Discrete Methods in Epidemiology

James Abello

Graham Cormode

DIMACS and Ask.com Research
E-mail address: abello@dimacs.rutgers.edu

DIMACS and Bell Labs — Lucent Technologies
E-mail address: graham@dimacs.rutgers.edu

2000 *Mathematics Subject Classification.* Primary 92D30 Epidemiology,
Secondary 68R01 Discrete Mathematics in Relation to Computer Science

Contents

Foreword

The DIMACS working group on Data Mining and Epidemiology held a meeting on March 18-19, 2004 at Rutgers University. We would like to express our appreciation to James Abello, Graham Cormode, Kenton Morgan and David Ozonoff for their efforts to organize and plan this successful conference.

The meeting was part of the 2002-2007 Special Focus on Computational and Mathematical Epidemiology, and was organized by one of a number of special focus research groups called "working groups" as part of the special focus. We extend our thanks to Martin Farach-Colton, Sunetra Gupta, Donald Hoover, David Krakauer, Simon Levin, Marc Lipsitch, David Madigan, Megan Murray, S. Muthukrishnan, David Ozonoff, Fred Roberts, Burton Singer and Daniel Wartenberg for their work as special focus organizers.

The meeting brought together researchers who approach the study of epidemiology from a variety of disciplines, some applied and some theoretical. These included computer scientists, mathematicians, statisticians, and biologists together with both descriptive and analytical epidemiologists. The goal of the meeting as well as of this volume is the exploration of cross-disciplinary approaches to the study of epidemiology.

DIMACS gratefully acknowledges the generous support that makes these programs possible. Special thanks go to the National Science Foundation, New Jersey Commission on Science and Technology, Office of Naval Research, Alfred P. Sloan Foundation, Burroughs-Wellcome Fund, and to DIMACS partners at Rutgers, Princeton, AT&T Labs - Research, Bell Labs, NEC Laboratories America, and Telcordia Technologies, and affiliate partners Avaya Labs, HP Labs, IBM Research, Microsoft Research, and Stevens Institute of Technology.

Fred S. Roberts
DIMACS Director

Robert Tarjan
Co-Director for Princeton

Preface

Faced with the question of the intended audience for a collection such as the one assembled here, we have been asking ourselves the following questions:

(a) What is epidemiology? Who is an epidemiologist?

(b) What are the flavors of epidemiology?

(c) What are the types of questions epidemiologists work on?

Dave Ozonoff provided to us the following quote by Rothman that sheds some light on possible answers to the questions posed above:

> "Unfortunately, there seem to be more definitions of epidemiology than there are epidemiologists. Some have defined it in terms of its methods. While the methods of epidemiology may be distinctive, it is more typical to define a branch of science in terms of its subject matter rather than its tools... If the subject of epidemiologic inquiry is taken to be the occurrence of disease and other health outcomes, it is reasonable to infer that the ultimate goal of most epidemiologic research is the elaboration of causes that can explain patterns of disease occurrence."

In general terms, epidemiology deals with populations rather than individuals. One of its goals is to study the frequency of occurrences of health related events. It has a major but not exclusive concern with causes and determinants of disease patterns in populations. The premise is that a systematic investigation of different populations can identify causal and preventive factors. Epidemiology is an observational rather than an experimental science. Sample questions take the form of:

- Does population exposure to x increase the risk of a disease w?
- Are dietary supplements $\{x, y, z\}$ beneficial in lowering the risk of malady w?
- Do behavioral interventions reduce risk behaviors?

We have observed that occurrence measures, causal inference and study designs play prominent roles in the daily endeavors of a typical epidemiologist. Descriptive and analytical epidemiology are two overlapping flavors of this discipline.

Descriptive epidemiology attempts to describe patterns of disease according to spatial and temporal information about the members of a population. These patterns are described by tabulations or summaries of surveys and polls or by parametric or non-parametric population models. Models are in general global descriptions of the major part of a data set. Patterns on the other hand are local features of the data that can be described by association rules, mode or gaps in density functions, outliers, inflection points in regressions, symptom clusters, geographic hot spots, etc. Some epidemiologists appear more interested in local patterns rather than in global structure. This raises questions of how "realistic" certain patterns are.

Analytical Epidemiology attempts to explain and predict the state of a population's health. A typical goal is to summarize the relationship between exposure and disease incidence by comparing two measures of disease frequency. These comparisons may be affected by chance, bias and by the presence or absence of an effect. This explains naturally why statistical methods play a major role in Epidemiology since bias is a central preoccupation of its practitioners. Bias means a systematic error that results in an incorrect or invalid estimate of the measure of association. This can create or mask associations. Selection and information bias are two of the main bias types. In particular, selection shall be independent of exposure if the purpose of the study is to explain the relationship between exposure and disease occurrence. In summary, one of the central themes in analytical epidemiology is to understand the roles of bias, chance and real effect in the understanding of populations health.

To evaluate the role of chance, statistical hypothesis testing and estimation appear to be the tools of choice. On the other hand, generative models offer a way to describe infectious disease dynamics. Since disease patterns are of primary interest, data mining algorithms and detection of rules for pattern formation have a lot to offer. Classification and taxonomies are useful tools to develop predictive models. In general we believe that some questions addressed by epidemiologists benefit from viewing them in a mathematical and algorithmic context. This volume is a first attempt to bridge the gap between the two communities. Its main emphasis is on discrete methods that have successfuly addressed some epidemiological question. We begin by providing introductory chapters, on some of the key methods from discrete data mining, by a selection of researchers in this area; and on descriptive epidemiology by Schneider. These collect, in a digested form, what we believe are among the most potentially useful concepts in data mining and epidemiology.

Next there are two chapters reporting work in epidemiology that suggest a discrete, analytical approach: Shannnon on challenges in molecular data analysis, and Hirschman and Damianos on a system for monitoring news wires for indications of disease outbreaks. The remainder of the volume draws out further some of the key areas in the intersection between epidemiology and discrete methods. The technique of formal concept analysis, and the amazing depth of mathematical structure that arises from it is explored in chapters by Ozonoff, Pogel and Hannan, and Abello and Pogel. The dynamics of disease transmission can be modeled in a variety of ways, but often involves setting up systems of differential equations to model the ebb and flow of infection, as demonstrated by Desai, Boily, Mâsse and Anderson, and Vázquez, in the context of quite different problems. Eubank, Kumar, Marathe, Srinivasan and Wang study massive interaction graphs and give results by a combination of combinatorial methods and simulation; Abello and Capalbo focus on properties of graphs generated by an appropriate random model; while Hartke takes a combinatorial model of disease spread on tree graphs. Finally, we see two applications of Support Vector Machines to epidemiological data sets, from Li, Muchnik and Schneider (using breast cancer data from the SEER database) and from Fradkin, Muchnik, Hermans and Morgan (using data on disease in chickens). Some other potential areas of interest that we have not touched in this collection relate to patient confidentiality, coding and cryptography and multiscale inference.

We hope the volume helps to foster cooperation between epidemiologists, computer scientists and mathematicians. We believe this will help elucidate the main algorithmic and mathematical issues. In a relatively brief period of time we noticed a variety of interconnections between the disciplines, far richer than we ever dreamed of. We trust that the papers included here are a good indicator of the possibilities that discrete mathematical thinking can offer to a variety of epidemiological questions.

James Abello
Graham Cormode
Piscataway, NJ, 2005

Acknowledgments

This volume grew out of the activities of the DIMACS Working Group on Data Mining and Epidemiology. This working group is part of the Special Focus on Computer Science and Epidemiology funded by NSF EIA Grant 02-05116. The themes of the working group and associated events can be accessed by visiting
`http://dimacs.rutgers.edu/Workshops/WGDataMining/` and
`http://dimacs.rutgers.edu/SpecialYears/2002_Epid/episeminars.html`.

The papers appearing in this volume are chiefly the results of interactions promoted by the working group. The editors want to express their appreciation to the DIMACS staff for their assistance with the logistics of organizing meetings of the Working Group, the DIMACS directorate for the funding, the members of the DIMACS Computational Methods Group for their insightful talks, Dave Ozonoff and Dona Schneider for sharing their epidemiological expertise, F. Roberts, T. Imielinski, Apostolos Gerasoulis, S. Muthukrishnan, and S. Sudarsky for their continued support.

All the speakers at meetings of the working group, contributing authors, the many anonymous referees of chapters, and working group participants deserve special credit for helping in one way or another make this volume possible.

J. Abello, DIMACS and Ask.com Research
G. Cormode, DIMACS and Bell Labs — Lucent Technologies