



DIMACS

Series in Discrete Mathematics
and Theoretical Computer Science

Volume 70

Discrete Methods in Epidemiology

James Abello
Graham Cormode
Editors



American Mathematical Society

Discrete Methods in Epidemiology

This page intentionally left blank

DIMACS

Series in Discrete Mathematics
and Theoretical Computer Science

Volume 70

Discrete Methods in Epidemiology

James Abello
Graham Cormode
Editors

Center for Discrete Mathematics
and Theoretical Computer Science

A consortium of Rutgers University, Princeton University,
AT&T Labs–Research, Bell Labs (Lucent Technologies),
NEC Laboratories America, and Telcordia Technologies
(with partners at Avaya Labs, HP Labs, IBM Research,
Microsoft Research, and Stevens Institute of Technology)



American Mathematical Society

This DIMACS volume presents the proceedings from the DIMACS workshop on Data Mining and Epidemiology held at the DIMACS Center, Rutgers University, March 18–19, 2004.

2000 *Mathematics Subject Classification*. Primary 92D30; Secondary 68R01.

Library of Congress Cataloging-in-Publication Data

Discrete methods in epidemiology : DIMACS Workshop, Data Mining and Epidemiology, March 18–19, 2004, DIMACS Center, Rutgers University/James Abello, Graham Cormode, editors.

p. cm. — (DIMACS series in discrete mathematics and theoretical computer science; 70)

“Volume grew out of the activities of the DIMACS Working Group on Data Mining and Epidemiology”—Acknowledgement p.

Includes bibliographical references and index.

ISBN 0-8218-3754-0 (alk. paper)

1. Epidemiology—Congresses. 2. Epidemiology—Mathematics—Congresses. 3. Computer science—Congresses. I. Abello, James M. II. Cormode, Graham, 1977- III. DIMACS Workshop, Data Mining and Epidemiology (2004 : Rutgers University) IV. DIMACS Working Group on Data Mining and Epidemiology (2004 : Rutgers University) V. Series.

RA651.D57 2006
614.4—dc22

2005057092

AMS softcover ISBN 978-0-8218-4379-6

Copying and reprinting. Material in this book may be reproduced by any means for educational and scientific purposes without fee or permission with the exception of reproduction by services that collect fees for delivery of documents and provided that the customary acknowledgment of the source is given. This consent does not extend to other kinds of copying for general distribution, for advertising or promotional purposes, or for resale. Requests for permission for commercial use of material should be addressed to the Acquisitions Department, American Mathematical Society, 201 Charles Street, Providence, Rhode Island 02904-2294, USA. Requests can also be made by e-mail to reprint-permission@ams.org.

Excluded from these provisions is material in articles for which the author holds copyright. In such cases, requests for permission to use or reprint should be addressed directly to the author(s). (Copyright ownership is indicated in the notice in the lower right-hand corner of the first page of each article.)

© 2006 by the American Mathematical Society. All rights reserved.

Reprinted by the American Mathematical Society, 2007.

The American Mathematical Society retains all rights
except those granted to the United States Government.

Copyright of individual articles may revert to the public domain 28 years
after publication. Contact the AMS for copyright status of individual articles.

Printed in the United States of America.

∞ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.

Visit the AMS home page at <http://www.ams.org/>

10 9 8 7 6 5 4 3 2 1 12 11 10 09 08 07

Contents

Foreword	vii
Preface	ix
Acknowledgments	xi
Selected Data Mining Concepts JAMES ABELLO, GRAHAM CORMODE, DMITRIY FRADKIN, DAVID MADIGAN, OFER MELNIK, and ILYA MUCHNIK	1
Descriptive Epidemiology: A Brief Introduction DONA SCHNEIDER	41
Biostatistical Challenges in Molecular Data Analysis WILLIAM D. SHANNON	63
Mining Online Media for Global Disease Outbreak Monitoring LYNETTE HIRSCHMAN and LAURIE E. DAMIANOS	73
Generalized Contingency Tables and Concept Lattices DAVID OZONOFF, ALEX POGEL, and TIM HANNAN	93
Graph Partitions and Concept Lattices JAMES ABELLO and ALEX POGEL	115
Using Transmission Dynamics Models to Validate Vaccine Efficacy Measures Prior to Conducting HIV Vaccine Efficacy Trials KAMAL DESAI, MARIE-CLAUDE BOILY, BENOÎT MÂSSE, and ROY M. ANDERSON	139
Causal Tree of Disease Transmission and The Spreading of Infectious Diseases ALEXEI VÁZQUEZ	163
Structure of Social Contact Networks and Their Impact on Epidemics STEPHEN EUBANK, V. S. ANIL KUMAR, MADHAV V. MARATHE, ARAVIND SRINIVASAN, and NAN WANG	181
Random Graphs (and the Spread of Infections in a Social Network) JAMES ABELLO and MICHAEL CAPALBO	215
Attempting to Narrow the Integrality Gap for the Firefighter Problem on Trees STEPHEN G. HARTKE	225

Influences on Breast Cancer Survival via SVM Classification in the SEER Database	
JIXIN LI, ILYA MUCHNIK, and DONA SCHNEIDER	233
Validation of Epidemiological Models: Chicken Epidemiology in the UK	
DMITRIY FRADKIN, ILYA MUCHNIK, PATRICK HERMANS, and KENTON MORGAN	243
Index	257

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

This page intentionally left blank

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, modes 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 successfully 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 D. 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: Shannon 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
G. Cormode, DIMACS and Bell Labs — Lucent Technologies

This page intentionally left blank

This page intentionally left blank

Index

- Activity Duration, 193
- Adaboost, 25
- Agent, 164
- Alembic, 80
- Alerting, 82
- Alias-I, 80
- Association Rules, 10–12, 109

- B-Tree, 32
- Batched, 30
- Bayesian Analysis, 21–23, 68, 92
- Biclique, 102, 118
- Bigraph Inducing Formal Context, 118
- Binary Relation, 100, 116
- Binning, 75
- Birth, *see also* Natality
- Boolean
 - Algebra, 126
 - Queries, 88
 - Variable, 227, 236–238
- Boosting, 25
- Bootstrap Method, 147, 255
- Bradford-Hill, Sir Austin, 43, 60
- Branching Process, 164
- Browsing, 83
- Buffer Trees, 34

- Cancer
 - Breast, 53, 233, 236, 255
 - Colon-Rectum, 134
 - Liver, 45
 - Lung, 27, 43, 56, 64
 - Skin, 64
- Case
 - Report, 51
 - Series, 51
- Categorical Values, 2, 101
- Causal Tree, 163
- Census Data, 46, 184
- Character Set Encodings, 87
- Chernoff Bounds, 219, 221
- Choropleth, 48
- Chung-Lu Model, 181, 183, 194, 195, 206, 210–212

- Classification
 - Accuracy, 27, 76, 241
 - Cross-classification, 93
 - Model, 235, 246
 - of Text, 13, 27
 - Rule, 27
- Classifier
 - Binary, 18
 - Linear, 14
- Clinical and Molecular Data, Correlation of, 68
- Closed Set Contingency Table, 102
- Closed Sets, 100
- Closure System, 102
- Clustering
 - Algorithm, 4
 - Analysis, 9, 68
- Clustering Coefficient, 188, 189, 193, 210, 212
- Complete
 - Bipartite Graph, 118
 - Lattice, 102
- Complete Mixing, 183
- Concept Explorer, 114, 125
- Concept Lattice, 11, 93, 115
- Concurrency, 168
- Conditional Event, 10, 120
- Confidence Interval, 22, 147, 238
- Configuration Model, 183, 194, 195, 210, 212
- Confounding, 42, 43, 60, 155
- Contact Process, 163
- Contingency Table, 64, 93, 115
- Cross-language Retrieval, 83
- Cross-sectional Studies, 50–52
- Cross-tabulation (Cross-tabs), 93
- Cross-validation, 244, 245
- Cumulative Incidence Ratio (CIR), 140

- Data Capture, 73
- Data Cleaning, 3
- Data Collection, 2, 155
- Demographic Data, 46, 58
- Demographic Mixing, 183, 188, 194

- Diagnosis, 25, 49, 58, 67, 236
- Differential-equation Models, 181
- Discrete-event Simulations, 181
- Disease Mapping, 47
- Disease Model, 187, 207
- Disk Striping, 31
- Dose-Response, 43, 48, 52
- Dot Mapping, 48

- Ecologic Studies, 52
- Ecological Fallacy, 52
- Edge Expansion, 190, 191
- Entity Tagging, 80
- Epidemic Outbreak, 163
- Epidemiological Factors, 233, 234
- EpiSims, 28, 183, 187, 188, 207, 212, 230
- Epistasis (Gene-gene Interaction), 67
- Equivalence Relation, 109
- Erdős-Rényi Random Graph, 208, 215
- Ethnicity, 46, 53, 60, 239
- Expectation Maximization, 3
- Expected Outbreak Size, 173
- Extent (of a concept), 125

- F-measure, 80
- Fast Fourier Transform, 30, 31
- Feature Selection, 244, 247, 251
- Force of Infection, 142
- Formal Concept Analysis (FCA), 101
- Formal Context, Core of, 131
- Frequent Closed Itemsets, 114, 117

- Gaussian Distribution, *see also* Normal Distribution
- Generalized Contingency Table, 115
- Generating Function, 173
- Generation Time, 165
- Generators, 117
- Genotype, 63
- Geographic Scale, 48
- Giant Component, 188, 195, 215–217
- GPHIN, 74
- Graph
 - Bipartite, 102, 117, 186, 188, 189, 193, 211
 - Degree Distribution, 136, 164, 188, 190, 196, 208–211
 - Diameter, 129, 171, 195
 - Induced Subgraph, 118
 - Tree, 225
- Graph Partitions, 115
- Graph-connected Binary Relation, 117
- Greatest Lower Bound, 102, 119
- Greedy Algorithm, 206, 227

- Hasse Diagrams, 104
- Hazard Rate Ratio, 140
- Hazard Regression (HARE), 139
- Henle-Koch Postulates, 42

- Heterogeneous Data, 2
- Hierarchical Clustering, 4, 68
- HIV/AIDS, 55, 57, 140
- HTML, 80
- Human Language Technology, 77
- Hypercube, 96
- Hyperplane, 14, 235

- Icerberg Lattice, 106, 135
- Incubation Period, 207
- Independent Probability Space, 217
- Indirect Standardization, 54
- Infectious Period, 207
- Infimum, 102
- Information Retrieval, 78
- Integer Program, 225
- Integrality Gap, 225
- Intent (of a Concept), 102, 119
- International Classification of Diseases (ICD), 44, 45
- Isopleth, 48

- k-center, 4
- k-means Method, 6
- k-median, 4
- Kaplan-Meier Method, 58
- Kernel Trick, 13, 19
- Key Sets, 117

- Language Identification, 79
- Lattice Theory, 93, 137
- Lead Time Bias, 58
- Least Upper Bound, 102, 119
- Length Bias, 58
- Lexicon, 79
- Life Tables, 58, 59, 64
- Linear Classifier, 14, 19, 246
- Linear Program, 227
- Linear Statistical Models, 100
- Linearly Separable Set, 14, 15
- Load Balancing, 30, 31
- Local Exposure, 215–217
- Logarithmic Method, 33
- Logistic Model, 243
- Logistic Regression, 64, 246

- Machine Learning, 25, 234
- Machine Translation, 77
- Mantel correlation, 69, 147
- Maps
 - Choropleth, 48
 - Dot, 49
- Margin, 14, 26, 235, 246
- Maximal Biclique, 119
- MEDLEE, 76
- Mercer's theorem, 19
- Meta-data, 78
- MetaCarta, 81
- Metric Space, 3

- Minimum Spanning Forest, 32
- Missing Data, 45, 46, 236, 237
- MjTAP, 73
- Mitigation, 207
- Model of Vaccine Action, 141
- Model Validation, 243
- Molecular Biology, Central Dogma of, 65
- Molecular Epidemiology, 63
- Monte Carlo simulation, 142
- Mortality Data, 44, 45
- Multi-document Summarization, 79
- Multi-lingual Information, 76
- Multiple Testing, 69

- Natality, 45
- Near Implication, 109, 120
- Neighborhood, 47, 117, 208
- News
 - Browser, 83
 - Group, 78
 - Server, 77
- News Wire, 73
- NNTP, 83
- Normal Distribution, 8, 23, 143, 207
- Normalization, 237, 249, 251
- Novikov's theorem, 14, 15
- NP-Complete, 203, 204, 226

- Odds Ratio (OR), 64, 112
- Open Source Information, 73
- Overlap Ratio, 188, 193

- P-Value, 22, 68
- Patient Record, 73
- Peeled Subcontext, 131
- Peeling, 115
- Perceptron Algorithm, 14, 15
- Perinatal, 45, 46
- Phenotype, 63
- Place Variables, 47, 51
- Poisson, 8
- Polynomial Growth, 177
- Portland, 184, 187
- Posterior Distribution, 22, 68
- Poultry, 73, 244
- Power Law Tail, 164
- Powerset, 97
- Precision, 80, 139
- Preferential Attachment Model, 182, 209
- Prevalence, 54, 98, 141
- Prior Distribution, 22
- Prodromal Period, 207
- ProMED, 75
- Proportionate Mortality Ratio, 56
- Protein, 63

- Quantile Volcano Plots, 70
- Quarantining, 199

- R-Trees, 34
- Race and Ethnicity, 46, 240
- Random Graphs, *see also* Erdős-Rényi
- Randomization, 32
- Rate
 - Absorption, 188
 - Adjusted, 53, 54
 - Case Fatality, 57
 - Crude, 53
 - Incidence, 54–56, 143
 - Shedding, 188
 - Specific, 53
- Recall, 80
- Recursive Feature Elimination (RFE), 247
- Recursive Partitioning, 67
- Relevance Ranking, 81
- Reproductive Number, 165
- Risk Factor, 42, 63, 240, 247
- RODS, 75
- Rooted Level Aware Breadth First Search, 115

- Sampling, 21, 93, 190, 245
- Screening Programs, 58
- Search Engine, 78
- SEER Database, 44, 136, 233
- Semi-External, 31
- Sensitivity, 87, 207, 235, 251, 254
- Sensor Placement, 181, 183
- Shattering, 195
- Sick, Infected Recovering (SIR), 182
- Single nucleotide polymorphism (SNP), 65
- Slack Variables, 18, 19
- Social Contact Networks, 181–184, 222
- Socioeconomic Status, 46
- Specificity, 235, 251
- Speech Transcription, 90
- Standardized Mortality or Morbidity Ratio (SMR), 54
- Stemming, 87
- Stop Word, 87
- Stratified Cross-validation, 246, 247
- Study Design, 51, 93, 140
- Subcontext, 119
- Subgroup Identification, 70
- Summarization, 81
- Support, 10, 11, 98, 119
- Support Push, 107
- Support Vector Machines (SVM), 13, 26, 241, 243
- Supremum, 102
- Surveillance, Epidemiology and End Results, *see also* SEER Database
- Survival
 - Measures, 58
 - Median —, 59
 - Relative —, 59
 - Time, 58, 158, 234

- Symmetrization (of a binary relation), 118
- t-test, 68, 235
- Temporal Degree Distributions, 196–198
- Text Mining, 73
- Time Duration of Disease, 55
- TIRR, 85
- Topped Intersection Structure, 102
- Trajectory, 186
- TRANSIMS, 181
- Transmission Dynamics Model, 140
- Transmission Probability, 168, 215
- Triage Report, 75
- Two-stage Analysis, 68
- Two-way Contingency Table, 95

- Unbiased Probability Space, 215
- Usability, 89

- Vaccination, 140, 182, 225
- Vaccine Efficacy, 139
- Vaccine Trials, 141
- Vapnik-Chervonenkis (VC) dimension, 15
- Venn Diagram, 104
- Veterinary Epidemiology, 244
- Visualization, 90, 105
- Vital Statistics Data, 44, 46

- Wet Litter, 244
- Withdrawal, 208
- Word Segmentation, 88

- Years of Potential Life Lost (YPLL), 57

Studies of the spread and containment of disease rely at heart on a variety of mathematical and computational techniques. This collection aims to introduce the fundamentals of epidemiology and to showcase contemporary work using discrete mathematical techniques. Introductory chapters explain the fundamental concepts of epidemiology, the basic tools provided by mathematics and computer science, and some of the outstanding open problems in the area. Contributed articles then highlight particular problems in monitoring disease outbreaks, vaccination strategies, and modelling disease survival factors, and successfully apply techniques such as formal concept analysis, support vector machines, random graph models, and systems of differential equations.

ISBN 978-0-8218-4379-6



9 780821 843796

DIMACS/70.S

AMS *on the Web*
www.ams.org