement*, Vol. I, Academic Press, New York.

Kretzschmar, M., and Morris, M. (1996), "Measures of concurrency in networks and the spread of infectious disease," *Mathematical Biosciences*, **133**, 165-95.

Kuznetsov, E., and Muchnik, I. (1982), "Analysis of the distribution of functions in an organization," *Automation and Remote Control*, **43**, 1325-1331.

Lambert, D., Pinheiro, J., and Sun, D. (1999), "Reducing transaction databases, without lagging behind the data or losing information," Technical Report, Statistics and Data Mining Research Department, Bell Labs, Lucent Technologies, Murray Hill, NJ.

Lamport, L., Shostak, R., and Pease, M. (1982), "The Byzantine generals problem," *ACM Trans. Of Programming Languages and Systems*, **4,** 382-401.

Langlotz, C.P., Shortliffe, E.H., and Fagan, L.M. (1986), "A methodology for computer-based explanation of decision analysis," Stanford University, KSL-86-57.

Lapointe, F-J., and Cucumel, G. (2002), "How many consensus trees?" in M. Janowitz, F-J. LaPointe, F.R. McMorris, B. Mirkin, and F.S. Roberts (eds.), *Bioconsensus*, DIMACS Series, American Mathematical Society, Providence, RI, to appear.

Latane, B. (1996), "The emergence of clustering and correlation from social interaction," in R. Hegelmann and H.O. Peitgen (eds.), *Models of Social Dynamics: Order, Chaos, and Complexity*, Holder, Picker, Tempsky, Vienna, 79-104

Latane, B., and Fink, E. (1996), "Symposium: Dynamical social impact theory and communication," *J. of Communication*, **46**, 4-77.

Latane, B., and Nowak, A. (1994), `"Attitudes as catastrophes: From dimensions to categories with increasing involvement," in R.R. Vallacher and A. Nowak (eds.), *Dynamical Systems in Social Psychology*, Academic Press, New York, 219-249.

Levin, S.A., and Durrett, R. (1997), "From individuals to epidemics," *Phil. Trans. Roy. Soc. Lond.*, **B351**, 1615-1621.

Levin, S.A., Grenfell, B., Hastings, A., and Perelson, A.S. (1997), "Mathematical and computational challenges in population biology and ecosystems science," *Science*, **17**, 334-343.

Linial, N. (1994), "Games computers play: Game-theoretic aspects of computing," in R.J. Aumann and S. Hart (eds.), *Handbook of Game Theory with Economic Applications*, vol. II, chapter 38, pp. 1340-1395.

Linial, N., Peleg, D., Rabinovitch, Y., and Saks, M. (1993), "Sphere packing and local majorities in graphs," *Proc. 2nd ISTCS*, 141-149.

Lipsitch, M., and Levin, B.R. (1997), "The population dynamics of antimicrobial chemotherapy," *Antimicrobial Agents and Chemotherapy*, **41**, 363-370.

Lipsitch, M., and Levin, B.R. (1998), "Population dynamics of tuberculosis treatment," *Int. J. Tuberc. Lung Dis.*, **2**, 187-199.

Loehman, E., and Whinston, A. (1974), "An axiomatic approach to cost allocation for public investment," *Public Finance Q.*, **2**, 236-251.

Lucas, W.F. (1981), “Applications of cooperative games to equitable allocations,” *Proc. of Symposia in Applied Mathematics*, American Mathematical Society, pp. 19-36.

Luce, R.D., Krantz, D.H., Suppes, P., and Tversky, A. (1990), *Foundations of Measurement*, Vol. III, Academic Press, New York.

Luccio, F. (1998), “Almost exact minimum feedback vertex sets in meshes and butterflies,” *Information Processing Letters*, **66**, 59-64.

Marcotorchino, P, and Michaud, P., (1981), “Heuristic approach of the similarity aggregation problem,” *Methods of Oper. Res.*, **43,** 395-404.

Mahon, B.E., Rohn, D.D., Pack, S.R., and Tauxe, R.V. (1995), “Electronic communication facilitates investigation of a highly dispersed foodborne outbreak: Salmonella on the superhighway,” *Emerging Infectious Diseases*, **1**, 94-95.

Marcus-Roberts, H., and Roberts, F.S. (1987), "Meaningless statistics," *J. Educ. Stat.,* **12**, 383-394.

McConachy, R., Korb, K.B., and Zuckerman, I. (1998), "Deciding what not to say: An attentional probabilistic approach to argument presentation," *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (CogSci 98).

Merrill, S. III (1988), *Making Multicandidate Elections More Democratic*, Princeton University Press, Princeton, NJ.

Merrill, S. III (1999), *A Unified Theory of Voting: Directional and Proximity Spatial Models*, Cambridge University Press, New York.

Mirkin, B., and Roberts, F.S. (1993), “Consensus functions and patterns in molecular sequences,” *Bull. Math. Bio.*, **55**, 695-713.

Mollison, D. (ed.) (1995), *Epidemic Models: Their Structure and Relation to Data*, Cambridge University Press.

Morris, M., and Dean, L. (1994), "Effects of sexual behavior change on long-term human immuno-deficiency virus prevalence among homosexual men," *Amer. J. Epidemiol.,* **140,** 217-231.

Moulin, H. (1988), *Axioms of Cooperative Decision Making*, Cambridge.

Moulin, H. (1999), “Incremental cost sharing: characterization by coalition strategy-proofness,” *Social Choice and Welfare*, **16**, 279-320.

Muchnik, I., and Shvartser, L. (1990), "Maximization of generalized characteristics of functions of monotone systems," *Automation and Remote Control*, **51**, 1562-1572.

Munzner, T. (1998), “Exploring large graphs in 3D hyperbolic space,” *IEEE Computer Graphics and Applications*, **18**, 18-23.

Murray, J.D. (1980), *Mathematical Biology*, Springer-Verlag, Berlin.

Murray, M. (2002), “Determinants of cluster distribution in the molecular epidemiology of tubercuolosis,” *Proc. Natl. Acad. Sci. USA*.

Mustafa, N., and Pekec, A. (2001), “Majority consensus and the local majority rule,” *Proceedings of 28th International Coloquium on Automata, Languages and Programming ICALP '01*, Lecture Notes in Computer Science, Springer-Verlag, Vienna, 530-542.

Myerson, R. (1991), *Game Theory*, Harvard University Press, Cambridge, MA.

Narayana, T. (1979), "Lattice path combinatorics with statistical applications," *Mathematical Expositions*, **23**, U. of Toronto Press, Toronto.

Neil, M.D. (1992), ``Multivariate assessment of software products," *J. of Software Testing*, **1**, 17-37.

Ngo, H.Q., and Du, D-Z. (2000), “A survey on combinatorial group testing algorithms with applications to DNA library screening,” in Du, D-Z., Pardalos, P.M., and Wang, J. (eds.), *Discrete Mathematical Problems with Medical Applications*, DIMACS Series, **55**, American Mathematical Society, Providence, RI.

Nowak, M.A., Krakauer, D.C., Klug, A., and May, R.M. (1998), "Prion infection dynamics," *Integrative Biology*, **1**, 3-15.

Nowak, M.A., and May, R.M. (2000), *Virus Dynamics*, University Press, Oxford.

Nowak, M.A., May, R.M., and Anderson, R.M. (1990), "The evolutionary dynamics of *HIV-1* quasispecies and the development of immunodeficiency disease," *AIDS*, **4**, 1095-1103.

Nuckols, J.R., Ellington, D., and Faidi, H. (1996), "Addressing the non-point source implications of conjunctive water use with a geographic information system (GIS)," *Application of Geographic Information Systems in Hydrology and Water Resource Management* (HydroGIS 96), Vienna, 341-348.

Okada, N., Hashimoto, T., and Young, P. (1982), “Cost allocation in water resources development,” *Water Resources Research*, **18**, 361-373.

O'Neill, P.D., Balding, D.J., Becker, N.G., Eerola, M., and Mollison, D. (2000), “Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods”, *J. Roy. Statist. Soc. C,* **49**, 517-542.

Openshaw, S., Turton, I. and MacGill, J. (1999), "Using the geographic analysis machine to analyze limiting long-term illness census data," *Geographic and Environmental Modeling*, **3**, 83-99.

Ozonoff, D., and Webster, T. (1997), "The Lattice diagram and 2x2 tables," in Johnson, B.L., Xintaras, C., and Andrews, J.S. (eds.), *Hazardous Waste: Impacts on Human and Ecological Health* (Proceedings of the 2nd International Congress on Hazardous Waste), Princeton Scientific, 441-458.

Papadimitriou, C.H., and Yannakakis, M. (1991), “On the value of information in distributed decision-making,” *PODC 91*.

Pardalos, P.M., Shalloway, D., and Xue, G. (1996) (eds.), *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, DIMACS Series, **23**, American Mathematical Society, Providence, RI.

Patil, G.P. (1991), "Encountered data, statistical ecology, environmental statistics, and weighted distribution methods," *Environmetrics*, **2**, 377-423.

Paulu, C., and Ozonoff D. (1998), "Exploring associations between residential history and breast cancer risk in a case-control study," *International Society for Environmental Epidemiology/International Society of Exposure Analysis, Annual Conference*, Boston, MA, August 15-18, 1998.

Pekec, A. (1995), “Monochromatic outcomes in majority processes,” preprint, Department of Mathematics, Rutgers University, New Brunswick, NJ, 1995.

Peleg, D. (1996), ``Local majority voting, small coalitions and controlling monopolies in graphs: A review," *Proc. 3rd Colloq. on Structural Information and Communication Complexity*, 170-179.

Peleg, D. (1998), “Size bounds for dynamic monopolies,” *Discrete Applied Math.*, **86**, 263-273.

Pendharkar, P.C., Rodger, J.A., Yaverbaum, G.J., Herman, H. and Benner, M. (1999), "Association, statistical, mathematical and neural approaches for mining breast cancer patients," *Expert Systems with Applications*, **17**, 223-232.

Perelson, A.S., Neumann, A.U., Markowitz, M., Leonard, J.M., and Ho, D.D. (1996), "*HIV*-1 dynamics in vivo: Virion clearance rate infected cell life span and viral generation time," *Science*, **271**, 1582-1586.

Podding, T., and Huber, C. (1999), “Data mining for the detection of turning points in financial time series,” *Proc. Advances in Intelligent Data Analysis*, Lecture Notes in Computer Science, **1642**, 427-436.

Poljak, S., and Sura, M. (1983), “On periodical behavior in society with symmetric influences,” *Combinatorica*, **3**, 119-121.

Proctor, S.P., Heeren, T., White, R.F., Wolfe, J., Borgos, M.S., Davis, J.D., Pepper, L., Clapp, R., Sutker, P.B., Vasterling J.J., and Ozonoff, D. (1998), "Health status of Persian Gulf War veterans: Self-reported symptoms, environmental exposures and the effect of stress," *International Journal of Epidemiology*, **27,** 1000-1010.

Province, M.A., and Single, A. (2000), "Sequential, genome-wide test to identify simultaneously all promising areas in a linkage scan," *Genet. Epid.*, **19**, 301-322.

Pybus, O. Rambaut, A., and Harvey, P.H. (2000), "An integrated framework for the inference of viral population history from reconstructed genealogies," *Genetics* **155**, 1429-1437.

Rambaut, A. (2000), "Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies," *Bioinformatics*, **16**, 395-399.

Rambaut, A., Robertson, D.L., Pybus, O.G., Peeters, M., and Holmes, E. (2001), "Phylogeny and the origin of *HIV-1*," *Nature,* **410**, 1047-1048.

Rand, D.A., Keeling, M., and Wilson, H.B. (1995), "Invasion, stability and evolution to criticality in spatially extended, artificial host-pathogen ecologies," *Proc. R. Soc. Lond.*, **259**, 55-63.

Rawlins, M.D. (1994), "Pharmacovigilance: Paradise lost, regained or postponed?," The William Withering Lecture," *J. R. Coll. Physicians London*, **29**, (1995), 41-49.

Regenwetter, M., Falmagne, J-C., and Grofman, B. (1999), "A stochastic model of preference change and its application to 1992 presidential election panel data," *Psychological Review*, **2**, 362-384.

Regenwetter, M., and Grofman, B. (1998), "Approval voting, Borda winners and Condorcet winners: evidence from seven elections," *Management Science*, **44**, 520-533.

Richards, G., Rayward-Smith, V.J., Sonksen, P.H., Carey, S., and Weng, C. (2001), "Data mining for indicators of early mortality in a database of clinical records," *Artif Intell Med*, **22**, 215-231.

Roberts, F.S. (1976), *Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems*, Prentice-Hall, Englewood Cliffs, NJ.

Roberts, F.S. (1979a), "Graph theory and the social sciences," in R. Wilson and L. Beineke (eds.), *Applications of Graph Theory*, Academic Press, 255-291.

Roberts, F.S. (1979b), *Measurement Theory, with Applications to Decisionmaking, Utility, and the Social Sciences*, Addison-Wesley, Reading, MA.

Roberts, F.S. (1994), "Limitations on conclusions using scales of measurement," in Barnett, A., Pollock, S.M., and Rothkopf, M.H. (eds.), *Operations Research and the Public Sector*, Elsevier, Amsterdam, 621-671.

Roberts, F.S. (1999), "Meaningless statements," in R.L. Graham, J. Kratochvil, J. Nesetril, and F.S. Roberts (eds.), *Contemporary Trends in Discrete Mathematics*, DIMACS Series, **49**, American Mathematical Society, Providence, RI, 257-274.

Rodrigo, A.G. (1999), "*HIV* evolutionary genetics [Commentary]," *Proceedings of the National Academy of Sciences*, **96**, 10559-10561.

Rodrigo, A.G., and Felsenstein, J. (1999), "Coalescent approaches to *HIV*-1 Population Genetics," in K.A. Crandall (ed.), *Molecular Evolution of HIV*, Johns Hopkins University Press.

Rodrigo, A.G., Shpaer, E.G., Delwart, E.L., Iversen, A.K.N., Gallo, M.V., Jürgen Brojatsch, J., Hirsch, M.S., Walker, B.D., and Mullins, J.I. (1999), "Coalescent estimates of *HIV*-1 generation time in vivo," *Proceedings of the National Academy of Science*, **96**, 2187-2191.

Rohani, P., Earn, D.J.D., and Grenfell, B.T. (2000), "Impact of immunisation on pertussis transmission in England and Wales," *LANCET*, **355**, 285-286.

Saari, D.G. (2001), *Chaotic Elections! A Mathematician Looks at Voting*, American Mathematical Society, Providence, RI.

Saccone, N.L., Kwon, J.M., Corbett, J., Goate, A., Rochberg, N., Edenberg, H.J., Foroud, T., Li, T., Begleiter, H., Reich, T., and Rice, J.P. (2000), "A genome screen of maximum number of drinks as an alcoholism phenotype," *Am. J. Med. Genet.*, **96**, 632-637.

Schuette, M.C. and Hethcote, W.H. (1999), "Modeling the effects of varicella vaccination programs on the incidence of chickenpox and shingles," *Bull. Math. Biol.*, **61**, 1031-1064.

Sen, A.K. (1970), Collective Choice and Social Welfare, Holden-Day.

Shankarappa, R., Margolick, R.B., Gange, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.-L., and Mullins, J.I. (1999), "Consistent viral evolutionary changes associated with progression of *HIV*-1 infection," *Journal of Virology*, **73**, 10489-10502.

Sharkey, W. (1995), “Network models in economics,” in M.O. Ball, et al . (eds.), *Network Routing*, Handbook in Operations Research and Management Science, Vol. 8, North-Holland, Amsterdam, 1995.

Shenker, S. (1995), “Making greed work in networks,” *ACM/IEEE Transactions on Networking*, **3**, 819-831.

Shenker, S., and Moulin, H. (2001), “Strategyproof sharing of submodular costs: Budget balance versus efficiency,” *Economic Theory*, **18**, 511-533.

Shubik, M. (1962), “Incentives, decentralized controls, the assignment of joint costs and internal pricing,” *Management Sci.*, **8**, 325-343.

Singh, M., Berger, B., and Kim, P.S. (1999), "LearnCoil-VMF: Computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins," *J. Mol. Bio.*, **290**, 1031-1041.

Singh, M., Berger, B., Kim, P.S., Berger, J.M., and Cochran, A.G. (1998), "Computational learning reveals coiled coil-like motifs in histidine kinase linker domains," *Proc. Natl. Acad. Sci. USA*, **95**, 2738-2743.

Stauffer, D., and A. Aharony, A. (1992), *Introduction to Percolation Theory*, 2nd edition, Taylor and Francis, London.

Steel, M.A., and Székely, L.A. (1999), "Inverting random functions," *Annals of Combinatorics*, **3**, 103-113.

Steel, M.A., and Székely, L.A. (2001), "Inverting random functions II: Explicit bounds for parametric and non-parametric MLE, with applications", preprint.

Stolfo, S., Fan, W., Lee, W., Prodromidis, A., and Chan, P. (2000), "Cost-based modeling for fraud and intrusion detection: Results from the JAM project," *Proc. DARPA Information Survivability Conference and Exposition*, IEEE Computer Press, vol. 2, pp. 130-144.

Straffin, P.D., and Heaney, J.P. (1981), "Game theory and the Tennessee Valley Authority," *Int. J. of Game Theory*, **10**, 35-43.

Strom, B.L., and Tugwell, P. (1990), "Pharmacoepidemiology: Current status, prospects, and problems," *Ann. Intern. Med.*, **113**, 179-181.

Suppes, P, Krantz, D.H., Luce, R.D., and Tversky, A. (1989), *Foundations of Measurement*, Vol. II, Academic Press, New York.

Szklo, M., and Nieto F.J. (2000), *Epidemiology, Beyond the Basics*, Aspen, Gaithersberg, Maryland.

Talbot, T. O., Forand, S. P., and Haley, V. B. (1998), "Geographic analysis of childhood lead exposure in New York State," *3rd National GIS in Public Health Conference*.

Thomas, A. (1980), *A Behavioral Analysis of Joint Cost Allocation and Transfer Pricing*, Arthur Andersen and Company Lecture Series, Stipes Publishing Company.

Trotter, W.T. (1992), *Combinatorics and Partially Ordered Sets: Dimension Theory*, The Johns Hopkins University Press, Baltimore, MD.

Tu, X., Litvak, E., and Pagano, M. (1994), "Screening tests: Can we get more by doing less," Studies of Aids and HIV Surveillance, *Statistics in Medicine*, **13**, 1905-1919.

Ullman, J.D. (1999), "The MIDAS data-mining project at Stanford," *Proc. 1999 IEEE Symp. on Database Engineering and Applications*, pp. 460-464.

VanLehn, K., and Niu, Z. (to appear), "Bayesian student modeling, user interfaces and feedback: A sensitivity analysis," *International Journal of Artificial Intelligence in Education*, **12**.

Vieira , V., Webster, T, Aschengrau, A, and Ozonoff, D. (2001), "A method for spatial analysis of risk in a population-based case-control study, *International Journal of Hygiene and Environmental Health*, submitted.

Vine, M., Degnan, D., and Hanchette, C. (1997), "Geographic information systems: Their use in environmental epidemiologic research," *Environ. Health Perspect.*, **105**, 598-605.

Walker, M.A. (2000), "An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email," *Journal of Artificial Intelligence Research*, **12**, 387-416.

Wartenberg D. (1994), "Use of geographic information systems for risk screening and epidemiology," in J.S. Andrews, Jr., H. Frumkin, B.L. Johnson, M.A. Mehlman, C. Xintaras, and J.A. Bucsela (eds.), *Hazardous Waste and Public Health: Intl. Congress on Health Effects of Hazardous Waste,* Princeton Scientific Publishing Co., Princeton, NJ, 853-859.

Wartenberg, D., Greenberg, M., Lathrop, R., Manning, R., and Brown, S. (1997), "Using a geographic information system to identify populations living near high voltage electric power transmission lines in New York State," *World Computer Graphics Foundation*, 20.

Waterman, M.S. (1989), "Consensus patterns in sequences," in M.S. Waterman (ed.), *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton, FL, 93-115.

Waterman, M.S., Galas, D., and Arratia, R. (1984), "Pattern recognition in several sequences: Consensus and alignment," *Bull. Math. Biol.,* **46**, 515-527.

Watts, D.J. (1999), *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, NJ.

Watts, D.J., and Strogatz, S.H. (1998), "Collective dynamics of 'small-world' networks," *Nature*, **393**, 440-442.

Wigginton, J.E., and Kirschner, D.E. (2001), "A model to predict cell-mediated immune regulatory mechanisms during human infection with mycobacterium tuberculosis," *J. Immunology*, **166**, 1951-1967.

Wilkinson, M., (1996), "Majority-rule reduced consensus methods and their use in bootstrapping," *Molecular Biology and Evolution*, **13**, 437-444.

Wilkinson, M., and Thorley, J. (2002), "Reduced consensus and supertree methods," in M. Janowitz, F-J. LaPointe, F.R. McMorris, B. Mirkin, and F.S. Roberts (eds.), *Bioconsensus*, DIMACS Series, American Mathematical Society, Providence, RI, to appear.

Yoon, S.S. (1995), "Geographical information systems: A new tool in the fight against schistosomiasis," in M.J.C. de Lepper, H.J. Scholten, and R.M. Stern (eds.), *The Added Value of Geographical Information Systems in Public and Environmental Health*, Kluwer Academic Publishers, Boston, 201-213.

Young, H.P. (1985), *Cost Allocation: Methods, Principles, Applications*, North-Holland, Amsterdam.

Zuckerman, I., McConachy, R., and Korb, K.B. (1998), "Bayesian reasoning in an abductive mechanism for argument generation and analysis," *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (AAAI 98).