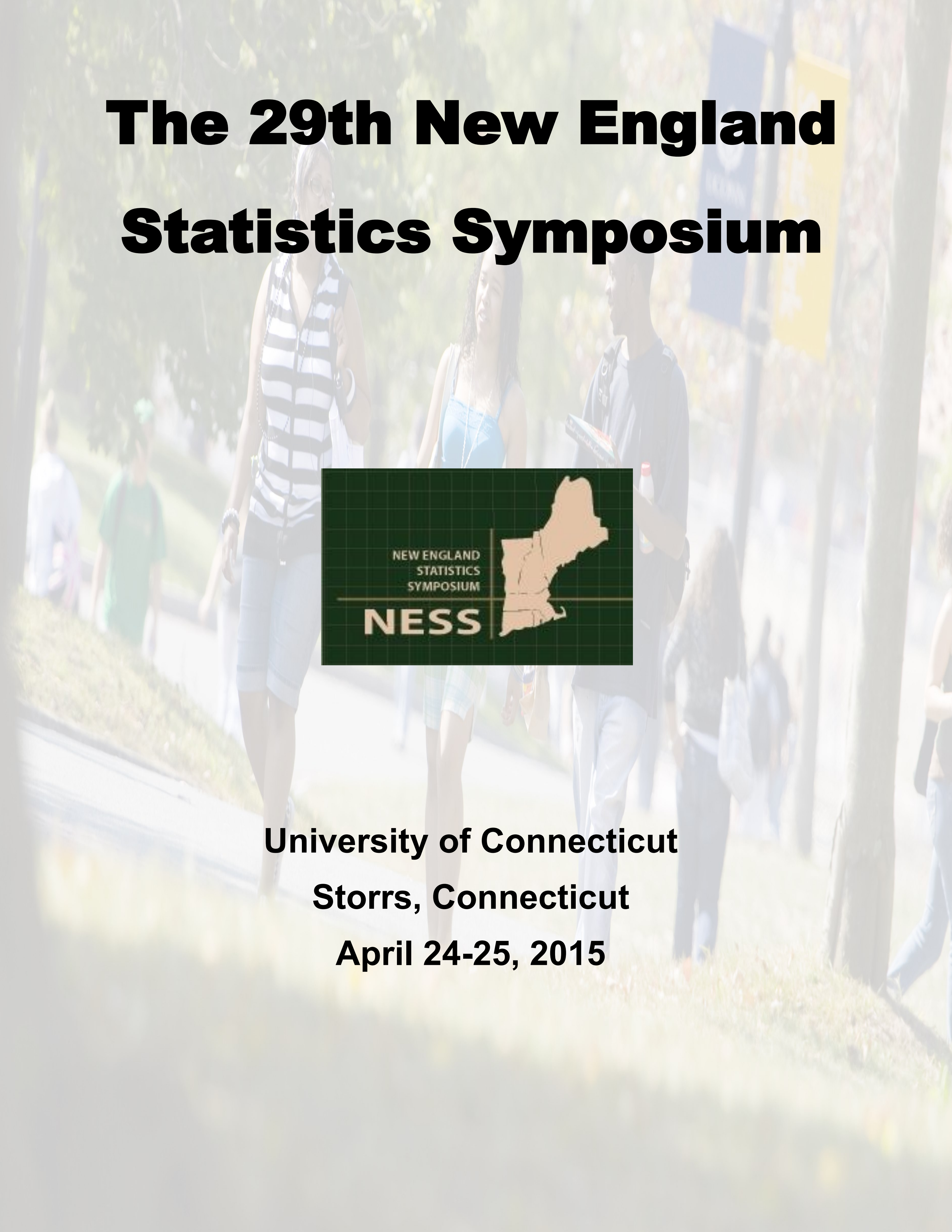

The 29th New England Statistics Symposium

NEW ENGLAND
STATISTICS
SYMPOSIUM
NESS

University of Connecticut

Storrs, Connecticut

April 24-25, 2015

The 2015 New England Statistical Symposium would not be possible without the generous support of our sponsors. Please join us in thanking the following organizations for their contributions to this year's symposium.

TRAVELERS

Boehringer Ingelheim

THE HARTFORD

IBM®

IBM T. J. Watson Research Center

NISS National Institute of Statistical Sciences

ASA AMERICAN STATISTICAL ASSOCIATION

ASA Connecticut Chapter

Google™

Pfizer

University of Connecticut Department of Statistics
University of Connecticut Office of the Vice President for Research
University of Connecticut College of Liberal Arts & Sciences Dean's Office,

NEW ENGLAND STATISTICS SYMPOSIUM NESS

The first New England Statistics Symposium was held at the University of Connecticut. We have been continuing the tradition of hosting the Symposium on alternate years while inviting other institutions throughout New England to host it in years in between.

NESS 2015 Program

Contents

Welcoming Remarks

It is my great pleasure to welcome you to the 29th New England Statistics Symposium. On behalf of the Department of Statistics at the University of Connecticut, I would like to wish all the delegates of the 29th New England Statistics Symposium (NESS) an enjoyable visit to our department and the campus. It is hard to believe that 28 years have passed since we have hosted the 1st NESS here at UConn in 1987, with the enthusiastic support from the academic institutions in New England and our colleague, Professor Herman Chernoff, Harvard University. We started our first New England Statistics Symposium with about 50 participants and now the participation has grown to over 200 from academic institution, industry and government agencies, from all across New England and the US. We are thrilled to have this year two distinguished keynote speakers: UConn Board of Trustees Distinguished Professor and Marianne E. Klewin Professor of Engineering, Professor Yaakov Bar-Shalom, University of Connecticut, and Professor Adrian Raftery, University of Washington, who is a Fellow of the American Academy of Arts and Sciences and member of the United States National Academy of Sciences.

This year we offer in conjunction with NESS an extensive educational program, which includes three one-day short- courses on major areas of current interest in statistical science. These short courses are being offered on Friday, April 24, 8:30am–5pm, at the Student Union, University of Connecticut. Professor Peter Müller, Department of Mathematics and McCombes School of Business, University of Texas at Austin, will offer a course entitled: "Bayesian Biostatistics: Design of Clinical Trials and Subgroup Analysis." Professors Kun Chen and Jun Yan, Department of Statistics, University of Connecticut, will offer a course entitled: "Modern Multivariate Statistical Learning: Methods and Applications." Yihui Xie, Ph.D., RStudio, Inc., will offer a course on "Boosting R Skills and Automating Statistical Reports." I would like to thank the instructors for offering these exciting short courses, which will enrich the knowledge of graduate students and statistical scientists as well.

My sincere thanks go to the leadership of Professors Ming-Hui Chen (Chair), Ofer Harel and Jun Yan for organizing this symposium. I further thank Ms. Megan Petsa, Ms. Tracy Burke and many graduate students who helped at each stage of the planning for this symposium. I am confident that the scientific program for this symposium will be of high quality and lead to the development of exchange of ideas and fruitful interactions between the statistical scientists and graduate students throughout New England and the rest of the country.

I would like to take this opportunity and acknowledge our generous sponsors: Office of the UConn Vice President for Research, College of Liberal Arts & Sciences Dean's Office, UConn Department of Statistics, Connecticut Chapter of the American Statistical Association, Boehringer Ingelheim Pharmaceuticals, Inc., Google, IBM T.J. Watson Research Center, National Institute for Statistical Sciences, Pfizer Inc., The Hartford, and Travelers Insurance Company. I would like to thank Google and the IBM T.J. Watson Research Center, for sponsoring this year the Student Awards for presentations at the 29th NESS. The IBM student awards initiated in 2005 have attracted numerous submissions from graduate students.

I would like to share with you some recent news and accomplishments of our department. Our department was founded in 1963, and is currently one of the major Statistics Departments in New England. Two years ago on November 1–3, 2013, we celebrated the 50th Anniversary of the Department. We were fortunate to have in attendance the founder of our department and its first Department Head, Professor Robert H. Riffenburgh, San Diego State University, to present a plenary lecture: "The Birth of the UConn Statistics Department: From Clay to Sculpture."

The research initiatives and the quality of output of the department continue to soar. We enjoy research funding from a variety of sources including NSF, NIH and private companies. International and national visibility of the department also continues to grow with our faculty's participation and visits at conferences and major universities all over the world. We have developed a strong interdisciplinary research program within UCONN. Recently we have received high ranks in NRC as well as US News and World News. I am proud to mention that 7 of our current faculty, one emeritus professor, are fellows of the ASA, 5 of our current faculty and one emeritus faculty are fellows of the IMS, and 4 of our current faculty are elected members of the Connecticut Academy of Arts and Sciences or Sciences and Engineering. Many of our faculty received intramural and extramural awards in research excellence. Our faculty members continue to hold prestigious editorial and editorial board positions in major journals in probability and statistics.

The department has a great interest in global initiatives promoted by UCONN. The department has co-signed, jointly with CLAS, Graduate School and Provost's Office an MOU with the Department of Mathematics, Shanghai Jiao Tong University, PRC, to collaborate on research projects and establish educational and research exchange programs for faculty and graduate students. We have also signed several additional MOU's, with major universities in China, to collaborate on joint educational programs for senior undergraduate students and graduate students. Our graduate and undergraduate programs continue to expand. Currently we have about 100 graduate students and over 100 undergraduate students majoring in statistics. We have launched this year a new Professional MS degree program in Biostatistics.

I am glad to report that our graduate programs have been ranked recently as 18th in the US by a nationwide poll of graduate students. The following link presents the full survey: `http://www.graduateprograms.com/best-statistics-programs/`.

I am grateful for the support our department has been receiving from the College of Liberal Arts and Sciences and the University. The new faculty positions that have been allocated to our department in the last three years will continue to strengthen research, undergraduate and graduate programs and enhance our department national and international standing.

Welcome to UConn and Storrs, the basketball capital of the world!

With best wishes,
Joseph Glaz, Professor and Head

Keynote Speakers

Adrian E. Raftery
Professor of Statistics and Sociology
University of Washington

Adrian E. Raftery is Professor of Statistics and Sociology at the University of Washington in Seattle. He was born in Dublin, Ireland, and obtained a B.A. in Mathematics (1976) and an M.Sc. in Statistics and Operations Research (1977) at Trinity College Dublin. He obtained a doctorate in mathematical statistics in 1980 from the Université Pierre et Marie Curie in Paris, France, under the supervision of Paul Deheuvels.

He was a lecturer in statistics at Trinity College Dublin from 1980 to 1986, and then an associate (1986–1990) and full (1990–present) professor of statistics and sociology at the University of Washington. He was the founding Director of the Center for Statistics and Social Sciences (1999–2009).

Professor Raftery has published over 170 refereed articles in statistical, sociological and other journals. His research focuses on Bayesian model selection and Bayesian model averaging, model-based clustering, inference for deterministic simulation models, and the development of new statistical methods for sociology, demography, and the environmental and health sciences.

He is a member of the United States National Academy of Sciences, a Fellow of the American Academy of Arts and Sciences, an Honorary Member of the Royal Irish Academy, a member of the Washington State Academy of Sciences, a Fellow of the American Statistical Association, a Fellow of the Institute of Mathematical Statistics, and an elected Member of the Sociological Research Association. He has won the Population Association of America's Clifford C. Clogg Award, the American Sociological Association's Paul F. Lazarsfeld Award for Distinguished Contribution to Knowledge, the Jerome Sacks Award for Outstanding Cross-Disciplinary Research from the National Institute of Statistical Sciences, and the Parzen Prize for Statistical Innovation. He is also a former Coordinating and Applications Editor of the Journal of the American Statistical Association and a former Editor of Sociological Methodology.

He was identified as the world's most cited researcher in mathematics for the decade 1995–2005 by Thomson-ISI. Twenty-six students have obtained Ph.D.'s working under Raftery's supervision, of whom 18 hold university faculty positions.

Keynote Lecture: Probabilistic Population Projections for All Countries

Projections of countries' future populations, broken down by age and sex, are widely used for planning and research. They are mostly done deterministically, but there is a widespread need for probabilistic projections. I will describe a Bayesian statistical method for probabilistic population projections for all countries. These new methods have been used by the United Nations to produce their most recent population projections for all countries.

Yaakov Bar-Shalom
Professor of Engineering
University of Connecticut

Yaakov Bar-Shalom was born on May 11, 1941. He received the B.S. and M.S. degrees from the Technion, Israel Institute of Technology, in 1963 and 1967 and the Ph.D. degree from Princeton University in 1970, all in electrical engineering. From 1970 to 1976 he was with Systems Control, Inc., Palo Alto, California. Currently he is Board of Trustees Distinguished Professor in the Dept. of Electrical and Computer Engineering and Marianne E. Klewin Professor in Engineering at the University of Connecticut. He is also Director of the ESP (Estimation and Signal Processing) Lab.

His current research interests are in estimation theory, target tracking and data fusion. He has published over 500 papers and book chapters in these areas and in stochastic adaptive control. He coauthored the monograph Tracking and Data Association (Academic Press, 1988), the graduate texts Estimation and Tracking: Principles, Techniques and Software (Artech House, 1993; translated into Russian, MGTU Bauman, Moscow, Russia, 2011), Estimation with Applications to Tracking and Navigation: Algorithms and Software for Information Extraction (Wiley, 2001), the advanced graduate texts MultitargetMultisensor Tracking: Principles and Techniques (YBS Publishing, 1995), Tracking and Data Fusion (YBS Publishing, 2011), and edited the books MultitargetMultisensor Tracking: Applications and Advances (Artech House, Vol. I, 1990; Vol. II, 1992; Vol. III, 2000).

He has been elected Fellow of IEEE for "contributions to the theory of stochastic systems and of multi-target tracking". He has been consulting to numerous companies and government agencies, and originated the series of MultitargetMultisensor Tracking short courses offered via UCLA Extension, at Government Laboratories, private companies and overseas.

During 1976 and 1977 he served as Associate Editor of the IEEE Transactions on Automatic Control and from 1978 to 1981 as Associate Editor of Automatica. He was Program Chairman of the 1982 American Control Conference, General Chairman of the 1985 ACC, and CoChairman of the 1989 IEEE International Conference on Control and Applications. During 198387 he served as Chairman of the Conference Activities Board of the IEEE Control Systems Society and during 198789 was a member of the Board of Governors of the IEEE CSS. He was a member of the Board of Directors of the International Society of Information Fusion (1999–2004) and served as General Chairman of FUSION 2000, President of ISIF in 2000 and 2002 and Vice President for Publications in 2004-2013.

In 1987 he received the IEEE CSS Distinguished Member Award. Since 1995, he is a Distinguished Lecturer of the IEEE AESS and has given numerous keynote addresses at major national and international conferences. He is corecipient of the M. Barry Carlton Award for the best paper in the IEEE Transactions on Aerospace and Electronic Systems in 1995 and 2000 and recipient of the 1998 University of Connecticut AAUP Excellence Award for Research. In 2002 he received the J. Mignona Data Fusion Award from the DoD JDL Data Fusion Group. He is a member of the Connecticut Academy of Science and Engineering.

In 2008 he was awarded the IEEE Dennis J. Picard Medal for Radar Technologies and Applications, and in 2012 the Connecticut Medal of Technology. He has been listed by academic.research.microsoft (top authors in engineering) as #1 among the researchers in Aerospace Engineering based on the citations of his work.

Keynote Lecture: How to Get the Most Out of Your Sensors (and make a living out of this)

This talk discusses the issues related to information extraction and data fusion from multiple sensors. The goal of extracting the maximum possible amount of information from each sensor requires the use of appropriate sensor and target models. In these models one has to quantify the corresponding uncertainties. The issues related to data association and multiple target behavior models are discussed together with some practical algorihms and their implementations for Low Observable targets. The fusion of the information from various sources has to account for their uncertainties as well as the interrelationship—crosscorrelations—between the uncertainties across sources. The "Track-to-Track Fusion" and "Centralized Fusion" configurations are discussed.

Schedule

Friday, April 24, 2015

8:30am–5:00pm **NESS short courses**

Saturday, April 25, 2015

8:30am **Registration & Refreshments**

8:30am **Poster Session**, AUST 313 and 344, and the third floor hallway

9:30am **Welcome and Opening Remarks**, AUST 108
Ming-Hui Chen
Chair of the 29th NESS Planning Committee

Joseph Glaz
Professor and Department Head
University of Connecticut

Jeremy Teitelbaum
Dean of CLAS and Professor of Mathematics
University of Connecticut

Kent Holsinger
Vice Provost for Graduate Education and Dean of the Graduate School
University of Connecticut

9:45am–10:45am **Keynote Presentation**, AUST 108

Probabilistic Population Projections for All Countries
Adrian E. Raftery
University of Washington

Introduction by Dipak K. Dey, University of Connecticut

11:00am–12:45pm **Invited Paper and Poster Oral Sessions**

Session 1: Frontiers in Sequential Analysis with Applications
AUST 103

Session 2: Statistical Challenges in Modeling and Applications
AUST 202

Session 3: Statistical Innovations in Biomedical Research
AUST 102

Session 4: Recent Advances in Spatial Statistics
AUST 445

Session 5: Statistical Inference in Time Series and Machine Learning
AUST 163

Session 6: Recent Developments in Analyzing Survival Endpoint Methods After Taking Alternative Therapy/Treatment Switching in Oncology Trials
AUST 434

Session 7: Boehringer Ingelheim and Travelers Sponsered Poster Session I
AUST 105

Session 8: Boehringer Ingelheim and Travelers Sponsered Poster Session II
AUST 108

Session 9: Boehringer Ingelheim and Travelers Sponsered Poster Session III
AUST 110

12:45pm–2:00pm **Lunch**, Wilbur Cross Reading Room North

1:00pm–2:10pm **Poster Session (continued)**, AUST 313 and 344, and the third floor hallway

2:15pm–3:15pm **Keynote Presentation**, AUST 108

How to Get the Most Out of Your Sensors (and make a living out of this)
Yaakov Bar-Shalom
University of Connecticut

Introduction by Joseph Glaz, University of Connecticut

3:15pm–3:30pm **Coffee Break**

3:30pm–5:15pm **Invited Paper Sessions**

Session 10: Statistical Methods and Computing with Big Data
AUST 102

Session 11: Statistical Applications and Practice
AUST 105

Session 12: Recent Advances in Subgroup Analyses
AUST 103

Session 13: Pharmaceutical Applications
AUST 108

Session 14: Design and Analysis of Complex Experiments
AUST 202

Session 15: Association and Correlation Analysis for Big Data
AUST 110

Session 16: Career Development
AUST 445

Session 17: Advances in Molecular Evolution and Statistics Genetics
AUST 164

Session 18: Data Analytics at IBM Research
AUST 163

Session 19: Probability and Related Topics – in memory of Evarist Giné
AUST 434

5:15pm–6:00pm **Student Paper and Poster Awards, Closing Reception**
Noether Lounge, AUST 326

Dipak K. Dey
Associate Dean of CLAS, University of Connecticut

Yasuo Amemiya
IBM T. J. Watson Research Center

Xiaojing Wang
Google, Inc.

6:30pm **Dinner at Sichuan Pepper restaurant**
435G Hartford Turnpike, Vernon, CT 06066

Detailed Program

Session 1: Frontiers in Sequential Analysis with Applications — Invited
Time and Location: 11:00am–12:45pm in AUST 103
Organizer: Aleksey Polunchenko, *State University of New York at Binghamton*
Chair: Grigory Sokolov, *State University of New York at Binghamton*

Talks (abstracts on page 22):

1. Quickest Detection with Post-Change Distribution Uncertainty
   Heng Yang, *City University of New York, Graduate Center*

2. An Analytic Expression for the Distribution of the Generalized Shiryaev-Roberts Diffusion
   Aleksey Polunchenko, *State University of New York at Binghamton*

3. Detecting Changes in Complex Networks: Challenges and New Directions
   Vasanthan Raghavan, *Qualcomm Flarion Technologies, Inc., Bridgewater, NJ*

4. Multisensor Quickest Detection
   Grigory Sokolov, *State University of New York at Binghamton*

5. Recent Development of First Crossing Times of Compound Poisson Processes with Two Piecewise Linear Boundaries
   Yifan Xu, *Case Western Reserve University*

Session 2: Statistical Challenges in Modeling and Applications — Invited
Time and Location: 11:00am–12:45pm in AUST 202
Organizer and Chair: Erin Conlon, *University of Massachusetts - Amherst*

Talks (abstracts on page 24):

1. On Missing Data Mechanism in Two-Way Incomplete Contingency Tables
   Daeyoung Kim, *University of Massachusetts - Amherst*

2. Quantile Regression for Survival Data with Delayed Entry
   Jing Qian, *University of Massachusetts - Amherst*

3. Estimating Population Susceptibility in Dynamic Models of Infectious Disease
   Nicholas Reich, *University of Massachusetts - Amherst*

4. Variable Selection in Single-index Varying Coefficient Models
   Anna Liu, *University of Massachusetts - Amherst*

5. Semi-parametric Time to Event Models in the Presence of Error-prone, Self-reported Outcomes - with Application to the Womens Health Initiative
   Raji Balasubramanian, *University of Massachusetts - Amherst*

Session 3: Statistical Innovations in Biomedical Research — Invited
Time and Location: 11:00am–12:45pm in AUST 102
Organizers: Yuping Zhang, *University of Connecticut* and Hongyu Zhao, *Yale University*
Chair: Yuping Zhang, *University of Connecticut*

Talks (abstracts on page 25):

1. A Nonparametric Approach to Comparing Diagnostic Accuracies in a Multi-Reader, Multi-Test Design
   Eunhee Kim, *Brown University*

2. Effective Detection of Differences in Related Mixture Distributions
   Li Ma, *Duke University*

3. LD Score: An Approach to Interpret the Polygenic Contribution of Common Variation
   Benjamin Neale, *Massachusetts General Hospital*

4. Analysis of Gene Expression at the Single-cell Resolution
   Guo-Cheng Yuan, *Dana-Farber Cancer Institute*

Session 4: Recent Advances in Spatial Statistics — Invited
Time and Location: 11:00am–12:45pm in AUST 445
Organizer and Chair: Cici Bauer, *Brown University*

Talks (abstracts on page 26):

1. Compact, Disjoint Bases for Spatio-Temporal Point Processes
   Luke Bornn, *Harvard University*

2. Bayesian Spatial Modeling of the Local Persistence of PCV-Targeted Pneumococcal Serotypes among Adults in Connecticut, 1998-2009
   Joshua Warren, *Yale University*

3. Bayesian Spatial Hierarchical Models for Small Estimation with Complex Survey Designs
   Cici Bauer, *Brown University*

4. Disease Risk Estimation by Combining Spatial Case-Control Data with Aggregated Information on the Population at Risk
   Xiaohui Chang, *Oregon State University*

5. Bayesian Spatio-Temporal Point Level Modelling of Air Pollution Concentration Levels for Estimating Long Term Exposure in Coarser Administrative Geographies in the UK
   Sujit K. Sahu, *University of Southampton, UK*

Session 5: Statistical Inference in Time Series and Machine Learning — Invited
Time and Location: 11:00am–12:45pm in AUST 163
Organizer and Chair: Luis Carvalho, *Boston University*

Talks (abstracts on page 28):

1. Statistical Inference for Perturbations of Multiscale Dynamical Systems
   Siragan Gailus, *Boston University*

2. Parameter Estimation for Continuous-time Stationary Models with Memory
   Mamikon Ginovyan, *Boston University*

3. Feature Vector Denoising with Prior Network Structures
   Mark A. Kon, *Boston University*

4. Semiparametric Model Building for Regression Models with Time-Varying Parameters
   Ting Zhang, *Boston University*

Session 6: Recent Developments in Analyzing Survival Endpoint Methods After Taking Alternative Therapy/Treatment Switching in Oncology Trials — Invited
Sponsor: ICSA New England Chapter
Time and Location: 11:00am–12:45pm in AUST 434
Organizer: Jing Xu, *Takeda Pharmaceuticals, Inc.* and Huyuan Yang, *Takeda Pharmaceuticals, Inc.*
Chair: Huyuan Yang, *Takeda Pharmaceuticals, Inc.*

Talks (abstracts on page 30):

1. Analyzing Time to Event Data in the Presence of Informative Censoring - a Regulatory Submission Case Study
   Jing Xu, *Takeda Pharmaceuticals, Inc.*

2. The g-formula to Adjust for Noncompliance in Randomized Clinical Trials
   Jessica Young, *Harvard T.H. Chan School of Public Health*

3. Comparing the Performance of Several Re-Censoring Rules in the RPSFT Method for the Analysis of OS Data Adjusting for Treatment Switching in the Phase III Oncology Trials
   Huyuan Yang, *Takeda Pharmaceuticals, Inc.*

Session 10: Statistical Methods and Computing with Big Data — Invited
Time and Location: 3:30pm–5:15pm in AUST 102
Organizers: John Emerson, *Yale University* and Jun Yan, *University of Connecticut*
Chair: John Emerson, *Yale University*

Talks (abstracts on page 31):

1. Computational Tools for Bayesian Methods in Big Data and Data Science
   Erin Conlon, *University of Massachusetts - Amherst*

2. High Performance Data I/O
   Taylor Arnold, *AT&T Labs Research*

3. Scalable, Exact Approaches to Fitting Linear Models when $n \gg p$
   Michael Kane, *Yale University*

4. Online Updating of Statistical Inference in the Big Data Setting
   Elizabeth Schifano, *University of Connecticut*

5. Further Growth of the Bigmemory Family of Packages
   John Emerson, *Yale University*

Session 11: Statistical Applications and Practice — Invited
Time and Location: 3:30pm–5:15pm in AUST 105
Organizers: Xiaojing Wang, *Google, Inc.* and Jun Yan, *University of Connecticut*
Chair: Jun Yan, *University of Connecticut*

Talks (abstracts on page 32):

1. A Metamodeling Approach to the Valuation of Large Variable Annuity Portfolio under Nested Simulation
   Guojun Gan, *University of Connecticut*

2. Measuring Online Audiences Using Logs Data
   Xiaojing Wang, *Google, Inc.*

3. The Optimal Mix of TV and Online Ads
   Georg Goerg, *Google, Inc.*

4. Statistical Applications in Property and Casualty Insurance
   James Landgrebe, *Travelers Insurance*

Session 12: Recent Advances in Subgroup Analyses — Invited
Time and Location: 3:30pm–5:15pm in AUST 103
Organizer: Xiaojing Wang, *University of Connecticut*
Chair: Yuefeng Lu, *Sanofi*

Talks (abstracts on page 33):

1. Subgroup Reporting Using Nonparametric Bayesian Inference
   Peter Müller, *University of Texas at Austin*

2. Bayesian Approaches to Subgroup Analysis and Selection Problems in Drug Development
   David Ohlssen, *Novartis*

3. Effective Subgroup Identification by Systematically Utilizing Multiple Baseline Characteristics
   Lihui Zhao, *Northwestern University*

4. A Bayesian Approach for Subgroup Analysis
   Xiaojing Wang, *University of Connecticut*

5. Biomarker Stratified Adaptive Basket Designs for Multiple Cancers
   Lorenzo Trippa, *Dana-Farber Cancer Institute*

Session 13: Pharmaceutical Applications — Invited
Sponsor: ASA CT Chapter
Time and Location: 3:30pm–5:15pm in AUST 108
Organizer: Daniel Meyer, *Pfizer, Inc.*
Chair: Joeseph Cappelleri, *Pfizer, Inc.*

Talks (abstracts on page 35):

1. Questionnaire on Network Meta-Analysis to Assess Its Relevance and Credibility
   Joseph Cappelleri, *Pfizer, Inc.*

2. An Introduction to Multiple Testing Procedures with an Example Using a Graphical Approach
   Nate Bennett, *Ingelheim Pharmaceuticals, Inc.*

3. Characterizing Disease Progression with Multivariate Longitudinal Models: An Example in Alzheimer's Disease
   James Rogers, *Metrum Research Group*

4. Comparison of Treatment Response and Loss of Response Definitions in Enriched Enrollment Randomized Withdrawal Design Pain Studies
   Birol Emir, *Pfizer, Inc.*

Session 14: Design and Analysis of Complex Experiments — Invited
Time and Location: 3:30pm–5:15pm in AUST 202
Organizer and Chair: Edoardo Airoldi, *Harvard University*

Talks (abstracts on page 36):

1. Analysis of Cluster Encouragement Designs with Interference
   Fabrizia Mealli, *University of Florence and Harvard*

2. A Potential Outcomes Based Perspective of the Analysis of Complex Multi-Factor Experiments with Randomization Restrictions
   Tirthankar Dasgupta, *Haravrd University*

3. Causal Inference with Partially Revealed Interference
   Panos Toulis, *Harvard University*

Session 15: Association and Correlation Analysis for Big Data — Invited
Time and Location: 3:30pm–5:15pm in AUST 110
Organizers: Kun Chen, *University of Connecticut* and Jian Zou, *Worcester Polytechnic Institute*
Chair: Jian Zou, *Worcester Polytechnic Institute*

Talks (abstracts on page 37):

1. Inferring Anchor Links across Heterogeneous Social Networks.
   Xiangnan Kong, *Worcester Polytechnic Institute*

2. Covariance Matrix Estimation in Big Data: Approaches Based on Algebraic Properties
   Xi Luo, *Brown University*

3. A Bayesian Test of Independence for Each of Several Sparse Two-Way Contingency Tables
   Balgobin Nandram, *Worcester Polytechnic Institute*

4. Robust Principal Component Analysis for Detecting Sparsely Correlated Phenomena in Computer Networks
   Randy Paffenroth, *Worcester Polytechnic Institute*

5. High-Frequency Financial Risk Management and High Performance Computing
   Jian Zou, *Worcester Polytechnic Institute*

Session 16: Career Development — Invited
Time and Location: 3:30pm–5:15pm in AUST 445
Organizer and Moderator: Naitee Ting, *Boehringer Ingelheim Pharmaceuticals, Inc.*

In this session, each panelist will speak for about 5 to 10 min about how to prepare CV, how to handle phone screen, and interview. After that, the session is open for Q and A.

Panelists:

1. Yasuo Amemiya, *IBM Thomas J. Watson Research Center*

2. Lane G. Coonrod, *The Hartford*

3. Shuangge Steven Ma, *Yale University*

4. Daniel Meyer, *Pfizer, Inc.*

5. Christopher Parks, *Travelers*

6. Torey Strauser, *Valesta Clinical Research Solutions*

Session 17: Advances in Molecular Evolution and Statistics Genetics — Invited
Time and Location: 3:30pm–5:15pm in AUST 164
Organizers: Lynn Kuo, *University of Connecticut* and Zheyang Wu, *Worcester Polytechnic Institute*
Chair: Lynn Kuo, *University of Connecticut*

Talks (abstracts on page 39):

1. Mutation and the Branching Structure of Evolutionary Trees
   Forrest W. Crawford, *Yale University*

2. Estimation of Phylogenetic Power for All Symmetric Markov Models of Evolution and Any Number of Taxa
   Jeffrey Townsend, *Yale University*

3. Confidence Interval Estimation Using Approximate Likelihood of Divergence Time for a Coalescent-based Model
   Arindam RoyChoudhury, *Columbia University*

4. Detecting Tracts of Local Ancestry for Admixture Disease Mapping
   Matthias Steinrücken, *University of Massachusetts - Amherst*

5. Assessing the Probability That a Finding Is Genuine for Large-Scale Genetic Association Studies
   Chia-Ling Kuo, *UConn Health Center*

6. Approximating the Distributions of Optimal Goodness-of-Fit Tests with Applications in GWAS
   Zheyang Wu, *Worcester Polytechnic Institute*

Session 18: Data Analytics at IBM Research — Invited
Sponsor: IBM T. J. Watson Research Center
Time and Location: 3:30pm–5:15pm in AUST 163
Organizer: Yasuo Amemiya, *IBM T. J. Watson Research Center*
Chair: Beatriz E Etchegaray Garcia, *IBM T. J. Watson Research Center*

Talks (abstracts on page 41):

1. Statistical Challenges for a Crowd-Sourcing-Based Delivery Platform of Software Development
   Ta-Hsin Li, *IBM T. J. Watson Research Center*

2. Robust Compressed Least Squares
   Ban Kawas, *IBM T. J. Watson Research Center*

3. Co-Clustering Structural Temporal Data with Applications to Semiconductor Manufacturing
   Yada Zhu, *IBM T. J. Watson Research Center*

4. Process Monitoring Techniques Based on Likelihood Ratios
   Emmanuel Yashchin, *IBM T. J. Watson Research Center*

5. Predicting Demand for Optimal Opportunity Pipeline Management
   Beatriz E Etchegaray Garcia, *IBM T. J. Watson Research Center*

Session 19: Probability and Related Topics – in memory of Evarist Giné — Invited
Time and Location: 3:30pm–5:15pm in AUST 434
Organizer and Chair: Rick Vitale, *University of Connecticut*

Talks (abstracts on page 43):

1. Welcome
   Rick Vitale, *University of Connecticut*

2. Evarist as a Student, Teacher and Friend
   Richard M. Dudley, *Massachusetts Institute of Technology*

3. Dependence Measures: A Perspective
   Victor de la Pena, *Columbia University*

4. Evarist's Favorite Undergraduate Proof and Where It Got Me
   Iddo Ben-Ari, *University of Connecticut*

5. On the Sup-norm Behavior of the Bernstein Density Estimator
   Lu Lu, *Colby College*

6. Evarist - Reminiscences
   Molly Hahn, *Tufts University*, and others as they would like

Session 7: Boehringer Ingelheim and Travelers Sponsored Poster Session I
Time and Location: 11:00am–12:45pm in AUST 105
Moderator: Jun Yan, *University of Connecticut*

Posters (abstracts on page 44):

1. Pathoscope 2.0: Statistical and Computational Methods for Accurate Characterization of Microbes in Sequencing Samples
   Solaiappan Manimaran, *Boston University*
2. In-Stream Text Mining of Patient Narratives: Sweeping for Missed Endpoints
   Greg Cicconetti, *GlaxoSmithKline*
3. Analysis of Measurement Errors in Evaluating the Effectiveness of Graduated Driver's Licensing Program
   Yang Liu, *University of Connecticut*
4. Statistical Power as a Function of Internal Consistency of Instrument Questionnaire Items
   Moonseong Heo, *Albert Einstein College of Medicine*
5. Student Growth Percentiles in the Presence of Measurement Error
   Eugene Quinn, *Stonehill College*
6. Analyzing Single-Molecule Protein Transportation Experiments via Hierarchical Hidden Markov Models
   Yang Chen, *Harvard University*
7. ISAP-MATLAB Package for Sensitivity Analysis of High-Dimensional Stochastic Chemical Networks
   Weilong Hu, *University of Massachusetts - Amherst*
8. Order Restricted Inference in Regression
   Thelge (Buddika) Peiris, *Worcester Polytechnic Institute*
9. A Unified Theory of Testing and Confidence Regions for High Dimensional Estimating Equations
   Matey Neykov, *Harvard University*
10. Sparse Kernel Machine Regression for Ordinal Outcomes
    Yuanyuan Shen, *Harvard University*
11. Canonical Variate Regression
    Chongliang Luo, *Department of Statistics, University of Connecticut*
12. Penalized variable selection in competing risks regression
    Zhixuan Fu, *Yale University*
13. Information-Theoretic Characterization of Short-Memory and Long-Memory Gaussian Processes
    Gordon Chavez, *New York University*
14. Power Evaluation of Methods for Detecting Non-Independent Relationships
    Ruobin Gong, *Harvard University*
15. The Optimised Theta Method
    Jose Augusto Fioruci, *Federal University of Sao Carlos / University of Connecticut*
16. Negative-Binomial Cure Rate Models with Spatial Frailties for Interval-Censored Data
    Yiqi Bao, *Federal University of Sao Carlos / University of Connecticut*
17. Hierarchical Bayesian Models, Small Area Estimation and Dirichlet Processes
    Jiani Yin, *Worcester Polytechnic Institute*

18. Stable and Optimal: Implicit Stochastic Gradient Descent with Averaging
Dustin Tran, *Harvard University*

19. Cluster Forests
Donghui Yan, *University of Massachusetts - Dartmouth*

Session 8: Boehringer Ingelheim and Travelers Sponsored Poster Session II
Time and Location:: 11:00am–12:45pm in AUST 108
Moderator: Xiaojing Wang, *University of Connecticut*

Posters (abstracts on page 51):

1. Consistent Estimation for Multi-Graph Stochastic Block Models, with Applications to Dynamic And Multi-Layer Networks
Qiuyi Han, *Harvard University*

2. Implicit Stochastic Gradient Descent for Principled Estimation with Large Datasets
Panos Toulis, *Harvard University*

3. Estimation of Discrete Survival Function through the Modeling of Diagnostic Accuracy for Mismeasured Outcome Data
Hee-Koung Joeng, *University of Connecticut*

4. Causal Inference with Social Interference
Edward Kao, *Harvard University*

5. Bayesian Or's of And's for Interpretable Classification, with Application to Context-Aware Recommender Systems
Tong Wang, *Massachusetts Institute of Technology*

6. Marginal Likelihood Estimation of Variable Tree Topology in Phylogenetics
Daoyuan Shi, *University of Connecticut*

7. Group-Corrected Stochastic Blockmodels for Community Detection on Large-scale Networks
Lijun Peng, *Boston University*

8. The Impact of Missing Values on Different Measures of Uncertainty
Chantal Larose, *University of Connecticut*

9. The Application of Sparse Estimation of Covariance Matrix to Quadratic Discriminant Analysis
Jiehuan Sun, *Yale University*

10. Post-GWAS Prioritization Through Integrated Analysis of Genomic Functional Annotation
Qiongshi Lu, *Yale University*

11. Bayesian Forest Classifier: Building Tree Structures on Naive Bayes
Viktoriya Krakovna, *Harvard University*

12. Placebo Non-Response Measure in Sequential Parallel Comparison Design Studies
Denis Rybin, *Boston University*

13. Multivariate Temporal Dynamics of Gastropod Abundance in a Puerto Rican Tropical Forest
Volodymyr Serhiyenko, *University of Connecticut*

14. Onset Time of Chronic Pseudomonas Aeruginosa Infection in Two Cohorts of Cystic Fibrosis Patients with Interval Censored Data
Wenjie Wang, *University of Connecticut*

15. Hidden Population Size Estimation from Respondent-Driven Sampling: A Network Approach
Jiacheng Wu, *Yale University*

16. Selecting the Number of Largest Order Statistics in Extreme Value Analysis
Brian Bader, *University of Connecticut*

17. Prediction of RiboSNitches
Jianan Lin, *University of Connecticut*

18. Second Order Correctness of Perturbation Bootstrap M Estimator of Multiple Linear Regression Parameter
Debraj Das, *North Carolina State University*

19. An Alarm System for Flu Outbreaks Using Google Flu Trend Data
Gregory Vaughan, *University of Connecticut*

20. Achieving Optimal Misclassification Proportion in Stochastic Block Model
Anderson Y. Zhang, *Yale University*

Session 9: Boehringer Ingelheim and Travelers Sponsored Poster Session III
Time and Location:: 11:00am–12:45pm in AUST 110
Moderator: Ofer Harel, *University of Connecticut*

Posters (abstracts on page 59):

1. Assessing Covariate Effects with the Monotone Partial Likelihood Using Jeffreys' Prior in the Cox Model
Jing Wu, *University of Connecticut*

2. Rate-Optimal Graphon Estimation
Yu Lu, *Yale University*

3. The Informative g-Prior vs. Common Reference Priors for Binomial Regression in an Application to Hurricane Electrical Utility Asset Damage Prediction
Nathan Lally, *University of Connecticut*

4. Statistical Analysis of Gene-expression Networks
Seo-Jin Bang, *University of Connecticut*

5. Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models
Justin Yang, *Harvard University*

6. Falling Rule Lists
Fulton Wang, *Massachusetts Institute of Technology*

7. Bayesian Analysis of Joint Modeling of Response Times with Dynamic Latent Ability
Abhisek Saha, *University of Connecticut*

8. Fully-Specified Subdistribution Model Using Weibull Distribution
Fatemeh Sadat Hosseini-Baharanchi, *Tarbiat Modares University*

9. Application of Diagnostics for Respondent-Driven Sampling
Dongah Kim, *University of Massachusetts - Amherst*

10. A Statistical Analysis of Information Spread Across Social Networks
Michael Piserchia, *University of Rhode Island*

11. Predictive Network Modeling of the 2014-2015 Ebola Epidemic in Sierra Leone
Daven Amin, *University of Rhode Island*

12. A Terminal Trend Model for Longitudinal Medical Cost Data and Survival
Qian Yang, *Dartmouth College*

13. Network Analysis Applied to Stock Market Data
    Gregory Breard, *University of Rhode Island*

14. Methods of Adjusting for Misclassification in Respondent-Driven Sampling Data
    Isabelle Beaudry, *University of Massachusetts - Amherst*

15. A New Monte Carlo Method for Computing Marginal Likelihoods
    Yu-Bo Wang, *University of Connecticut*

16. Statistical Modeling of High-Throughput RNase Footprinting for Genome-Wide RNA Structure Inference
    Chenchen Zou, *The Jackson Lab*

17. Induction and Priors for Finite Populations
    Sudip Bose, *The George Washington University*

18. What Can We Tell about a Consumer's Political Affiliation based on Co-Purchases on Amazon?
    Gabriel De Pace, *University of Rhode Island*

19. Analysis and Comparison of Air Pollutants in Wuhan, China
    Zihao Zhang, *Brown University*

NESS 2015 Committees

UCONN NESS 2015 Planning Committee

Ming-Hui Chen (Chair)

Ofer Harel

Jun Yan

M. Henry Linder (Symposium website, IT and Program)

Megan Petsa (Department of Statistics Program Assistant)

Tracy Burke (Department of Statistics Secretary)

NESS 2015 IBM Student Paper Award Committee

Haim Bar, *University of Connecticut*

Kun Chen, *University of Connecticut*

Ming-Hui Chen (Coordinator), *University of Connecticut*

Zhiyi Chi, *University of Connecticut*

Beatriz E Etchegaray Garcia, *IBM T. J. Watson Research Center*

Lynn Kuo, *University of Connecticut*

Elizabeth D. Schifano, *University of Connecticut*

Xiaojing Wang, *University of Connecticut*

Jun Yan, *University of Connecticut*

NESS 2015 Google Student Poster Award Committee

Abidemi Adeniji, *Boehringer Ingelheim Pharmaceuticals, Inc.*

Kun Chen, *University of Connecticut*

Ming-Hui Chen (Coordinator), *University of Connecticut*

Zhiyi Chi, *University of Connecticut*

Beatriz E Etchegaray Garcia, *IBM T. J. Watson Research Center*

Ofer Harel, *University of Connecticut*

Lynn Kuo, *University of Connecticut*

Peter Müller, *University of Texas at Austin*

Nalini Ravishanker, *University of Connecticut*

Elizabeth D. Schifano, *University of Connecticut*

Jeffrey Townsend, *Yale University*

Xiaojing Wang, *University of Connecticut*

Jun Yan, *University of Connecticut*

UCONN NESS 2015 Student Committees

Dooti Roy (Chair)

Gregory Vaughan (Head, Registration Committee)

Hee-Koung Joeng, Chantal Larose

Hao Li, Dan Liu

Rouchen Zha, Fan Zhang

Ved Deshpande (Head, Poster Committee)

Abhishek Bishoyi, Anthony Labarga

Jin Liu, Volodymyr Serhiyenko

Daoyuan Shi

Brian Bader (Head, Parking Committee)

Yeongjin Gwon, Hongbing Jin

Aditya Mishra, Shariq Mohammad

Abhisek Saha, Bo Zhao

Chun Wang (Head, IT Committee)

Gyuhyeong Goh, Jun Hu

M. Henry Linder, Chongliang Luo

Qian Meng, Ellis Shaffer

Ellis Shaffer, Zhe Sun

Yu-Bo Wang, Qianzhu Wu

Yaohua Zhang

Bo Zhao (Head, Sign Committee)

Seo-jin Bang

Paul Mclaughlin

Abhisek Saha

Jing Wu (Head, Photography Committee)

Sudeep Bapat

Abstracts of Invited Papers

Session 1: Frontiers in Sequential Analysis with Applications

Quickest Detection with Post-Change Distribution Uncertainty

Heng Yang, *City University of New York, Graduate Center*
Olympia Hadjiliadis, *City University of New York, Brooklyn College*
Michael Ludkovski, *University of California Santa Barbara*

We consider the problem of quickest detection of an abrupt change when there is uncertainty about the post-change distribution. In particular we examine this problem in the continuous-time Wiener model where the drift of observations changes from zero to a drift randomly chosen from a collection. We set up the problem as a stochastic optimization in which the objective is to minimize a measure of detection delay subject to a frequency of false alarm constraint. We consider a composite rule involving the CUSUM reaction period, that is the time between the last reset of the CUSUM statistic process and the CUSUM alarm, and show that by choosing parameters appropriately, such a composite rule can be asymptotically optimal of third order in detecting the change point as the average time to the first false alarm increases without bound. Such a composite rule can also provide the information of the post-change distribution through the idea of its construction based on the uncertainty.

An Analytic Expression for the Distribution of the Generalized Shiryaev-Roberts Diffusion

Aleksey Polunchenko, *State University of New York at Binghamton*
Grigory Sokolov, *State University of New York at Binghamton*

We consider the quickest change-point detection problem where the aim is to detect the onset of a specified drift in "live"-monitored standard Brownian motion with the change-point assumed unknown (nonrandom). Given this minimax context, the topic of interest is the distribution of the Generalized Shryaev-Roberts (GSR) detection statistic set up to "sense" the presence of the drift. Specifically, we derive a closed-form formula for the transition probability density function (pdf) of the time-homogeneous Markov diffusion process generated by the GSR statistic when the Brownian motion under surveillance is "drift-free", i.e., in the pre-change regime; the GSR statistic's (deterministic) nonnegative headstart is assumed arbitrarily given. The transition pdf formula is found analytically, through direct solution of the respective Kolmogorov forward equation via the Fourier spectral method to achieve separation of the spacial and temporal variables. The obtained result generalizes the well-known formula for the (pre-change) stationary distribution of the GSR statistic: the stationary distribution is the temporal limit of the distribution sought in this work. To conclude, we exploit the obtained formula numerically and offer a brief study intended to characterize the pre-change behavior of the GSR statistic depending on three factors: (a) drift-shift magnitude, (b) time, and (c) the GSR statistic's headstart.

Recent Development of First Crossing Times of Compound Poisson Processes with Two Piecewise Linear Boundaries

Yifan Xu, *Case Western Reserve University*

Compound Poisson processes (CPP) are important in queueing theory, insurance risk modeling and SPRT. In sequential probability ratio tests of distributions from exponential families, with Poissonized sampling procedure, the first crossing time of CPP is used to calculate the exact power function of the test. In queueing theory, the crossing times reveal time dependent behaviors of M/G/1 queues with workload (virtual waiting

time in the first-come-first-served policy) restriction. To be more specific, the first upper crossing time is the first time that the workload of the queueing system exceeds a predetermined threshold; The lower crossing time is the busy period where the workload stays below the threshold.

In this talk we consider CPPs with positive absolutely continuous jumps. I will demonstrate the setup of the boundary crossing problem, and derive exact distributions of first crossing times of such CPPs with both upper and lower piece-wise linear boundaries. I will also show a numerical approximation method to reduce the calculation complexity in the case of parallel boundaries.

Multisensor Quickest Detection

Grigory Sokolov, *State University of New York at Binghamton*

Consider the multisensor quickest detection problem, in which the goal is to detect as soon as possible an abrupt change that occurs at some unknown time and modifies the observations of an unknown subset of sensors that monitor a system.

In this talk I will show the second-order asymptotic optimality of two families of detection rules: the Generalized CUSUM and a mixture CUSUM. Specifically, I will show that for each of these two schemes, the inflicted performance loss (relative to the optimal one that could be attained only if the identity of the affected subset was known) remains bounded as the rate of false alarms goes to 0 for any possible affected subset.

In the special case that it is known in advance that the change will affect exactly one sensor, I will revisit the multichart CUSUM, according to which an alarm is raised the first time a local CUSUM statistic exceeds a user-specified threshold. In this context, I will propose a family of thresholds that makes the multichart CUSUM uniformly second-order asymptotically optimal and, using insights from high-order asymptotic approximations, propose a specific selection of thresholds in this family that further robustifies this detection rule.

This is joint work with Georgios Fellouris (Department of Statistics, University of Illinois at Urbana-Champaign).

Detecting Changes in Complex Networks: Challenges and New Directions

Vasanthan Raghavan, *Qualcomm Flarion Technologies, Inc., Bridgewater, NJ*

Changepoint detection has a rich and multifaceted history with a rich array of problems having been explored over the past few decades. Much of this attention however has been on the i.i.d. problem, where the observations are independent and identically distributed, both before and after change and with the distributional parameters known. The increased focus on big data problems in statistics as well as emerging applications of changepoint detection in complex networks such as social and terrorist networks brings in a diverse array of problems, hitherto not studied. This talk discusses multiple challenges in this setting such as modeling the data, modeling the changepoint(s), tradeoffs between parametric vs. non-parametric approaches in detecting the change, etc.

Session 2: Statistical Challenges in Modeling and Applications

On Missing Data Mechanism in Two-Way Incomplete Contingency Tables

Daeyoung Kim, *University of Massachusetts - Amherst*
Seongyong Kim, *Hoseo University, Republic of Korea*

Contingency tables, displaying the frequency distribution of the categorical variables, are useful for analyzing the dependence between the variables. When some or all of the categorical variables have missing data, one needs to incorporate a missing data mechanism into a statistical model for valid inference on the association of incompletely observed categorical variables. In this talk we introduces a novel method of sensitivity analysis designed to aid in assessing the missing-at-random assumption in the two-way contingency table with one supplemental margin. The procedure involves a set of response odds and nonresponse odds computed from fully and partially observed counts. We illustrate the proposed method with real datasets.

Quantile Regression for Survival Data with Delayed Entry

Boqin Sun, *University of Massachusetts - Amherst*
Jing Qian, *University of Massachusetts - Amherst*

Delayed entry arises frequently in follow-up studies for survival outcomes, where additional study subjects enter during the study period. We propose a quantile regression model to analyze survival data subject to delayed entry and right-censoring. Such a model offers flexibility in assessing covariate effects on survival outcome and the regression coefficients are interpretable as direct effects on the event time. Under the conditional independent censoring assumption, we proposed a weighted martingale-based estimating equation, and formulated the solution finding as a ℓ_1-type convex optimization problem, which was solved through a linear programming algorithm. We established uniform consistency and weak convergence of the resultant estimators. We developed and justified a resampling inference procedure for variance and covariance estimation. The finite-sample performance of the proposed method was demonstrated via simulation studies. The proposed method was illustrated through an application to a real study.

Estimating Population Susceptibility in Dynamic Models of Infectious Disease

Nicholas Reich, *University of Massachusetts - Amherst*

A central challenge in modeling infectious diseases is accurately estimating the (largely unobservable) population of individuals who are susceptible to infection. In this talk, I will review methods used to infer the susceptible fraction from time-series data and incorporate these estimates into models of disease transmission. Additionally, I will propose a new approach to accounting for the susceptible population over time based on observed case data. This method provides a simple way to include complex dynamics in otherwise standard statistical time-series models. Using over four decades of surveillance data on dengue fever infections from the Ministry of Public Health in Thailand, I will compare the ability of these methods to draw inference about mechanistic disease transmission models.

Variable Selection in Single-Index Varying Coefficient Models

Anna Liu, *University of Massachusetts - Amherst*
Peng Wang, *Google, Inc.*

Single index varying coefficient model is an attractive statistical model with its ability to hand high dimensional data and its ease of interpretation. Motivated by a Geoscience project and a TV rating project from the advertisement industry, we study the problem of index variable selection in the single index varying coefficient model. We consider both regression and classification problems, and use LASSO type of penalty for the variable selection purpose. We propose a new and easy-to-implement algorithm for the optimization which consists of two steps alternating between estimating the coefficient functions, and selecting/estimating the single index. We illustrate our algorithm with the above mentioned applications and our R package.

Semi-parametric Time to Event Models in the Presence of Error-prone, Self-reported Outcomes - with Application to the Womens Health Initiative

Xiangdong Gu, *University of Massachusetts - Amherst*
Yunsheng Ma, *University of Massachusetts Medical School - Worcester*
Raji Balasubramanian, *University of Massachusetts - Amherst*

The onset of several silent, chronic diseases such as diabetes can be detected only through diagnostic tests. Due to cost considerations, self-reported outcomes are routinely collected in lieu of expensive diagnostic tests in large-scale, prospective investigations such as the Womens Health Initiative. However, self-reported outcomes are subject to imperfect sensitivity and specificity. Using a semi-parametric likelihood-based approach, we present time to event models to estimate the association of one or more covariates with a error-prone, self-reported outcome. We present simulation studies to assess the effect of error in self-reported outcomes with regard to bias in the estimation of the regression parameter of interest. We apply the proposed methods to prospective data from 152,830 women enrolled in the Womens Health Initiative to evaluate the effect of statin use with the risk of incident diabetes mellitus, among postmenopausal women. The current analysis is based on follow up through 2010, with a median duration of follow up of 12.1 years. The methods proposed in this paper are readily implemented using our freely available R software package *icensmis*, which is available at the Comprehensive R Archive Network (CRAN) website.

Session 3: Statistical Innovations in Biomedical Research

A Nonparametric Approach to Comparing Diagnostic Accuracies in a Multi-Reader, Multi-Test Design

Eunhee Kim, *Brown University*
Zheng Zhang, *Brown University*
Youdan Wang, *Brown University*
Donglin Zeng, *University of North Carolina at Chapel Hill*

Receiver operating characteristic (ROC) analysis is widely used to evaluate the performance of diagnostic tests with continuous or ordinal responses. A popular study design for assessing the accuracy of diagnostic tests involves multiple readers interpreting multiple diagnostic test results, called the multi-reader, multi-test design. I will introduce a novel power formula to compare the correlated areas under the ROC curves (AUC) in a multi-reader, multi-test design. In particular, I will describe a nonparametric approach to estimate and

compare the correlated AUCs by extending DeLong et al.s (1988) approach.

Effective Detection of Differences in Related Mixture Distributions

Li Ma, *Duke University*

In numerous scientific investigations, a central statistical problem is the comparison of two or more data sets to identify the difference in the underlying probability distributions. The distributions being compared are often related and structurally similar. In this talk, we consider the comparison of related mixture distributions. Instead of modeling the distributions separately, we construct a joint hierarchical model that incorporates a structural assumption that the distributions tend to share some mixture components (or subpopulations) though the shape of such a shared component may vary slightly due to cross-study variation. We show that such a joint modeling approach allows effective borrowing of information across the samples and thereby results in substantially improved power compared to methods that ignore the underlying structural relationship among the samples. We introduce and illustrate the work of our method through analyzing a flow cytometry data set.

Analysis of Gene Expression at The Single-Cell Resolution

Guo-Cheng Yuan, *Dana-Farber Cancer Institute*

Single-cell gene expression profiling has recently emerged as a powerful tool for the investigation of cellular heterogeneity. However, computational methods to analyze such data are still lacking. Here we present two new methods that, in combination, are able to not only detect cellular hierarchy but to quantify dynamic changes of gene expression patterns. In the first method, we use an ensemble-based approach to infer cell hierarchy. In the second method, we use tools from dynamical systems theory to identify the bifurcation events during cell differentiation and model the dynamic changes by using stochastic differential equations. We applied these methods to analyze publically available single-cell data obtained using Fluidigm and CyTOF technologies, and compared the performance of our methods with SPADE. We found that our methods improve the robustness of cellular hierarchy prediction and further provided temporal information. Our bifurcation analysis further identified the key initial events during cell differentiation, and correctly predicted the effect of perturbation of key regulators on cell-fate transitions. As such, we provide two useful computational tools for single-cell gene expression data analysis.

Session 4: Recent Advances in Spatial Statistics

Compact, Disjoint Bases for Spatio-Temporal Point Processes

Luke Bornn, *Harvard University*
Andrew Miller, *Harvard University*
Alex Franks, *Harvard University*
Alex D'Amour, *Harvard University*
Daniel Cervone, *Harvard University*

In this talk I will show how disjoint and compact bases provide a natural and intuitive approach to modeling spatial and spatio-temporal point processes. I will argue that inducing spatially compact bases allows for more intuitive representation of many real-world point processes. I will demonstrate how such a result can be obtained through non-negative matrix factorization, and subsequently extend this idea to a full generative model.

Bayesian Spatial Modeling of the Local Persistence of PCV-Targeted Pneumococcal Serotypes among Adults in Connecticut, 1998-2009

Joshua Warren, *Yale University*
Daniel Weinberger, *Yale University*

The pneumococcal conjugate vaccine (PCV) has been used in children for over a decade, and there is an ongoing debate about whether adults should also receive the vaccine. The original PCV targeted seven of the 90+ pneumococcal serotypes, and pediatric immunization had a strong indirect effect on the incidence of disease in adults. However, despite the impact of indirect protection, the burden of pneumococcal disease in adults remains high. Adult disease cases in the post-PCV period are caused by non-vaccine serotypes and, to a lesser degree, by the persistence of vaccine-targeted serotypes. The use of PCVs in adults would be justified if the vaccine-targeted serotypes persist at a high level in some communities. Using previously collected surveillance data from the state of Connecticut, 1998-2009, we introduce new methodology in the Bayesian setting to determine whether there are communities where vaccine-targeted serotypes cause a disproportionately large percentage of adult disease cases in the post-PCV period. Our individual-level probit regression model jointly allows for the possibility of spatially varying baseline probabilities of disease caused by vaccine-targeted serotypes, spatially varying rates of decline of these probabilities over time, and spatially varying dates at which this decline begins in the adult population. Results suggest that considering spatial variability in the parameters leads to improved model fit in terms of explanation and prediction of the process. We further explore the results to determine if the spatial variability in the dates where the decline begins can be explained by spatially varying covariates and to validate the proposed models.

Bayesian Spatial Hierarchical Models for Small Estimation with Complex Survey Designs

Cici Bauer, *Brown University*

Spatial hierarchical models have shown to be beneficial for small area estimation. However, the sampling weights that are required to reflect complex surveys are rarely considered in these models. In this talk, I will describe a method for incorporating the sampling weights for binary data when estimating, for example, small area proportions or predicting small area counts. Spatial hierarchical random effects are shown to be beneficial, with computation carried out using the integrated nested Laplace approximation, which is fast. Simulation results will be presented to show that estimation of mean squared error can be reduced when compared with more standard approaches. Bias reduction occurs through the incorporation of sampling weights, with variance reduction being achieved through hierarchical smoothing. The application of our proposed method with data taken from the Washington 2006 Behavioral Risk Factor Surveillance System will also be presented

Disease Risk Estimation by Combining Spatial Case-Control Data with Aggregated Information on the Population at Risk

Xiaohui Chang, *Oregon State University*

We propose a novel statistical framework by supplementing spatial case-control data with summary statistics on the population at risk for a subset of risk factors. Under a general spatial point process setting, our approach is to first form two unbiased estimating equations, one based on the case-control data and the other on both the case data and the summary statistics, and then optimally combine them to derive another estimating equation to be used for the estimation. The proposed method is computationally simple and more efficient than standard approaches based on case-control data alone. We also establish asymptotic properties of the resulting estimator, and investigate its finite-sample performance through simulation. As a substantive application, we apply the proposed method to investigate risk factors for endometrial cancer, by using data from a recently completed population-based case-control study and summary statistics from the Behavioral Risk Factor Surveillance System, the Population Estimates Program of the US Census Bureau, and the Connecticut Department of Transportation.

Bayesian Spatio-Temporal Point Level Modelling of Air Pollution Concentration Levels for Estimating Long Term Exposure in Coarser Administrative Geographies in the UK

Sujit K. Sahu, *University of Southampton, UK*
Sabyasachi Mukhopadhyay, *University of Southampton, UK*

Estimation of long term exposure to air pollution levels over a large spatial domain, such as the mainland UK, entails a challenging modelling task since exposure data are often only observed by a network of sparse monitoring sites with large percentages of missing data. This article develops and compares several flexible anisotropic and non-stationary hierarchical Bayesian models for four most harmful air pollutants in England and Wales during the five year period 2007–2011. The models make use of observed data from the UK's AURN (Automatic Urban and Rural Network) as well as output of an atmospheric air quality dispersion model developed recently especially for the UK. Land use information, incorporated as a predictor in the model, further enhances the accuracy of the model. Indeed, out-of-sample spatial predictions are found to be more accurate than both the air quality model output and kriging of the observations. Using daily data for four pollutants: nitrogen dioxide, ozone, and two variants of particulate matter, over the five year period we obtain empirically verified accurate maps of air pollution levels for England and Wales. Monte Carlo integration methods for model based spatial aggregation are developed and these allow us to obtain predictions, and their uncertainties, at the level of a given administrative geography. These estimates for local authority areas, available on request, can readily be used for many purposes such as modelling of aggregated health outcome data. This talk will also present preliminary results in estimating health effects of air pollution using a fully Bayesian two stage modelling approach for air pollution and health outcome data.

Session 5: Statistical Inference in Time Series and Machine Learning

Statistical Inference for Perturbations of Multiscale Dynamical Systems

Siragan Gailus, *Boston University*
Konstantinos Spiliopoulos, *Boston University*

In this paper, we study statistical inference for small-noise perturbations of multiscale dynamical systems. We prove the consistency and asymptotic normality of an appropriately constructed maximum likelihood estimator (MLE) for a parameter of interest, identifying precisely the limiting variance. We allow unbounded coefficients in the equation for the slow process and assume neither periodicity nor that the fast process is compact - ergodicity of the fast process is guaranteed by imposing a recurrence condition. The results provide a theoretical basis for calibration of small-noise perturbed multiscale dynamical systems and related diffusion processes. Data from numerical simulations are provided to supplement and illustrate the theory.

Parameter Estimation for Continuous-time Stationary Models with Memory

Mamikon Ginovyan, *Boston University*

Let $\{X(t),\ t \in \mathbb{R}\}$ be a zero mean real-valued continuous-time stationary process with spectral density $f(\lambda, \theta)$, where $\theta := (\theta_1, \dots, \theta_p) \in \Theta \subset \mathbb{R}^p$ is an unknown vector parameter. The problem of interest is to investigate whether various statistical estimators of θ, constructed on the basis of an observed finite realization $X_T := \{X(u),\ 0 \le u \le T\}$, possess "nice" statistical properties (consistency, asymptotic normality, asymptotic efficiency), depending on the memory structure of the model $X(t)$ and the smoothness of its spectral density f.

As an estimator for unknown θ, we consider the Whittle-type estimator $\hat{\theta}_W(X_T)$, the statistic that minimizes the weighted Whittle contrast functional, and derive conditions (in terms of spectral density f) under which the estimator $\hat{\theta}_W(X_T)$ possesses the above "nice" properties.

To prove the asymptotic properties of the estimator $\hat{\theta}_W(X_T)$, we first establish central limit theorems for empirical spectral functionals, which are Toeplitz-type quadratic functionals of the process $X(t)$.

Feature Vector Denoising with Prior Network Structures

Mark A. Kon, *Boston University*

Problems in machine learning (ML) often involve noisy input data, and ML classification methods have in some cases reached limiting accuracies when they are based on standard ML data sets, e.g. consisting of feature vectors and classes. An important step toward greater accuracy in ML will require incorporation of prior structural information on data into learning. We will denote methods that regularize feature vectors as unsupervised regularization methods, analogous to supervised regularization methods used to estimate functions in machine learning. We study regularization (denoising) of ML feature vectors using analogues of Tikhonov and other regularization methods for functions on $\mathbb{R}^n$. A feature vector $\mathbf{x} = (x_1, \ldots, x_n) = \{x_q\}_{q=1}^n$ is viewed as a function of its index q, and smoothed using some prior information on the structure of the feature vector. This can involve a penalty functional on feature vectors analogous to those in statistical learning, or use of some proximity (e.g. graph) structure on the set of indices q (the *index space*). Such feature vector regularization inherits a property from function denoising on $\mathbb{R}^n$, in that denoising accuracy is non-monotonic in the denoising (regularization) parameter α. We show that the best reconstruction accuracy occurs at a positive α in index spaces with graph structures. We adapt three standard function denoising methods used on $\mathbb{R}^n$, local averaging, kernel regression, and support vector regression. In general the index space can be any discrete set with a notion of proximity, e.g. a metric space, a subset of $\mathbb{R}^n$, or a graph/network, with feature vectors as functions with some notion of continuity. We show this improves feature vector recovery, and thus the subsequent classification or regression done on them. We give an example in gene expression analysis for cancer classification, with the genome as an index space with a network structure based on prior knowledge of protein-protein interactions.

Semiparametric Model Building for Regression Models with Time-Varying Parameters

Ting Zhang, *Boston University*

We consider the problem of semiparametric model building for linear regression models with potentially time-varying coefficients. By allowing the response variable and explanatory variables be jointly a nonstationary process, the proposed methods are widely applicable to nonstationary and dependent observations. We propose a local linear shrinkage method that can simultaneously achieve parameter estimation and variable selection. Its selection consistency and the favorable oracle property are established. Due to the fear of losing efficiency, an information criterion is further proposed for distinguishing between time-varying and time-constant components. Numerical examples are presented to illustrate the proposed methods.

Session 6: Recent Developments in Analyzing Survival Endpoint Methods After Taking Alternative Therapy/Treatment Switching in Oncology Trials

Analyzing Time to Event Data in the Presence of Informative Censoring - A Regulatory Submission Case Study

Jing Xu, *Takeda Pharmaceuticals, Inc.*

One of the important assumptions in survival analysis on time to event endpoints is that the censoring and endpoint outcome are independent. When this assumption is not met, the results may be biased or even not valid. In clinical trials, informative censoring may occur for time to event data. In this case study, we present such a case, and summarize how we conducted sensitivity analysis using IPCW approach to address the concern from regulatory agency.

The *g*-formula to Adjust for Noncompliance in Randomized Clinical Trials

Jessica Young, *Harvard T.H. Chan School of Public Health*

Causal inference based on randomized clinical trials may be compromised when some subjects fail to follow their assigned treatment strategy at some time after the randomization. In this talk, we review Robins' g-formula (1986) as a means of identifying and estimating the causal effect of following different treatment strategies in a randomized trial where treatment switching or any form of noncompliance occurs. Specifically, we review sufficient identifying conditions in a general time-varying setting where (i) measured patient characteristics at a given time are associated with both future outcomes and future treatment and (ii) these characteristics are themselves affected by past treatment. We also review some approaches to effect estimation motivated by these identification results which may be applied in typical high-dimensional settings. These include the parametric g-formula and inverse probability weighted estimation.

Comparing the Performance of Several Re-Censoring Rules in the RPSFT Method for the Analysis of OS Data Adjusting for Treatment Switching in the Phase III Oncology Trials

Huyuan Yang, *Takeda Pharmaceuticals, Inc.*
Zheng Yuan, *Agios Pharmaceuticals, Inc.*

RPSFT (Rank Preserving Structural Failure Time) method is one of the widely used methods in analyzing adjusted OS data after patients progressed and switched from control group to active group in the Phase III oncology trials. The current recommended re-censoring rule for censored patients of RPSFT has limitation. It might be too conservative causing bias toward active treatment arm and inflated type I error rate based on the model based variance estimator. On the other hand, the current recommended method for variance estimation is through re-sampling based bootstrap, which leads to large p-values and wide 95% CI's for the hazard ratios (HRs); as a result, even when HR is in the 0.5 range, the p-values are still greater than 0.05 in a lot of cases. Therefore, here we consider several different ways of doing re-censoring and compare their performance in terms of power and type I error rate through simulations based on a Phase 3 randomized study for multiple myeloma patients. The comparison results demonstrate that the current re-censoring rule could be improved to reduce the inflated type I error rate.

Session 10: Statistical Methods and Computing with Big Data

Computational Tools for Bayesian Methods in Big Data and Data Science

Erin Conlon, *University of Massachusetts*

Recently, new Bayesian statistical approaches have been developed for big data sets that are too large to analyze in their entirety, due to restrictions on either computer memory or storage capacity. These methods partition big data sets into subsets, and implement parallel Bayesian Markov chain Monte Carlo analyses independently on the subsets. The independent subset posterior samples are then combined to produce estimated posterior densities based on the complete data set. There are several approaches to combining the subset samples, including averaging, weighted averaging (the Consensus Monte Carlo approach) and kernel smoothing techniques. Here, we introduce computational tools to carry out these methods, and compare results of the different strategies using both simulated and real data examples.

High Performance Data I/O

Taylor Arnold, *AT&T Labs Research*
Michael Kane, *Yale University*
Simon Urbanke, *AT&T Labs Research*

Anyone dealing with large data knows that stock tools in R are bad at loading (non-binary) data to R. We present the iotools package that provides high-performance parsing tools to minimize copying and avoid the use of strings when possible. To allow processing of arbitrarily large files we have added way to process chunk-wise input, making it possible to compute on streaming input as well as very large files. Finally, the hmr package is an extension of iotools that wraps the fast loading and chunkwise processing to allow for seamless execution over files stored in a Hadoop cluster.

Scalable, Exact Approaches to Fitting Linear Models when $n \gg p$

Michael Kane, *Yale University*

A common challenge in data analysis is the regression of p variables onto a single variable where the number of samples (n) may be very large. Traditional, numerically stable methods pose challenges since a QR decomposition of the entire matrix is feasible but may be time consuming and difficult to implement, especially in distributed settings. Chunk-wise implementations that directly solve the normal equations provide a simpler approach, trading numerical stability for performance. However, current chunk-wise implementations face two challenges. The performance is limited by disk I/O and they are written either for a single machine or for a cluster. This talk explores these challenges and shows how the ioregression package addresses these challenges by making use of the iotools package, which can speed up disk access by up to and order of magnitude and by providing regression kernels that have been used to implement chunk-wise regressions for any number of configurations.

Online Updating of Statistical Inference in the Big Data Setting

Elizabeth Schifano, *University of Connecticut*
Jing Wu, *University of Connecticut*
Chun Wang, *University of Connecticut*
Jun Yan, *University of Connecticut*
Ming-Hui Chen, *University of Connecticut*

We present statistical methods for big data arising from online analytical processing, where large amounts of data arrive in streams and require fast analysis without storage/access to the historical data. In particular, we develop iterative estimating algorithms and statistical inferences for linear models and estimating equations that update as new data arrive. The online updating framework introduces predictive residuals that can be used to test the goodness-of-fit of the hypothesized model. We also present a new online-updating estimator under the estimating equation setting. In simulation studies and real data applications, our estimator compares favorably with competing approaches.

Further Growth of the Bigmemory Family of Packages

John Emerson, *Yale University*

This talk will outline work in progress to extend bigmemory's big.matrix objects for support of (1) massive vectors of character strings and (2) massive data.frame-like objects. With an assist from Steve Weston's foreach and iterators packages, these new packages allow convenient new big.read.csv() functionality. The term "massive" here includes "larger-than-RAM', and these objects are also available in shared memory for efficient parallel computing.

Session 11: Statistical Applications and Practice

A Metamodeling Approach to the Valuation of Large Variable Annuity Portfolio under Nested Simulation

Guojun Gan, *University of Connecticut*

A variable annuity (VA) is equity-linked annuity product that has rapidly grown in popularity around the world in recent years. Research up to date on VA largely focuses on the valuation of guarantees embedded in a single VA contract. However, methods developed for individual VA contracts based on option pricing theory cannot be extended to large VA portfolios. Insurance companies currently use nested simulation to valuate guarantees for VA portfolios but efficient valuation under nested simulation for a large VA portfolio has been a real challenge. The computation in nested simulation is highly intensive and often prohibitive. In this talk, I will introduce a metamodeling approach that combines a clustering technique with a functional data analysis technique to address the issue.

Measuring Online Audiences Using Logs Data

Xiaojing Wang, *Google, Inc.*
Jim Koehler, *Google, Inc.*

Advertisers would like to understand attributes of the audience their ads reach. TV marketers, in particular, are accustomed to metrics such as the reach and frequency for their TV ads, and would welcome digital counterparts. Existing digit reporting practices, however, measure cookies rather than people and thus present significant technical challenges: a cookie does not identify a person, and anyone who uses multiple accounts, devices, or browsers has multiple cookies; a cookie can expire or be deleted; the demographic information attached to some cookies may be of questionable quality. We introduce a method to correct these issues in aggregate and measure the reach and frequency of online ad campaigns by audience attributes. It reduces demographic biases and adjusts cookie counts to unique people counts by using data from probability-recruited online panels.

The Optimal Mix of TV and Online Ads

Georg Goerg, *Google, Inc.*

Brand marketers often wonder how they should allocate budget between TV and online ads in order to maximize reach or maintain the same reach at a lower cost. We use probability models based on historical cross media panel data to suggest the optimal budget allocation between TV and online ads to maximize reach to the target demographics. We take a historical TV campaign and estimate the reach and GRPs of a hypothetical cross-media campaign if some budget was shifted from TV to online. The models are validated against simulations and historical cross-media campaigns. They are illustrated on one case study to show how an optimized cross-media campaign can obtain a higher reach at the same cost or maintain the same reach at a lower cost than the TV-only campaign.

Statistical Applications in Property and Casualty Insurance

James Landgrebe, *Travelers Insurance*

This presentation will provide an overview of some common analytical challenges faced by actuaries and statisticians working in the property and casualty insurance industry. It will highlight how reliance on advanced analytic expertise continues to expand, encompassing an increasingly broad range of practical business issues.

Session 12: Recent Advances in Subgroup Analyses

Subgroup Reporting using Nonparametric Bayesian Inference

Peter Müller, *University of Texas at Austin*

We discuss Bayesian inference for subgroups in clinical trials. The key feature of the proposed approach is that we separate the decision problem of selecting subgroups for reporting and the probability model for prediction. For the latter we use a flexible nonparmetric Bayesian model, while the decision problem is based on a parsimonious description of the subgroups and a stylized utility function.

Bayesian Approaches to Subgroup Analysis and Selection Problems in Drug Development

David Ohlssen, *Novartis*

In drug development setting the challenge of dealing with selection problems regularly arises. For example: the interpretation of pre-planned or post-hoc exploratory subgroup analyses; safety signal detection, which can involve a large number of outcomes related to different system organ classes; estimating a treatment effect when a trial stops early for success in a group sequential design and selecting the dose to take to phase III following a phase II dose-finding study. From a statistical perspective these types of problems can be divided between the need to estimate an effect of interest accounting for a potential selection or random high bias and dealing with multiplicity when examining numerous potential signals or subgroups. When considering the former, the Bayesian framework provides the ability to incorporate priors with a degree of skepticism, a natural framework for forming models with exchangeability or shrinkage and the possibility to form realistically complex models allowing synthesis of information from a variety of sources. While in the case of the latter, Bayesian approaches to hypothesis testing and extensions of false discovery rate provide potential techniques to handle multiplicity. In this talk we shall provide an overview of selection and subgroup problems occurring in medical product development. We shall also briefly review some key techniques from the Bayesian framework that can help tackle such problems. The final part of the presentation will look more in depth at a series of examples

Effective Subgroup Identification by Systematically utilizing Multiple Baseline Characteristics

Lihui Zhao, *Northwestern University*
Lu Tian, *Stanford University*
Tianxi Cai, *Harvard University*
Brian Claggett, *Harvard Medical School*
Lee-Jen Wei, *Harvard University*

When comparing a new treatment with a control in a randomized clinical study, the treatment effect is generally assessed by evaluating a summary measure over a specific study population. The success of the trial heavily depends on the choice of such a population. In this research, we show a systematic, effective way to identify a promising subgroup, for which the new treatment is expected to have a desired benefit, using the data from current studies involving similar comparator treatments. Specifically, we first create a parametric scoring system using multiple covariates to estimate subject-specific treatment differences. Using this system, we specify a desired level of treatment difference and create a subgroup of patients, defined as those whose estimated scores exceed this threshold. An empirically calibrated group-specific treatment difference curve across a range of threshold values is constructed. The population of patients with any desired level of treatment benefit can then be identified accordingly. To avoid any "self-serving" bias, we utilize a cross-training-evaluation method for implementing the above procedure. At the final stage, we validate such a subgroup selection via two-sample inference procedures for assessing the treatment effectiveness statistically and clinically with a holdout sample. The proposals are illustrated with real data from a cardiovascular study.

A Bayesian Approach for Subgroup Analysis

Xiaojing Wang, *University of Connecticut*

This talk discusses subgroup analysis, the goal of which is to determine the heterogeneity of treatment effects across subpopulations. Searching for differences among subgroups is challenging because it is inherently a multiple testing problem with the complication that test statistics for subgroups are typically highly dependent, making simple multiplicity corrections such as the Bonferroni correction too conservative. In this article, a Bayesian approach to identify subgroup effects is proposed, with a scheme for assigning prior probabilities to possible subgroup effects that accounts for multiplicity and yet allows for (preexperimental) preference to specific subgroups. The analysis utilizes a new Bayesian model selection methodology and, as a by-product, produces individual probabilities of treatment effect that could be of use in personalized medicine. The analysis is illustrated on an example involving subgroup analysis of biomarker effects on treatments.

This is the joint work with Dr. James O. Berger and Dr. Lei Shen.

Session 13: Pharmaceutical Applications

Questionnaire on Network Meta-Analysis to Assess Its Relevance and Credibility

Joseph Cappelleri, *Pfizer Inc*

Despite the great realized or potential value of network meta-analysis (indirect treatment comparisons) of randomized controlled trial evidence to inform health-care decision making, many decision makers might not be familiar with these techniques. A task force from the International Society for Pharmacoeconomics and Outcomes Research developed a consensus-based 26-item questionnaire to help decision makers assess the relevance and credibility of network meta-analysis to help inform health-care decision making (Jansen et al. Value in Health 2014; 17:157-173).

The relevance domain of the questionnaire (4 questions) calls for assessments about the applicability of network meta-analysis results to the setting of interest to the decision maker. The remaining 22 questions belong to an overall credibility domain and pertain to assessments about whether the network meta-analysis results provide a valid answer to the question they are designed to answer by examining 1) the evidence base used, 2) analysis methods, 3) reporting quality and transparency, 4) interpretation of findings, and 5) conflicts of interest.

The questionnaire aims to help readers of network meta-analysis opine about their confidence in the credibility and applicability of the results of a network meta-analysis, and help make decision makers aware of the subtleties involved in the analysis of networks of randomized trial evidence. It is anticipated that user feedback will permit periodic evaluation and modification of the questionnaire.

An Introduction to Multiple Testing Procedures with an Example using a Graphical Approach

Nate Bennett, *Boehringer Ingelheim*

In pharmaceutical research, there are often many competing hypotheses to test. To help ensure that a finding is real (and was not found by chance), we need to control for multiple comparisons. There are many reasons for the multitude of tests, ranging from new genetic testing capabilities to looking at different population subgroups to look for different safety or efficacy results.

When comparing multiple hypotheses, there are many different instruments to use. In this talk we will recap many of the commonly taught methods from statistics courses. We will expand the list with techniques developed in the last generation before coming to a very new approach which can be illustrated graphically.

Characterizing Disease Progression with Multivariate Longitudinal Models: An Example in Alzheimer's Disease

Dan Polhamus, *Metrum Research Group*
James Rogers, *Metrum Research Group*
Jonathan French, *Metrum Research Group*

The effects of therapeutic intervention are difficult to assess for a number of diseases that progress through qualitatively different stages over several years. Different outcome variables, all understood to be reflective of a single disease process, may be more or less informative depending on the disease stage. In Alzheimer's Disease for example, certain memory assessments are very sensitive to change at the incipient and early stages of the disease and less so at later stages, while for instruments measuring other aspects of cognition, the reverse may be true. In this context, comprehensive analysis of data spanning multiple stages of the disease requires a joint (multivariate) longitudinal model. Our approach to this problem is to posit a latent (univariate) disease state that progresses continuously over time and that determines (via link functions) location parameters for individual outcome variables. We present an application of this approach in Alzheimer's Disease. Brief remarks will be made regarding model identifiability, diagnostics, and the "value added" over separate univariate longitudinal models.

Comparison of Treatment Response and Loss of Response Definitions in Enriched Enrollment Randomized Withdrawal Design Pain Studies

Birol Emir, *Pfizer, Inc.*
Ed Whalen, *Pfizer, Inc.*

Enriched enrollment randomized withdrawal (EERW) studies are composed of an enrollment period during which responders to the treatment are identified and randomized into the treatment arms of the withdrawal period. This analysis was conducted to assess the different clinically relevant definitions of treatment response and loss of therapeutic response (LTR) in EERW design studies. Two treatment response randomization criteria, $\geq 30\%$ and $\geq 50\%$ improvement in daily pain diary scores, were examined. LTR was defined using baseline pain scores compared to the moving averages of daily pain scores during the RW period (RWMA). 3-day, 5-day, and 7-day RWMAs were examined and compared to either baseline pain score pre-EE or at randomization to generate 6 different LTR definitions. All LTRs were analyzed using standard time-to-event methods. To compare these different definitions of LTR, simulations using randomly selected patients from 5 pain studies that used different patient populations were conducted, with studies of size n=75, 100 and 125 per treatment (active and placebo). Risk of type I error (false positive finding) was assessed by repeating the process using data from patients randomized to placebo. We will discuss our simulation strategy, why we choose the one we used and show our recommendation for a future study.

Session 14: Design and Analysis of Complex Experiments

Analysis of Cluster Encouragement Designs with Interference

Laura Forastiere, *Harvard University*
Fabrizia Mealli, *University of Florence and Harvard University* Tyler VanderWeele, *Harvard University*

Encouragement design studies arise frequently when the treatment cannot be enforced because of ethical or practical constrains and an encouragement intervention is conceived with the purpose of increasing the uptake of the treatment of interest. By design, encouragements always entail the complication of non-compliance. Encouragements can also give rise to a variety of mechanisms, particularly when assigned at the cluster level. Social interactions among units within the same cluster can result in spillover effects. Disentangling the effect of encouragement through spillover effects from that through the enhancement of the treatment is our goal. We use the principal stratification framework to define stratum-specific causal effects, showing how stratum-specific causal effects are related to the decomposition commonly used in the literature and provide flexible homogeneity assumptions under which an extrapolation across principal strata allows to disentangle the effects. Estimation of causal estimands is performed with Bayesian inferential methods using hierarchical models to account for clustering. We illustrate the proposed methodology analyzing a cluster randomized experiment implemented in Zambia and designed to evaluate the impact on malaria prevalence of an agricultural loan program intended to increase the bed net coverage.

A Potential Outcomes Based Perspective of the Analysis of Complex Multi-Factor Experiments with Randomization Restrictions

Tirthankar Dasgupta, *Harvard University*
Joseph Lee, *Harvard University*
Anqi Zhao, *Harvard University*
Peng Ding, *Harvard University*
Donald B. Rubin, *Harvard University*

The first formal notation for potential outcomes was introduced by Neyman (1923) for randomization-based inference in randomized experiments. Potential outcomes were subsequently used by several authors during 1930–1960 for causal inference from randomized experiments. The lack of adequate computational power and the difficulty in asymptotic analysis associated with complex experiments were probable causes behind the gradual reduction of interest in such randomization-based methods. In this talk, we re-visit the randomization analysis of multi-treatment and multi-factor experiments with randomization restrictions and discuss analyses of such experiments using the Fisherian and Neymanian approaches.

Causal Inference with Partially Revealed Interference

Panos Toulis, *Harvard University*
Edoardo M. Airoldi, *Harvard University*
Donal B. Rubin, *Harvard University*

In many practical applications it is suspected that the treatment of one unit actively affects the outcomes of other units. In such cases, in addition to observing outcome data, we can also observe *connections* between units whose treatments affect other units.

In this work, we extend the potential outcomes model of causal inference to assess causal effects when such forms of interference exist, which we term *causal cupid effects* for the following reason. When two units are not connected, then there is no cupid effect, and one unit's treatment only affects that unit's outcome. However, when two units are connected, one unit's treatment assignment generally affects both of their outcomes. Thus, the connections create a binary directed network among units that is only partially observed. We illustrate our causal framework in applications where such forms of interference are ubiquitous, but currently not adequately addressed. Our analysis of data generated by such situations, uses both Bayesian and frequentist methods. For example, our hypothesis test for causal cupid effects relies on Bayesian posterior predictive p-values, and entails repeated imputation of the missing network of connections between units under the null hypothesis of no cupid effect, using a suitable test statistic, for example, the MLE of a causal parameter under an alternative hypothesis.

Session 15: Association and Correlation Analysis for Big Data

High-Frequency Financial Risk Management and High Performance Computing

Jian Zou, *Worcester Polytechnic Institute*

Financial statistics covers a wide array of applications in the financial world, such as (high frequency) trading, risk management, pricing and valuation of securities and derivatives, and various business and economic analytics. In this article, we focus on the portfolio allocation problem using high-frequency financial data, and propose a hybrid parallelization solution to carry out efficient asset allocations in a large portfolio via intra-day high-frequency data. We exploit a variety of HPC techniques, including parallel R, Intel Math

Kernel Library, and automatic offloading to Intel Xeon Phi coprocessor in particular to speed up the simulation and optimization procedures in our statistical investigations. Using a combination of software and hardware parallelism, we demonstrate a high level of performance on high-frequency financial statistics.

A Bayesian Test of Independence for Each of Several Sparse Two-Way Contingency Tables

Balgobin Nandram, *Worcester Polytechnic Institute*
Dalho Kim, *Kyungpook National University*

We consider testing for independence in each of several sparse two-way contingency tables. Techniques of small area estimation are used to construct a pooled Bayes test via a hierarchical Bayesian model. We use the Bayes factor to construct the pooled Bayes test and the required marginal likelihoods are estimated in a unified manner. We show how to obtain the Monte Carlo estimator of the Bayes factor and its standard error. An example is used to compare the pooled Bayes test with two direct (no pooling) tests, Cressie-Read power divergence test and a Bayes test. The three tests give similar evidence against independence for moderate to large areas, but there are some differences for smaller areas. As expected, while the direct Bayes test is sensitive to the prior specifications, the pooled Bayes test is not so sensitive. Moreover, for relatively small areas the pooled Bayes test has higher power than the two competitors. Finally, we discuss how to improve the pooled Bayes test even further using surrogate sampling for covariates and the Dirichlet process for robustification.

Covariance Matrix Estimation in Big Data: Approaches Based on Algebraic Properties

Xi Luo, *Brown University*

There is a growing need to understand the complex relationships among a humongous number of variables. Classical approaches, such as penalized likelihood, have provided powerful tools in understanding lower dimensional structures in multivariate data, but are facing increasing challenges in computation and theory. In this talk, I will present a framework for sparse inverse covariance estimation that exploits matrix algebraic properties. These properties lead us to develop simple (convex) optimization criteria, which enjoy a few advantages. The optimization problems are decomposable, and we develop efficient algorithms that can solve large-scale problems, even if the number of variables is in hundreds of thousands. The optimization problems are also designed to achieve faster convergence rates than classical approaches. I will illustrate the numerical merits of this framework using a few real data examples as well as simulations.

Inferring Anchor Links across Heterogeneous Social Networks

Xiangnan Kong, *Worcester Polytechnic Institute*
Jiawei Zhang, *University of Illinois at Chicago*
Philip S. Yu, *University of Illinois at Chicago*

Online social networks can often be represented as heterogeneous information networks containing abundant information about: who, where, when and what. Nowadays, people are usually involved in multiple social networks simultaneously. The multiple accounts of the same user in different networks are mostly isolated from each other without any connection between them. Discovering the correspondence of these accounts across multiple social networks is a crucial prerequisite for many interesting inter-network applications, such as link recommendation and community analysis using information from multiple networks. In this paper, we study the problem of anchor link prediction across multiple heterogeneous social networks, i.e., discovering the correspondence among different accounts of the same user. Unlike most prior work on link prediction and network alignment, we assume that the anchor links are one-to-one relationships (i.e., no two edges share

a common endpoint) between the accounts in two social networks, and a small number of anchor links are known beforehand. We propose to extract heterogeneous features from multiple heterogeneous networks for anchor link prediction, including user's social, spatial, temporal and text information. Then we formulate the inference problem for anchor links as a stable matching problem between the two sets of user accounts in two different networks. An effective solution, MNA (Multi-Network Anchoring), is derived to infer anchor links w.r.t. the one-to-one constraint. Extensive experiments on two real-world heterogeneous social networks show that our MNA model consistently outperform other commonly-used baselines on anchor link prediction.

Robust Principal Component Analysis for Detecting Sparsely Correlated Phenomena in Computer Networks

Randy Paffenroth, *Worcester Polytechnic Institute*

In this talk we will present theory and algorithms for detecting weak distributed patterns in network data. The patterns we consider are sparse correlations between signals recorded at sensor nodes across a network. We use robust matrix completion and second order analysis to detect distributed patterns that are not discernible at the level of individual sensors. When viewed independently, the data at each node cannot provide a definitive determination of the underlying pattern, but when fused with data from across the network the relevant patterns emerge. We are specifically interested in detecting weak patterns in computer networks where the nodes (terminals, routers, servers, etc.) are sensors that provide measurements (of packet rates, user activity, CPU usage, etc.). The approach is applicable to many other types of sensor networks including wireless networks, mobile sensor networks, and social networks where correlated phenomena are of interest.

Session 17: Advances in Molecular Evolution and Statistics Genetics

Mutation and the Branching Structure of Evolutionary Trees

Forrest W. Crawford, *Yale University*
Willem H. Mulder, *University of West Indies, Mona*
Ignacio Quintero, *Yale University*

Statistical methods for phylogenetic reconstruction and evolutionary inference rely on simple Markov models of speciation, DNA sequence mutation, continuous trait evolution, and geographic movement over evolutionary timescales. Speciation is modeled as a stochastic branching process that gives rise to a phylogenetic tree. A second stochastic process representing trait, sequence, or location change then evolves over the branches of this tree, producing dependencies in the observations at the tips. In this talk, I will outline recent results on Markov models for evolutionary processes on random trees. The results give insight into phylogenetic reconstruction and evolutionary inference for applications in experimental design, mutation rate estimation, and phylogeography.

Estimation of Phylogenetic Power for All Symmetric Markov Models of Evolution and Any Number of Taxa

Jeffrey Townsend, *Yale University*
Tony Su, *Yale University*
Christophe Leuenberger, *University of Fribourg, Switzerland*

The advent of big data in phylogenomics has ushered in conflicting big data results that are each bulwarked by the traditional hallmarks of strong support. Resolution of this conflict requires investigation of the relative power of data to address phylogenetic hypotheses. Publicly available sequenced genes and genomes can provide estimates of relevant parameters such as the rate of evolution of each molecular character, and therefore can provide guidance as to how to optimally design a study. We have derived theory that applies such exogenous data to prioritize gene sequencing and taxon sampling, projecting potential to resolve nodes onto the molecular evolutionary or chronological time scale of interest. It weighs the linear accumulation of signal with internode length, versus both the linear accumulation of bias and the nonlinear accumulation of noise with increasing length of subtending branches of the phylogenetic tree. We have relaxed previous model assumptions of four-taxon trees, deriving theory for all symmetric models of molecular evolution applied to any taxon-sampling scheme.

Confidence Interval Estimation Using Approximate Likelihood of Divergence Time for a Coalescent-based Model

Arindam RoyChoudhury, *Columbia University*

We present confidence interval estimation of divergence time using approximate likelihood based on a coalescent model. In this model, we sum the probability of coalescent trees, taking into account the effect of incomplete lineage sorting. Maximum likelihood estimator based on this model has been computed previously; however, a formula for estimator of confidence interval has never been presented for this model. This is because the expression of the likelihood makes this estimation difficult to compute. Our formula can be readily coded into a program. We did not use a simulation or resampling-based approach and therefore our method is fast and less computationally intensive. We demonstrate that our method is much faster and as accurate a simulation-based approach.

Detecting Tracts of Local Ancestry for Admixture Disease Mapping

Matthias Steinrücken, *University of Massachusetts - Amherst*

The complex demographic history of modern humans has had a substantial impact on the genetic variation we observe today. Due to the process of chromosomal recombination the genomes of contemporary individuals can be mosaics comprised of different DNA segments originating from diverged subpopulations. This is of particular interest when studying variation related to genetic diseases. On the one hand, one has to account for neutral background variation resulting from the demographic history, but on the other hand, knowledge about the distribution of these ancestry segments can also be used to identify causal variants.

In this talk, I present a new method to detect tracts of local ancestry in genomic sequence data of modern humans, and demonstrate its accuracy and efficiency on simulated data. Explicitly modeling the underlying demographic history allows detection under very general scenarios. I will discuss extensions of the method and potential applications using the local ancestry information to foster the detection of functional genetic variation. The distribution of these tracts can also be used to infer features of the demographic history.

Assessing the Probability That a Finding Is Genuine for Large-Scale Genetic Association Studies

Chia-Ling Kuo, *UConn Health Center*

In large-scale genetic association studies, typically a univariate test is applied to each of the variants, the results are sorted by P-value, and the top hits that pass a threshold are identified statistically significant. Focusing on significant P-values provides a filtering mechanism for spurious signals. However, the significance threshold has nothing to do with the odds that the findings are genuine. It is now appreciated that a statistically significant P-value does not imply a high chance that a finding is genuine. The probability in fact depends on the parameters such as the sample size and the effect size distribution. Bayesian methods have been developed for assessing the probability, but they typically require considerable subjective input for the prior distributions. In contrast, we propose a simple, but accurate method that works off P-values and only requires specification of one or few typical effect size values and their abundance.

Approximating the Distributions of Optimal Goodness-of-Fit Tests with Applications in GWAS

Zheyang Wu, *Worcester Polytechnic Institute*

Goodness-of-fit tests (GOFs) are used to test against joint null hypothesis in broad applications such as meta-analysis and signal detection. Among these, Higher Criticism test (HC), Berk-Jones test (B-J), and some other phi-divergence tests have been proved asymptotically optimal for detecting weak and sparse signals. Such property is very attractive in the practice of big data analysis, for example, in detecting subtle disease genes by large-scale genetic and genomic data. However, it is challenging to calculate the p-values and statistical power of GOFs. We provide an analytical solution that can well approximate the entire distributions of a broad family of goodness-of-fit tests. It allows calculating p-values as well as statistical power for arbitrary null and alternative hypotheses, as long as they are continuous. This technique also helps the study of the finite-sample performances for tests that are all asymptotically optimal. The application of GOFs and the p-value calculation technique are illustrated in a real genome-wide association studies (GWAS) of Crohn's disease.

Session 18: Data Analytics at IBM Research

Statistical Challenges for a Crowd-Sourcing-Based Delivery Platform of Software Development

Ta-Hsin Li, *IBM T. J. Watson Research Center*

An emerging business model in application software development in large enterprises is to employ a flexible workforce, or a resource pool, which consists of vetted freelancers, to support the application development process, including software design, coding, and application testing. The success of this model depends crucially on having the right participants at the right time when their skills are needed. However, the need for each set of skills fluctuates over time, depending on the software development activities of the business; the number of participants also fluctuates because participation is entirely voluntary and performed via self-selection of work. Therefore, maintaining the appropriate capacity, or supply, of the resource pool is an important and challenging problem for the service provider who utilizes this type of delivery platform. This talk will discuss statistical challenges in supply predictive analytics for a resource pool operation in IBM's Global Business Services Division.

Robust Compressed Least Squares

Ban Kawas, *IBM T. J. Watson Research Center*
Stephen Becker, *University of Colorado at Boulder*
Marek Petrik, *IBM T. J. Watson Research Center*
Karthikeyan N. Ramamurthy, *IBM T. J. Watson Research Center*

Randomized matrix compression techniques such as the Johnson-Lindenstrauss transform have emerged as an effective method for accelerating classical numerical linear algebra tasks, such as solving least squares problems. Yet, the aggressive compression introduces noise in the problem solution which can significantly degrade its quality. In this paper, we propose mechanisms for improving the solution quality of matrix

sketching for least squares problems. We introduce tools from robust optimization together with a form of partial compression to improve the error-time tradeoff of sketching-based compressed least squares solvers. We develop an efficient algorithm to solve the resulting robust optimization formulation. Empirical results comparing numerous alternatives suggest that aggressive randomized transforms are effectively insulated with partial compression and robustification.

Co-Clustering Structural Temporal Data with Applications to Semiconductor Manufacturing

Yada Zhu, *IBM T. J. Watson Research Center*
Jingrui He, *Arizona State University*

Recent years have witnessed data explosion in semiconductor manufacturing due to advances in instrumentation and storage techniques. The large amount of data associated with process variables monitored over time form a rich reservoir of information, which can be used for a variety of purposes, such as anomaly detection, quality control and fault diagnostics. In particular, following the same recipe for a certain IC device, multiple tools and chambers can be deployed for the production of this device, during which multiple time series can be collected, such as temperature, impedance, gas flow, electric bias, etc. These time series naturally fit into a two-dimensional array (matrix), i.e., each element in this array corresponds to a time series for one process variable from one chamber. To leverage the rich structural information in such temporal data, in this paper, we propose a novel framework named C-Struts to simultaneously cluster on the two dimensions of this array. In this framework, we interpret the structural information as a set of constraints on the cluster membership, introduce an auxiliary probability distribution accordingly, and design an iterative algorithm to assign each time series to a certain cluster on each dimension. To the best of our knowledge, we are the first to address this problem. Extensive experiments on synthetic, benchmark, as well as manufacturing data sets demonstrate the effectiveness of the proposed method.

Process Monitoring Techniques Based on Likelihood Ratios

Emmanuel Yashchin, *IBM T. J. Watson Research Center*

Traditional approaches to Statistical Process Control (SPC) rely on sets of rules (such as Western Electric or Juran rules) to establish what processes should be declared Out of Control. These techniques are based on the concept of statistical significance and the constituent rules correspond to various statistical tests of a simple null hypothesis H0 against the alternative ”not H0”. In todays data-rich environments there is a need for an alternative methodology focused on detecting unfavorable deviations from in-control conditions that are of practical, rather than statistical, significance. In this talk we discuss approaches to monitoring based on various types of likelihood ratio tests and issues related to their implementation in practice.

Predicting Demand for Optimal Opportunity Pipeline Management

Beatriz E Etchegaray Garcia, *IBM T. J. Watson Research Center*

Sales transaction support centers reduce the amount of transaction work performed by sellers allowing them to spend more time with customers and increase revenue generation. Adequate staffing ensures that staff with appropriate skills is available to support incoming service requests. We describe a suite of analytical tools including hierarchical forecasting of demand to determine staffing requirements. These tools were deployed across multiple geographies and lines of business.

Session 19: Probability and Related Topics – in memory of Evarist Giné

Evarist as Student, Teacher, and Friend

Richard M. Dudley, *Massachusetts Institute of Technology*

Originally from Catalonia, Evarist arrived at MIT as a graduate student in 1970. He taught me from early on distinctions between Catalan and Spanish.

He finished his Ph.D. thesis (on testing for uniformity in compact Riemannian manifolds) in 1973 and published it in Annals of Statistics in 1975. After he took a job at Storrs much later, we began to have joint seminars at different campuses in southern New England and still later, mainly alternating between MIT and Storrs. Besides mathematics, I'll mention a hike up Mt Monadnock in May 2007 by several friends from the seminar where we saw a lot of a remarkable flower, *Rhodora*.

Among the topics of joint interest later on were: central limit theorems in Banach spaces and the bootstrap.

Dependence Measures: A Perspective

Ying Liu, *Google, Inc.*
Victor de la Pena, *Columbia University*
Tian Zheng, *Columbia University*

In recent years there has been an increasing interest in the development of new measures of dependence. In this talk I will provide an overview of some of these results including work developed using copulas as well as the distance covariance. Finally, I will introduce a general framework that includes several of the known dependence measures.

Evarist's Favorite Undergraduate Proof and Where It Got Me

Iddo Ben-Ari, *University of Connecticut*

Although we never worked on a joint project, I was very fortunate to spend many enjoyable and memorable hours with Evarist, during which we discussed probability (among other things). One time, several years ago, Evarist enthusiastically told me about his favorite proof in undergraduate probability. I vividly remember this conversation. In this talk I'll reminisce about this conversation and how it got me to start working on a new major line of research.

On the Sup-norm Behavior of the Bernstein Density Estimator

Lu Lu, *Colby College*

The well-known kernel density estimator has serious boundary bias when the underlying density has compact support (e.g., $[0, 1]$). The estimator based on Bernstein polynomials is a promising alternative in this case. In this talk, we will consider its stochastic error under the sup norm. The results imply that the Bernstein density estimator does not attain optimal rate of convergence under the sup norm. This is part of the PhD thesis under the guidance of Prof. Evarist Giné.

Abstracts of Posters

Session 1: Boehringer Ingelheim and Travelers Sponsored Poster Session I

Pathoscope 2.0: Statistical and Computational Methods for Accurate Characterization Of Microbes in Sequencing Samples

Solaiappan Manimaran, *Boston University*
Evan Johnson, *Boston University*

The rapid identification and quantification of pathogens present in a clinical sample is of high importance in controlling contagious diseases during an outbreak. For example, during the European E. coli outbreak of 2011, there was a 3 week delay in the correct identification of the pathogen strain O104:H4 which caused 3,800 infections and 54 deaths. Here, we present Pathoscope 2.0, a complete software package for rapidly identifying and quantifying the microbial strains present in environmental or clinical sequencing samples. Pathoscope utilizes a Bayesian statistical methodology based on a penalized mixture modeling approach to accurately identify and quantify the pathogens. We also present a confidence region for the identified pathogens so that accurate diagnosis and the best possible treatment can be provided. We simulated sequencing reads from 25 strains of bacteria that are commonly found in humans. Our method was able to accurately identify and quantify the pathogen strain both in pure samples with single strain and in mixture samples with multiple strains of bacteria. Our method performed well even with low read coverage and in mixture samples with multiple closely related strains. We also tested on samples from fecal specimens obtained from the 2011 outbreak of Shiga-toxigenic Escherichia coli (STEC) O104:H4, where a study was done previously, and again Pathoscope 2.0 performed better than other approaches such as RINS and ReadScan.

In-stream Text Mining of Patient Narratives: Sweeping for Missed Endpoints

Greg Cicconetti, *GlaxoSmithKline*
Shani Sampson, *GlaxoSmithKline*
David Wade,

During the course of a large cardiovascular program roughly 35k patient narratives were collected. Among these are those describing events associated with the primary endpoint (a composite of myocardial infarction, stroke and cardiovascular death), non-cardiovascular deaths, limb amputations, and other secondary endpoints of interest. The task of bolstering study integrity in anticipation of regulatory submission motivated this exercise. We wish to sweep through the full corpus of narratives and identifying for query those which should have links elsewhere in the electronic case report forms (eCRF). E.g., a narrative paragraph that describes a subject's fatal myocardial infarction needs to have specific information captured elsewhere if statistical analysis is to include that event. Without loss of generality, assume narratives can be partitioned into two sets: Fatal and Non-fatal. The task at hand: Given a collection of narratives known (or presumed known) to be true examples of fatal and non-fatal narratives, deploy an algorithm that ranks narratives based on the likelihood of being "Fatal". Narratives sorted on this metric are then passed to clinical colleagues for review, who then make decision whether or not to raise queries with participating sites to correct database as appropriate. Facets of this exercise to be described:

Challenges in pre-processing data

Algorithm considerations (deployed algorithm: support vector machine)

Reporting provided for clinical review

In-stream logistical considerations

Estimation of the number of missed queries due to curtailment of clinical review

Analysis of Measurement Errors in Evaluating the Effectiveness of Graduated Driver's Licensing Program

Yang Liu, *University of Connecticut*
Peng Zhang, *Zhejiang University*
Juxin Liu,

Accidental deaths from car crashes claim more teenager lives than any other causes in the United States. Graduated Driver's Licensing (GDL) is one of the effective policies adopted in North America to reduce the occurrences of fatal car accidents for teenage-drivers. Many studies have reported the effectiveness of GDL in the US with temporal models and recently by also considering spatial variations. This article studies the effectiveness of GDL with spatial random effects models fitting on data from the state of Michigan collected from databases publicly available. When a multiplicative measurement error term, the rate of teenage-drivers in the teenage population in Michigan, is integrated in the model, the effectiveness of GDL is totally absorbed in that the reduction of fatal car accidents after the implementation of GDL is completely explained by the drop of teenage driver rates in the same period of time. Hence we are able to discover the mechanism in how GDL policies take effect. Identifying the holding rates of teenage-drivers license as a underlying factor for risk reduction, this article provides a new perspective for further policy refining.

Statistical Power as a Function of Internal Consistency of Instrument Questionnaire Items

Moonseong Heo, *Albert Einstein College of Medicine*
Namhee Kim, *Albert Einstein College of Medicine*
Myles Faith, *University of North Carolina–Chapel Hill*

Background: In countless number of clinical trials, measurements of outcomes rely on instrument questionnaire items which however often suffer measurement error problems which in turn affect statistical power of study designs. The Cronbach alpha or coefficient alpha, here denoted by C_α, is a measure of internal consistency of instrument items that are developed to measure a target outcome construct. Scale score for the target construct is often represented by the sum of the item responses. However, power functions based on C_α have been lacking for various study designs. Methods: We formulate a statistical model to derive power functions as a function of C_α under several study designs. To this end, we assume fixed construct variance assumption as opposed to usual fixed total variance assumption. That assumption is critical and practically relevant to show that smaller measurement errors are inversely associated with higher inter-item correlations, and thus that greater C_α is associated with greater statistical power. We compare the derived theoretical statistical power with empirical power obtained through Monte Carlo simulations for the following comparisons: one-sample comparison of pre- and post-treatment mean differences, two-sample comparison of pre-post mean differences between groups, and two-sample comparison of mean differences between groups. Results: It is shown that C_α is the same as a test-retest correlation of the scale scores, which enables testing significance of C_α. Closed-form power functions and samples size determination formulas are derived in terms of C_α, for all of the aforementioned comparisons. Power functions are shown to be an increasing function of C_α, regardless of comparison of interest. The derived power functions are well validated based on simulation studies that show that the magnitudes of theoretical power are virtually identical to those of the empirical power. Conclusion: Regardless of research designs or settings, in order to increase statistical power, development and use of instruments with greater C_α, or equivalently with greater inter-item correlations, is crucial for trials that intend to use questionnaire items for measuring research outcomes.

Student Growth Percentiles in the Presence of Measurement Error

Eugene Quinn, *Stonehill College*
Jeanette Hogan, *Stonehill College*
Colleen McLaughlin, *Stonehill College*

Since its acceptance by the U.S. Department of Education as a means of demonstrating student progress, the Student Growth Percentile (SGP) measure has been adopted by many states for this and other purposes including evaluation of teachers and schools. We examine the behavior of the SGP in the presence of measurement error using the Item Response Theory model of the Massachusetts Comprehensive Assessment System mathematics test as the data generating process to simulate repeatedly administering the MCAS to each student in a grade-sized cohort (70,000 students). Our results indicate that across replications for a specific student, the SGP measure is approximately uniformly distributed over its range.

Analyzing Single-Molecule Protein Transportation Experiments via Hierarchical Hidden Markov Models

Yang Chen, *Harvard University*
Samuel Kou, *Harvard University*

To maintain proper cellular functions, over 50% of proteins encoded in the genome need to be transported to cellular membranes. The molecular mechanism behind such a process, often referred to as protein targeting, is not well understood. Single-molecule experiments are designed to unveil the detailed mechanisms and reveal the functions of different molecular machineries involved in the process. The experimental data consist of hundreds of stochastic time traces from the fluorescence recordings of the experimental system. We introduce a Bayesian hierarchical model on top of hidden Markov models (HMM) to rigorously analyze these data and use the statistical results to answer the biological questions. In addition to resolving the biological puzzles and delineating the regulating roles of different molecular complexes, our statistical results enable us to propose a more detailed mechanism for the late stages of the protein targeting process.

ISAP-MATLAB Package for Sensitivity Analysis of High-Dimensional Stochastic Chemical Networks

Weilong Hu, *University of Massachusetts - Amherst*
Yannis Pantazis, *University of Massachusetts - Amherst*
Markos Katsoulakis, *University of Massachusetts - Amherst*

Stochastic simulation and modeling play an important role to elucidate the fundamental mechanisms in complex biochemical networks. The parametric sensitivity analysis of reaction networks becomes a powerful mathematical and computational tool, yielding information regarding the robustness and the identifiability of model parameters. However, due to overwhelming computational cost, parametric sensitivity analysis is a extremely challenging problem for stochastic models with a high-dimensional parameter space and for which existing approaches are very slow. Here we present an information-theoretic sensitivity analysis in path-space (ISAP) MATLAB package that simulates stochastic processes with various algorithms and most importantly implements a gradient-free approach to quantify the parameter sensitivities of stochastic chemical reaction network dynamics using the pathwise Fisher information matrix (PFIM). The sparse, block-diagonal structure of PFIM makes its computational complexity scale linearly with the number of model parameters. As a result of the gradient-free and the sparse nature of the PFIM, it is highly suitable for the sensitivity analysis of stochastic reaction networks with a very large number of model parameters, which are typical in the modeling and simulation of complex biochemical phenomena. Finally, the PFIM provides a fast sensitivity

screening method which allows it to be combined with any existing sensitivity analysis software.

Order Restricted Inference in Regression

Thelge (Buddika) Peiris, *Worcester Polytechnic Institute*

Abstract Regression analysis constitutes a large portion of the statistical repertoire in applications. In case where such analysis is used for exploratory purposes with no previous knowledge of the structure one would not wish to impose any constraints on the problem. But in many applications we are interested in a simple parametric model to describe the structure of a system with some prior knowledge of the structure. An important example of this occurs when the experimenter has the strong belief that the regression function changes monotonically in some or all of the predictor variables in a region of interest. The analyses needed for statistical inference under such constraints are nonstandard. The specific aim of this study is to introduce a technique which can be used for statistical inferences of a multivariate simple regression with some non-standard constraints.

A Unified Theory of Testing and Confidence Regions for High Dimensional Estimating Equations

Matey Neykov, *Harvard University*
Yang Ning, *Princeton University*
Jun S. Liu, *Harvard University*
Han Liu, *Princeton University*

We propose a new inferential framework of testing hypotheses and constructing confidence regions for high dimensional statistical models that can be fitted by solving a system of regularized estimating equations. Such an estimating equation based inferential framework is quite general and can be used for a wide variety of regularized estimators, including penalized M-estimators, constrained Z-estimators, and even greedy estimators. The key ingredient of this framework is a test statistic constructed by projecting the fitted estimating equations to a sparse direction obtained by solving a large-scale linear program. For hypothesis tests, we derive the limiting distribution of this proposed test statistic under both null and local alternative hypotheses. For confidence regions, we develop uniformly valid confidence intervals for low dimensional parameters of interest, and show their optimality under scenarios when the estimating equation is based on a log-likelihood function. To illustrate the usefulness of this framework, we further apply it to conduct inference for several constrained Z-estimators which have not been equipped with inferential power before, including the Dantzig selector for high dimensional regression, the LDP estimator for high dimensional discriminant analysis, the CLIME estimator for high dimensional graphical models, and a regularized transition matrix estimator for high dimensional vector autoregressive models. Compared with existing methods, our framework is the only one that is applicable for the latter three applications. We provide thorough numerical simulations and real data experiments to back up the developed theoretical results.

Sparse Kernel Machine Regression for Ordinal Outcomes

Yuanyuan Shen, *Harvard University*
Katherine Liao, *Brigham and Women's Hospital*
Tianxi Cai, *Harvard University*

Ordinal outcomes arise frequently in clinical studies when each subject is assigned to a category and the categories have a natural order. Classification rules for ordinal outcomes may be developed with commonly used regression models such as the full continuation ratio (CR) model (fCR), which allows the covariate

effects to differ across all continuation ratios, and the CR model with a proportional odds structure (pCR), which assumes the covariate effects to be constant across all continuation ratios. For settings where the covariate effects differ between some continuation ratios but not all, fitting either fCR or pCR may lead to suboptimal prediction performance. In addition, these standard models do not allow for non-linear covariate effects. In this paper, we propose a sparse CR kernel machine (KM) regression method for ordinal outcomes where we use the KM framework to incorporate non-linearity and impose sparsity on the overall differences between the covariate effects of continuation ratios to control for overfitting. In addition, we provide data driven rule to select an optimal kernel to maximize the prediction accuracy. Simulation results show that our proposed procedures perform well under both linear and non-linear settings, especially when the true underlying model is in-between fCR and pCR models. We apply our procedures to develop a prediction model for levels of anti-CCP among rheumatoid arthritis patients and demonstrate the advantage of our method over other commonly used methods.

Canonical Variate Regression

Chongliang Luo, *Department of Statistics, University of Connecticut*
Jin Liu, *Centre for Quantitative Medicine, Duke-NUS Graduate Medical School*
Dipak Dey, *Department of Statistics, University of Connecticut*
Kun Chen, *Department of Statistics, University of Connecticut*

In many fields, multi-view datasets, measuring multiple distinct but interrelated sets of characteristics on the same set of subjects, together with data on certain outcomes or phenotypes, are routinely collected. For example, in cancer study, microscopic organ tissue measurements, genetic profiles, and organ function test results may all be available. The objective in such a problem is often twofold: both to explore the association structures of multiple sets of measurements and to develop a parsimonious model for predicting the future outcomes. We study a unified canonical variate regression framework to tackle the two problems simultaneously, allowing them to flexibly borrow strength from each other and hence reinforce each other. The proposed criterion integrates multiple canonical correlation analysis with predictive modeling, balancing between the association strength of the canonical variates and their joint predictive power on the outcomes. Moreover, the proposed criterion seeks multiple sets of canonical variates simultaneously to enable the examination of their joint effects on the outcomes, and is able to handle multivariate and non-Gaussian outcomes through a general loss function formulation. An efficient algorithm based on variable splitting and Lagrangian multipliers is developed. Simulation studies show the superior performance of the proposed approach compared to existing alternative methods. We demonstrate the effectiveness of the proposed approach in an F2 intercross mice study and an alcohol dependence study.

Penalized Variable Selection in Competing Risks Regression

Zhixuan Fu, *Yale University*
Chirag R. Parikh, *Yale University*
Bingqing Zhou, *Yale University*

The penalized variable selection methods have been extensively studied for standard time-to-event data. Such methods, cannot be directly applied when subjects are at risk of several mutually exclusive events, known as competing risks. The proportional subdistribution hazard (PSH) model proposed by Fine and Gray has become a popular semi-parametric model for time-to-event data with competing risks. It allows for direct assessment of covariate effects on the cumulative incidence function. In this paper, we propose a general penalized variable selection strategy that simultaneously handles variable selection and parameter estimation in the PSH model. We rigorously establish the asymptotic properties for the proposed penalized estimators and present a numerical algorithm for implementing the variable selection procedure. Simulation studies are conducted to demonstrate the good performance of the proposed method. Diseased donor kidney transplant data from the United Network of Organ Sharing illustrate the utility of the proposed method.

Information-Theoretic Characterization of Short-Memory and Long-Memory Gaussian Processes

Gordon Chavez, *New York University*

We study the amount of available information about the future state of a Gaussian stochastic process. Using the asymptotic expansion for a long-memory autocovariance function found by Lieberman and Phillips (2008), we derive an expression for the mutual information between states as a function of lag. For stationary Gaussian processes, we derive an expression for the mutual information between two states, conditioned on the earlier state's infinite past, as a function of the moving average parameters. This gives an explicit connection between available information for prediction and the moving average parameters. We apply the above formulas to the study of the ARFIMA(p,d,q) process. We show revealing, information-theoretic characterizations of the transition from stationary long-memory to nonstationarity, and from long-memory to anti-persistence, as well as describing a regime where the mutual information is not summable. For comparison, we derive the same quantities for the short-memory, Markovian, AR(1) process. We describe an application of the above results to the optimization and evaluation of predictive models.

Power Evaluation of Methods for Detecting Non-Independent Relationships

Ruobin Gong, *Harvard University*
Bo Jiang, *Harvard University*
Jun Liu, *Harvard University*
Edoardo Airoldi, *Harvard University*

Automated or semi-automated dependence detection methods are widely employed in the exploratory analysis of large, high-dimensional datasets. They usually require little tuning from the user and aim to provide a one-stop solution. Without substantial care to design and modeling assumptions however, these methods may suffer from low statistical power, and blindly applying them may result in inefficient use of data. Seeing a direct correspondence between dependence detection and statistical hypothesis testing, we propose an assessment framework on the efficacy of large-scale inference methods of distinct rationales, utilizing Monte Carlo-based methods to numerically evaluate the statistical power. We present two simulation experiments to compare competing sets of methods designed for nonlinear functional relationship detection, including distance correlation (DCor), Maximal Information Coefficient (MIC), and Dynamic Slicing for sample variances (DSV). We observe that DCor is superior at detecting smooth and monotone relationships, and is DSV superior at spiking, periodic and oscillating relationships. MIC is dominated by either DCor or DSV in statistical power for all functional types tested, and is less robust to the corruption of heavy-tailed noise. From a practitioner's standpoint, it is the best to incorporate subject-matter scientific knowledge when choosing a suitable metric and constructing decision rules. We hope to see a convincing statistical power analysis to accompany all dependence detection methods proposed in order to illustrate its strengths and weaknesses, facilitating better-informed decisions for scientific discoveries.

The Optimised Theta Method

Jose Augusto Fioruci, *Federal University of Sao Carlos / University of Connecticut*
Tiago Pellegrini, *University of New Brunswick*
Francisco Louzada, *University of Sao Paulo*
Fotios Petropoulos, *Cardiff Business School*

Accurate and robust forecasting methods for univariate time series are very important when the objective is to produce estimates for a large number of time series. In this context, the Theta method called researchers attention due its performance in the largest up-to-date forecasting competition, the M3-Competition. Theta method proposes the decomposition of the deseasonalised data into two ”theta lines”. The first theta line removes completely the curvatures of the data, thus being a good estimator of the long-term trend component. The second theta line doubles the curvatures of the series, as to better approximate the short-term behaviour. In this work, we propose a generalisation of the Theta method by optimising the selection of the second theta line, based on various validation schemes where the out-of-sample accuracy of the candidate variants is measured. The recomposition process of the original time series builds on the asymmetry of the decomposed theta lines. An empirical investigation through the M3-Competition data set shows improvements on the forecasting accuracy of the proposed optimised Theta method.

Negative-Binomial Cure Rate Models with Spatial Frailties for Interval-Censored Data

Yiqi Bao, *Federal University of Sao Carlos / University of Connecticut*
Vicente Cancho, *University of Sao Paulo*
Dipak Dey, *University of Connecticut*

In this work, we extend a flexible cure rate model to allow for spatial correlations by including spatial frailties for the interval censored data setting. The parametric and semi-parametric cure rate models with the independent and dependent spatial frailties are proposed and compared. The proposed models encompass several known cure rate models as its particular cases. Moreover, since these cure models are obtained by considering the occurrence of an event of interest is caused by the present of any non-observed risks, we also studies the complementary cure model, that is, the cure models are obtained by assuming the occurrence of an event of interest is caused when all of non-observed risks are activated. The MCMC method is used in Bayesian inference approach and some Bayesian criterions are used for a comparison. Furthermore, we also conduct the influence diagnostic through the diagnostic measures in order to detect possible influential or extreme observations that can cause distortions on the results of the analysis. Finally, the proposed models are applies to analysis a real data set.

Hierarchical Bayesian Models, Small Area Estimation and Dirichlet Processes

Jiani Yin, *Worcester Polytechnic Institute*

We propose nonparametric hierarchical Bayesian models for multi-stage finite population sampling. Bayesian predictive inference for small area estimation is studied by embedding a parametric model in a nonparametric model. Survey data tend to have gaps, ties and outliers. Therefore, it is sensible to robustify Bayesian predictive inference. We exemplify by considering in detail two-stage and three-stage hierarchical Bayesian models with Dirichlet processes at various stages. Moreover we conduct model comparison by computing log pseudo marginal likelihood and Bayes factors; this is problematic however. The computational difficulties of the predictive inference when the population size is much larger than the sample size can be overcome by using the stick-breaking algorithm. We use simulated data and body mass index data to compare the performance of our nonparametric Bayesian models to the Bayesian baseline parametric models.

Stable and Optimal: Implicit Stochastic Gradient Descent with Averaging

Dustin Tran, *Harvard University*
Panos Toulis, *Harvard University*
Edoardo Airoldi, *Harvard University*

Stochastic gradient methods have increasingly become popular for large-scale optimization. However, they are often numerically unstable and statistically inefficient because of their sensitivity to the problem convexity and learning rate specification. We propose a new learning procedure, termed averaged implicit stochastic gradient descent (ai-SGD), which combines stability through proximal (implicit) updates and statistical efficiency through averaging of the iterates. In an asymptotic analysis we prove convergence of the procedure and show that it is statistically optimal, i.e., it achieves the Cramer-Rao lower variance bound. In a non-asymptotic analysis, we show that the apparent stability of ai-SGD is due to its robustness to misspecifications of the learning rate with respect to the convexity of the loss function. Our experiments demonstrate that ai-SGD performs on par with state-of-the-art learning methods. Furthermore, ai-SGD is more stable than averaging methods that do not utilize proximal updates, and it is simpler and computationally more efficient than methods that do employ proximal updates in an incremental fashion.

Cluster Forests

Donghui Yan, *University of Massachusetts - Dartmouth*

“Ensemble to achieve the best performance” has become the folklore in statistics, machine learning and data mining. Enormous success has been achieved in the context of classification or regression, for example the winning entry for the Netflix million dollars challenge is an ensemble of 107 models, and Random Forests (RF), one of the most successful inference tools in statistics is an ensemble of decision trees. The progress is slow, however, in the setting of clustering. Inspired by the success of Random Forests, we propose a clustering ensemble method – Cluster Forests (CF). Geometrically, CF randomly probes a high-dimensional data cloud, from many different angles, to obtain many “good local clusterings” and then aggregates via spectral clustering to obtain cluster assignments for the whole data. The search for a good local clustering is guided by a cluster quality measure kappa. CF progressively improves each local clustering in a fashion that resembles the tree growth in RF. Empirical studies on several real-world data under two different performance metrics show that CF compares favorably to its competitors. Theoretical analysis reveals that the kappa measure allows the growth of each local clustering in a “noise-resistant” fashion. A stochastic block model is studied to understand the behavior of cluster ensemble in CF, and a closed-form formula is obtained for the fast decay of “error rate” under such model. This yields new insights into the ensemble mechanism in CF as well as the success of spectral clustering.

Session 2: Boehringer Ingelheim and Travelers Sponsored Poster Session II

Consistent Estimation for Multi-Graph Stochastic Block Models, with Applications to Dynamic and Multi-Layer Networks

Qiuyi Han, *Harvard University*
Kevin Xu, *Technicolor*
Edoardo Airoldi, *Harvard University*

Significant progress has been made recently on theoretical analysis of estimators for the stochastic block model (SBM). In this paper, we consider the multi-graph SBM, which serves as a foundation for many

application settings including dynamic and multi-layer networks. We explore the asymptotic properties of two estimators for the multi-graph SBM, namely spectral clustering and the maximum-likelihood estimate (MLE), as the number of layers of the multi-graph increases. We derive sufficient conditions for consistency of both estimators and propose a variational approximation to the MLE that is computationally feasible for large networks. We verify the sufficient conditions via simulation and demonstrate that they are practical. In addition, we apply the model to two real data sets: a dynamic social network and a multi- layer social network with several types of relations.

Implicit Stochastic Gradient Descent for Principled Estimation with Large Datasets

Panos Toulis, *Harvard University*

Efficient optimization procedures, such as stochastic gradient descent, have been gaining popularity for estimation tasks with large amounts of data. Typically these procedures are iterative and use *explicit* iterates, where the update at every iteration depends on the previous iterate and one observed datapoint. In this paper, we introduce an *implicit* stochastic gradient descent estimation procedure that uses iterates that are implicitly defined. The implicit iterates are *shrinked* versions of explicit iterates, and it can be shown that the amount of shrinkage depends on the observed Fisher information, but this latter quantity needs not be directly computed. The implicit procedure is thus robust to the choice of a scalar hyper-parameter in stochastic gradient descent, known as the learning rate, that affects its asymptotic statistical properties. In contrast, explicit procedures require the learning rate to agree with the eigenvalues of the Fisher information matrix of the underlying model parameters in order to be stable. In the context of generalized linear models, we derive analytic formulas for the asymptotic bias and variance of both procedures as estimation methods, and quantify their efficiency loss compared to maximum likelihood. Our analysis naturally extends to exponential family models, and to a general class of estimation methods through Monte-Carlo stochastic gradient descent, in problems where the likelihood is hard to compute but where it is easy to sample from the underlying model. We demonstrate our theory in an extensive set of experiments involving real and simulated data. Implicit stochastic gradient descent compares favorably to popular estimation methods that use explicit iterates. Additionally, it provides principled estimators that are numerically stable and can be trusted to have the nominal theoretical variance.

Estimation of Discrete Survival Function through the Modeling of Diagnostic Accuracy for Mismeasured Outcome Data

Hee-Koung Joeng, *University of Connecticut*
Abidemi K. Adeniji, *Boehringer Ingelheim Pharmaceuticals, Inc*
Naitee Ting, *Boehringer Ingelheim Pharmaceuticals, Inc*
Ming-Hui Chen, *University of Connecticut*

Standard survival methods are inappropriate for mismeasured outcomes. Previous research has shown that outcome misclassification can bias estimation of the survival function. We develop methods to accurately estimate the survival function when the diagnostic tool used to measure the outcome of disease is not perfectly sensitive and specific. Since the diagnostic tool used to measure disease outcome is not the gold standard, the true or error-free outcomes are latent, they cannot be observed. Our method uses the negative predictive value (NPV) and the positive predictive values (PPV) of the diagnostic tool to construct a bridge between the error-prone outcomes and the true outcomes. We formulate an exact relationship between the true (latent) survival function and the observed (error-prone) survival function as a formulation of time-varying NPV and PPV. We specify models for the NPV and PPV that depend only on parameters that can be easily estimated from a fraction of the observed data. Furthermore, we conduct an in depth study to accurately estimate the latent survival function based on the assumption that the biology that underlies the disease

process follows a stochastic process. We further examine the performance of our method by applying it to the VIRAHEP-C data.

Causal Inference with Social Interference

Edward Kao, *Harvard University*
Panos Toulis, *Harvard University*
Edoardo Airoldi, *Harvard University*

The broad adoption of social media has led to new opportunities in understanding and leveraging peer influence. However, in order to quantify the causal effect of peer influence, we need to overcome some new challenges including: interference and dependencies between units in a social network, complex functions of unit response to treatments on the network, and network uncertainty. Here, we introduce novel causal estimands of peer-influence under a principled extension of the potential outcomes framework for causal inference, and develop two procedures to estimate them: a sampling procedure that requires knowledge of the network, but which can operate under complicated (e.g., non-linear) response, and a Bayesian procedure that accounts for social network uncertainty, but relies on a linear model for unit response. We are able to characterize bias-information trade-offs. In the sampling procedure, randomizations that identify more information about the causal estimand may lead to more biased estimates, depending on the network topology that is being conditioned on. In contrast, the Bayesian procedure utilizes information on the network topology to achieve more precise estimates; however, bias is introduced when the linearity assumption does not hold. We experiment on various popular network topologies and unit response functions, which highlights the comparative advantages of the two aforementioned procedures.

Bayesian Or's of And's for Interpretable Classification, with Application to Context-Aware Recommender Systems

Tong Wang, *massachusetts institute of technology*
Cynthia Rudin, *massachusetts institute of technology*
Finale Doshi-Velez, *Harvard University*
Yimin Liu, *Ford Motor Company*
Perry MacNeille, *Ford Motor Company*
Erica Klampfl, *eklampfl@ford.com*

We present a machine learning algorithm for building classifiers that are comprised of a small number of disjunctions of conjunctions (or's of and's). An example of a classifier of this form is as follows: If X satisfies (conditions A1, A2, and A3) OR (conditions B1 and B2) OR (conditions C1 and C2), then we predict that Y=1, ELSE predict Y=0. Models of this form have the advantage of being interpretable to human experts, since they produce a set of conditions that concisely describe a specific class. We present two probabilistic models for forming a pattern set, one with a Beta-Binomial prior, and the other with Poisson priors. In both cases, there are prior parameters that the user can set to encourage the model to have a desired size and shape, to conform with a domain-specific definition of interpretability. We provide two methods for finding maximum a posteriori solutions: a pattern level search, which involves association rule mining, and a literal-based search, both of which scale nicely. We provide theoretical bounds that motivate our optimization techniques, showing how stronger priors provably aid computation. We apply the BOA model to predict user behavior with respect to in-vehicle context-aware personalized recommender systems.

Marginal Likelihood Estimation of Variable Tree Topology in Phylogenetics

Daoyuan Shi, *University of Connecticut*

The marginal likelihood is commonly used for comparing different evolutionary models in Bayesian phylogenetics. But how to estimate the marginal likelihood of the variable tree topology is still challenging. We apply Larget's idea to Chib's method and then develop a new general method Weighted Chib's method. It uses more than one tree topology to estimate the variable tree topology marginal likelihood and gives more efficient result. We also compare our methods with Generalized Stepping Stone method for rbcL chloroplast DNA sequences example.

Group-Corrected Stochastic Blockmodels for Community Detection on Large-scale Networks

Lijun Peng, *Boston University*
Luis Carvalho, *Boston University*

Community detection in networks is becoming increasingly important in many applications, especially in social sciences. Today's era of big data proposes a new challenge in this field—how to efficiently detect community structure on large-scale social networks. In this paper, we propose a novel stochastic blockmodel based on a logistic regression setup with group correction terms to better address this problem and conduct exact inference based on a maximum a posteriori (MAP) estimator. We demonstrate the novel proposed model and estimation on large real-world networks as well as simulated benchmark networks, and show that the proposed estimator performs better when compared to the MAP estimator from classical degree-corrected stochastic blockmodels as well as other commonly used estimators suitable for large-scale networks.

The Impact of Missing Values on Different Measures of Uncertainty

Chantal Larose, *University of Connecticut*
Dipak K. Dey, *University of Connecticut*
Ofer Harel, *University of Connecticut*

How a researcher chooses to handle incomplete records can dramatically affect the result of statistical analyses. Differential entropy measures the uncertainty in a model which describes a data set, and common sense tells us this measure should increase when there are incomplete records. However, the effect of missing values on entropy has never been quantified. We utilize multiple imputation to examine the effect of missing values on the entropy of the bivariate normal data model. Results include new theorems to quantify entropy of incomplete data, simulation studies to illustrate the effect of a variety of missingness percentages on entropy, and a comparison of the behavior of entropy of the incomplete data model and the fraction of missing information, a bi-product of multiple imputation.

The Application of Sparse Estimation of Covariance Matrix to Quadratic Discriminant Analysis

Jiehuan Sun, *Yale University*
Hongyu Zhao, *Yale University*

Although Linear Discriminant Analysis (LDA) is commonly used for classification, it may not be directly applied in genomics studies due to the large p, small n problem in these studies. Different versions of sparse LDA have been proposed to address this significant challenge. One implicit assumption of various LDA-based methods is that the covariance matrices are the same across different classes. However, rewiring of genetic networks (therefore different covariance matrices) across different diseases has been observed in many genomics studies, which suggests that LDA and its variations may be suboptimal for disease classifications. However, it is not clear whether considering differing genetic networks across diseases can improve classification in genomics studies. In this manuscript, we proposed a sparse version of Quadratic Discriminant Analysis (SQDA) to explicitly consider the differences of the genetic networks across diseases. Both simulation and real data analysis are performed to compare the performance of SQDA with six commonly used classification methods. In our comparisons, SQDA provides more accurate classification results than other methods for both simulated and real data. Our method should prove useful for classification in genomics studies and other research settings, where covariances differ among classes.

Post-GWAS Prioritization through Integrated Analysis of Genomic Functional Annotation

Qiongshi Lu, *Yale University*
Xinwei Yao, *Yale College*
Yiming Hu, *Yale University*
Hongyu Zhao, *Yale University*

Genome-wide association study (GWAS) has been a great success in the past decade, with tens of thousands of loci identified associated with many complex diseases in humans. However, challenges still remain in both identifying new risk loci and interpreting results. Bonferonni-corrected significance level is a very conservative threshold for high dimensional hypothesis testing, leading to insufficient statistical power when the effect size is moderate at each risk locus. Complex structure of linkage disequilibrium also makes it challenging to distinguish causal variants from large haplotype blocks. In this paper, we propose a post-GWAS prioritization method that integrates genomic functional annotation and GWAS test statistics. We apply our method to GWAS results for Crohns disease and schizophrenia, and test the performance using the largest studies available. After prioritization, highly ranked loci show substantially stronger signals in the testing dataset than the top loci before prioritization. At the single nucleotide polymorphism (SNP) level, SNPs after prioritization have both higher replication rates and consistently stronger enrichment of eQTLs. Within each risk locus, our method is able to distinguish real signal sources from groups of correlated SNPs. Our method has the potential to be widely used to reveal functional spots at disease-associated risk loci and guide further studies such as resequencing analysis.

Bayesian Forest Classifier: Building Tree Structures on Naive Bayes

Viktoriya Krakovna, *Harvard University*
Jiong Du,
TengJiao Wang, *National University of Defense Technology*
Yuan Yuan,
Jun Liu, *Harvard University*

We focus on the high-dimensional classification and variable selection problem with categorical variables. Of particular interest are interactions among the variables that influence the outcome. We propose a novel Bayesian strategy to build a correlation structure on top of the naive Bayes classifier, resulting in an algorithm that provides both accurate classification and variable selection. We introduce a latent variable to partition all the covariates into two groups according to their relationships with the class variable, and conduct Bayesian variable selection and covariance structure modeling simultaneously. In order to achieve both model flexibility and parsimony, we use trees to approximate the dependency relationship among the covariates, and set a complexity-penalizing prior on the tree structure parameters. We use Gibbs sampling to explore the partition and tree structure space, and combine the predictions using Bayesian model averaging. Our method performs competitively with state-of-the-art classifiers, and provides insight into relevant

variables and variable interactions.

Placebo Non-Response Measure in Sequential Parallel Comparison Design Studies

Denis Rybin, *Boston University*

The Sequential Parallel Comparison Design (SPCD) is one of the novel approaches addressing placebo response. The analysis of SPCD data typically classifies subjects as ‘placebo responders’ or ‘placebo non-responders’. Most current methods employed for analysis of SPCD data utilize only a part of the data collected during the trial. A repeated measures model was proposed for analysis of continuous outcomes that permitted the inclusion of information from all subjects into the treatment effect estimation. We describe a new approach using a weighted repeated measures model that further improves the utilization of data collected during the trial, allowing the incorporation of information that is relevant to the placebo response, and dealing with the problem of possible misclassification of subjects. Our simulations show that when compared to the unweighted repeated measures model method, our approach performs as well or, under certain conditions, better, in preserving the type I error, achieving adequate power and minimizing the mean squared error.

Multivariate Temporal Dynamics of Gastropod Abundance in a Puerto Rican Tropical Forest

Volodymyr Serhiyenko, *University of Connecticut*
Nalini Ravishanker, *University of Connecticut*
Michael Willig, *University of Connecticut*
Brian Klingbeil, *University of Connecticut*

In many application areas, there is the necessity for accurate statistical modeling of multivariate counts as a function of relevant covariates. This paper investigates the multivariate Poisson Lognormal (MVPLN) modeling of time series of counts for abundance of gastropod species in Puerto Rico. The proposed multivariate model can account for the overdispersion as well as positive and/or negative association between counts of different species. This provides correct estimates for parameters in predicting counts. The proposed dynamic framework can incorporate time dependence together with location and time specific covariates which help the ecologists to understand the evolution and variation of the species abundance. The data are collected on 40 sites from 1995 to 2014 and on additional 111 sites for 1995, 1996, and 1997. We investigate the predictive capabilities of the model using temporal as well as location specific hold-out observations. We also take advantage of Approximate Bayesian inference via the Integrated Nested Laplace Approximations (INLA) which significantly decreases computational time and makes it attractive for interdisciplinary research.

Onset Time of Chronic Pseudomonas Aeruginosa Infection in Two Cohorts of Cystic Fibrosis Patients with Interval Censored Data

Wenjie Wang, *Department of Statistics, University of Connecticut*
Ming-Hui Chen, *Department of Statistics, University of Connecticut*
Sy Han Chiou, *Department of Mathematics and Statisitcs, University of Minnesota Duluth*
Hui-Chuan Lai, *Department of Nutritional Sciences, University of Wisconsin*
Xiaojing Wang, *Google, Inc.*
Jun Yan, *Department of Statistics, University of Connecticut. Institute for Public Health Research, University of Connecticut Health Center*
Zhumin Zhang, *Department of Nutritional Sciences, University of Wisconsin*

Chronic pseudomonas aeruginosa (PA) infection indicates lung function deterioration in children cystic fi-

brosis (CF) patients. Modeling the onset time of chronic PA infection is important for clinicians to devise better treatment plan and patient management. Due to the scheduled visits for patients in the cystic fibrosis foundation patient registry (CFFPR) and the definition of chronic PA infection, the onset time is only known up to a certain interval. The analysis is further challenged by the need to allow some risk factors to have time-varying effects on the onset time. This problem fits into the framework of Bayesian dynamic Cox model for interval censored data recently developed by Wang, Chen, and Yan (2013, Lifetime Data Analysis 19:297–316). Application of the methodology to the onset time of chronic PA infection of children CF patients revealed interesting findings. Compared with patients diagnosed via new born screening, patients diagnosed via meconium ileus or other symptoms had moderately higher risks of acquiring chronic PA infections before age two. Two cohorts of five years apart were compared, and patients in the more recent cohort were found to have lower risks of chronic PA infection throughout their first five years.

Hidden Population Size Estimation from Respondent-Driven Sampling: A Network Approach

Jiacheng Wu, *Yale University*
Forrest Crawford, *Yale University*
Mait Raag, *University of Tartu*
Robert Heimer, *Yale University*
Anneli UuskÃla, *University of Tartu*

Estimating the size of hidden and hard-to-reach populations such as people who inject drugs, sex workers, or men who have sex with men is important for public health interventions and resource allocation. Multiplier and capture-recapture methods have become standard tools to estimate population size but the requisite assumption of random sampling is rarely satisfied. Respondent-driven sampling (RDS) is a network-based survey method in which subjects recruit their social acquaintances from within the hidden population, and it has become the most widely used sampling method for epidemiological studies of risk behavior in hidden populations.

Current methods for estimating hidden population size using RDS require an additional data source, rely on unrealistic assumptions, or ignore most of the information in the RDS sample. In this work, we derive a method for estimating the size of hidden population that uses all the data typically collected in an RDS study. The recruitment chain, recruitment times, and network degrees of subjects provide information about the number of people in the hidden population who are not yet in the RDS sample. The observed recruitment information imposes topological constraints on the structure of the induced sub-network of sampled individuals. A computationally efficient Bayesian method is used to integrate over induced sub-networks of sampled individuals and calculate the posterior distribution of hidden population size. We apply the technique to estimate the number of people who inject drugs (PWID) in Kohtla-JÃrve, a collection of towns in northeastern Estonia. We estimate that the size of this population is 1082 people (95% posterior CI 1013 - 1165). We compare the results from multiplier method and successive sampling estimate (Gile et al 2015). The multiplier method estimate is 1350 (95% uncertainty region 876 - 2110). Successive sampling gives the estimate 959 (95% posterior CI 638 - 1597).

Selecting the Number of Largest Order Statistics in Extreme Value Analysis

Brian Bader, *University of Connecticut*
Jun Yan, *University of Connecticut*
Xuebin Zhang, *Environment Canada*

The r largest order statistics approach is widely used in extreme value analysis because it may use more information from the data than just the block maxima. In practice, the choice of r is critical. If r is too large, bias can occur; if too small, the variance of the estimator can be high. The limiting distribution of the

r largest order statistics, denoted by GEV$_r$, extends that of the block maxima. Two specification tests are proposed to select r sequentially. The first is a score test for the GEV$_r$ distribution, which is implemented via a fast weighted bootstrap or multiplier procedure. The second test uses the difference in estimated entropy between the GEV$_r$ and GEV$_{r-1}$ models, applied to the r largest order statistics and the $r-1$ largest order statistics, respectively. The asymptotic distribution is derived with the central limit theorem. In a large scale simulation study, both tests held their size and had substantial power to detect various misspecification schemes. The utility of the tests is demonstrated with environmental and financial applications.

Prediction of RiboSNitches

Jianan Lin, *University of Connecticut*
Zhengqing Ouyang, *The Jackson Laboratory*

RiboSNitches are single-nucleotide polymorphisms (SNPs) that alter the secondary structure of regulatory regions of RNAs. These variations of structure are often linked to the effect on phenotypes such as diseases. RNA-folding algorithm-based methods make it possible to predict RiboSNitches but they suffer from limited accuracy when only considering the dissimilarity of the base-pairing probability matrix (BPPM). Therefore we present a method that uses not only the distance score of BPPM, but also the features of the regions where the SNPs happen. Based on genome-wide RNA structure mapping datasets (Wan et al., Nature, 2014) , we trained and tested our model on different subsets of RiboSNitches and found that the introduced region features improve the accuracy of RiboSNitch prediction, which also provide us deeper understanding of the biological condition that makes a SNP a RiboSNitch.

Second Order Correctness of Perturbation Bootstrap M Estimator of Multiple Linear Regression Parameter

Debraj Das, *North Carolina State University*
Soumendra Lahiri, *North Carolina State University*

This work introduces Perturbation Bootstrap in Multiple Linear Regression setup. Residual Bootstrap method can already be found in the literature of Regression. Lahiri S. N. [AOC (1992) 1548-1570] has shown that Residual Bootstrap M-Estimator in multiple linear regression is second order correct. Our motivation behind this work is to show that this new bootstrap procedure in Regression setup is second order correct and in some cases Perturbation Bootstrap is expected to give better result than usual Residual Bootstrap in Regression setup.

Consider the Multiple Linear Regression model :

$$y_i = x_i^{'}\beta + \epsilon_i, \qquad i = 1, 2, \ldots.., n$$

where $\epsilon_1, \ldots.., \epsilon_n$ are independent and identically distributed (iid) random variables, $x_1, \ldots.., x_n$ are known non random design vectors and β is the p x 1 vector of parameters where p is fixed. Suppose $\bar{\beta}_n$ is the M-estimator corresponding to the score function ψ.

Let us introduce the perturbation random factors G_i^*'s where G_i^*'s are non-negative iid completely known random variables. Define, Perturbation Bootstrap M-estimator β_n^* of β to be the solution of

$$\sum_{i=1}^{n} x_i \psi(y_i - x_i'\beta) G_i^* = 0$$

We have obtained two term Edgeworth expansion of standardized and studentized β_n^* and using that we have shown second order optimality. Comparison between Perturbed and Residual Bootstrap to check the betterment of our new bootstrap method requires future investigation.

An Alarm System for Flu Outbreaks Using Google Flu Trend Data

Gregory Vaughan, *Department of Statistics, University of Connecticut*
Robert Aseltine Jr., *Institute for Public Health Research, University of Connecticut Health Center*
Sy Han Chiou, *Department of Mathematics and Statistics, University of Minnesota Duluth*
Jun Yan, *Department of Statistics, University of Connecticut*

Outbreaks of influenza pose a serious threat to communities and hospital resources. It is important for health care providers not only to know the seasonal trend of influenza, but also to be alarmed when unusual outbreaks occur as soon as possible for more efficient, proactive resource allocation. Google Flu Trends data showed a good match in trend patterns, albeit not in exact occurrences, with the hospitalization counts due to influenza from the Center for Disease Control, and, hence, provide a timely, inexpensive data source to develop an alarm system for outbreaks of influenza. For the State of Connecticut, using weekly Google Flu Trends data from 2003 to 2012, an exponentially weighted moving average control chart was developed after removing the trend from the observed data. The control chart was tested with the 2013 data, and was able to issue an alarm at the unusually earlier outbreak in January, 2013.

Achieving Optimal Misclassification Proportion in Stochastic Block Model

Chao Gao, *Department of Statistics, Yale University*
Zongming Ma, *Department of Statistics, University of Pennsylvania*
Anderson Y. Zhang, *Department of Statistics, Yale University*
Harrison Zhou, *Department of Statistics, Yale University*

Community detection is a fundamental statistical problem in network data analysis. Many algorithms have been proposed in literature to tackle this problem. However, these algorithms are not guaranteed to achieve the statistical optimality of the problem. Moreover, procedures that achieve information theoretic lower bounds for very general parameter spaces are not computationally tractable. In this paper, we present a computationally feasible two-stage method that achieves optimal statistical performance in misclassification proportion for stochastic block model under very weak regularity conditions. Our two-stage procedure consists of a generic refinement step that can take a wide range of weakly consistent clustering procedures as initializer, to which the refinement stage applies and outputs clustered nodes with optimal misclassification proportion. The practical effectiveness of the new algorithm is demonstrated by compatible numerical results.

Session 3: Boehringer Ingelheim and Travelers Sponsored Poster Session III

Assessing Covariate Effects with the Monotone Partial Likelihood Using Jeffreys' Prior in the Cox Model

Jing Wu, *University of Connecticut*
Mário de Castro, *Universidade de São Paulo*
Elizabeth D. Schifano, *University of Connecticut*
Ming-Hui Chen, *University of Connecticut*

In clinical trials, the monotone partial likelihood is frequently encountered in the analysis of time-to-event data using the Cox model. When there are zero events in one or more covariate groups, the resulting partial likelihood is monotonic and consequently, the covariate effects are difficult to estimate. In this paper, we develop both Bayesian and frequentist approaches using Jeffreys' prior to handle the monotone partial likelihood problem. We first carry out an in-depth examination of the conditions of the monotone partial likelihood and then characterize sufficient and necessary conditions for the propriety of Jeffreys' prior. We also study several theoretical properties of Jeffreys' prior for the Cox model. In addition, we propose two variations of Jeffreys' prior. An efficient Markov chain Monte Carlo algorithm is developed to carry out posterior computation. We perform extensive simulations to examine the performance of parameter estimates and demonstrate the applicability of the proposed method by analyzing real data from a prostate cancer study in details.

Rate-Optimal Graphon Estimation

Chao Gao, *Yale University*
Yu Lu, *Yale University*
Harrison Zhou, *Yale University*

Network analysis is becoming one of the most active research areas in statistics. Significant advances have been made recently on developing theories, methodologies and algorithms for analyzing networks. However, there has been little fundamental study on optimal estimation. In this paper, we establish optimal rate of convergence for graphon estimation. For the stochastic block model with k clusters, we show that the optimal rate under the mean squared error is $n^{-1}\log k + k^2/n^2$. The minimax upper bound improves the existing results in literature through a technique of solving a quadratic equation. When $k \leq \sqrt{n \log n}$, as the number of the cluster k grows, the minimax rate grows slowly with only a logarithmic order $n^{-1}\log k$. A key step to establish the lower bound is to construct a novel subset of the parameter space and then apply Fano's lemma, from which we see a clear distinction of the nonparametric graphon estimation problem from classical nonparametric regression, due to the lack of identifiability of the order of nodes in exchangeable random graph models. As an immediate application, we consider nonparametric graphon estimation in a Hölder class with smoothness α. When the smoothness $\alpha \geq 1$, the optimal rate of convergence is $n^{-1}\log n$, independent of α, while for $\alpha \in (0,1)$, the rate is $n^{-\frac{2\alpha}{\alpha+1}}$, which is, to our surprise, identical to the classical nonparametric rate.

The Informative g-Prior vs. Common Reference Priors for Binomial Regression in an Application to Hurricane Electrical Utility Asset Damage Prediction

Nathan Lally, *University of Connecticut*
Brian Hartman, *University of Connecticut*

Eliciting appropriate prior information from experts for a statistical model is no easy task. Expressing this information in terms of parameters of prior probability distributions on abstract model parameters can be nearly impossible, especially after a data augmentation/transformation procedure. In previous work on logistic and binomial regression models, Hanson 2014 assert that "experts are confident only in their assessment of the population as a whole" and propose a version of the g-prior which allows the statistician or practitioner to effectively place a standard beta distributed prior on the overall population probability of success. We explore the efficacy of using the informed g-prior in a real prediction problem involving electrical utility asset damages due to hurricanes in Connecticut. Prior information is elicited from a group of engineers at the electrical utility and several methods are used to select hyper-parameters for the g-prior. The out-of-sample predictive accuracy of these informed models is compared to the performance of models constructed under common reference priors (Jeffreys's, Gelman 2008, and a noninformative specification of the g-prior) using IS-LOO, root mean squared error (RMSE), and other statistics. We conclude that in this application with carefully selected hyper-parameters, binomial regression models using the informed g-prior match the predictive accuracy of common reference priors and offer no distinct advantage. Careless selection of hyper-parameters can however, lead to substantial reduction in predictive accuracy. Surprisingly, the noninformative specification of the g-prior performed marginally better than all other models tested in this paper; contradicting one of the findings in Hanson 2014. In addition, we show the predictive accuracy gained by modeling spatial correlation in the residuals and prove that such models substantially outperform a class of statistical learning models growing in popularity in this field.

Statistical Analysis of Gene-expression Networks

Haim Bar, *University of Connecticut*
Seo-Jin Bang, *University of Connecticut*

Gene co-expression network is an undirected graph with nodes corresponding to genes and edges indicating co-expression significance between a pair of genes. The biological interactions between genes are encoded by the co-expression measures such as Pearson's correlation coefficient, Mutual Information, Spearman's rank correlation coefficient and Euclidean distance. Then the encoded networks are used to define clusters and intra-modular hub genes, to explore the relationships between co-expression clusters of highly correlated genes and external sample traits, and to compare the network topology of different biological networks. For example, the weighted correlation network analysis (WGCNA) constructs weighted gene co-expression networks where each weight represents the significance of the co-expression relationship between a pair of genes. The weighted network is specified by the adjacency matrix in which cell (i, j) is a function of correlation coefficients r_{ij} between genes i and j.

We raise a number of concerns about the aforementioned approach. First, the correlation coefficient r_{ij} and the other similarity measures can only encodes direct relationships between pairs of genes, while indirect relationships are ignored. Second, the adjacency matrix is very large, and not sparse, although many pairs of genes have correlation values close to zero. This increases the computational complexity of the network analysis and may lead to spurious results. Third, existing methods identify *eigengenes* or *hub genes* within modules, and test whether these genes are differentially expressed in different treatment groups. It is not clear how one can test whether a treatment or condition is associated with the network structure

We propose a novel approach which addresses these concerns. To define a similarity measure, we fit a mixture model to the normalized correlation coefficients. This leads to shrinkage estimation, and to borrowing strength across all pairs. It allows us to compute the posterior probability that the connection between genes i and j is significant. This enables the construction of a sparse adjacency matrix, $\mathbf{A}$. Using this matrix we define a distance measure which takes into account both direct and indirect connections between a pair of genes. This measure, referred to as commuting time is a proper distance metric, and its computation is made much simpler since our mixture model induces sparsity in $\mathbf{A}$. This distance represents how long it takes to hit gene j and come back to gene i in a random walk on the gene co-expression network. The adjacency matrix also allows us to define a Markov chain transition probability matrix to construct the random walk model. This modeling approach allows us to perform hypothesis tests regarding the network structure. For instance, one can test whether a treatment is associated with a change in connectivity between a pair of genes or centrality of a gene.

We illustrate our approach using Broad TCGA GDAC breast invasive carcinoma and Glioma data. The structural difference of the two groups is explored using connectivity and centrality measures based on the commuting time.

Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models

Justin Yang, *Harvard University*
Samuel Kou, *Harvard University*

Shrinkage estimators have profound impacts in statistics and in scientific and engineering applications. In this article, we consider shrinkage estimation in the presence of linear predictors. We formulate two heteroscedastic hierarchical regression models and study optimal shrinkage estimators in each model. A class of shrinkage estimators, both parametric and semiparametric, based on unbiased risk estimate (URE) is proposed and is shown to be (asymptotically) optimal under mean squared error loss in each model. Simulation study is conducted to compare the performance of the proposed methods with existing shrinkage estimators. We also apply the method to real data and obtain encouraging and interesting results.

Falling Rule Lists

Fulton Wang, *Massachusetts Institute of Technology*
Cynthia Rudin, *Massachusetts Institute of Technology*

Falling rule lists are classification models consisting of an ordered list of if-then rules, where (i) the order of rules determines which example should be classified by each rule, and (ii) the estimated probability of success decreases monotonically down the list. These kinds of rule lists are inspired by healthcare applications where patients would be stratified into risk sets and the highest at-risk patients should be considered first. We provide a Bayesian framework for learning falling rule lists that does not rely on traditional greedy decision tree learning methods.

Bayesian Analysis of Joint Modeling of Response Times with Dynamic Latent Ability

Abhisek Saha, *University of Connecticut*
Xiaojing Wang, *University of Connecticut*
Dipak K. Dey, *University of Connecticut*

In measurement testing, inferences about latent ability of test takers have been mainly based on their responses to test items while the time taken to complete an item has been often ignored. With the advent of computerized testing, it becomes much easier to collect the response time of each item without additional cost. The separate analysis of response accuracy and response time in a test might be misleading. To better infer latent ability, a new class of state space models, conjointly modeling response time with time series of dichotomous responses, is put forward. The proposed models can entertain longitudinal observations at individually-varying and irregularly-spaced time points and can accommodate changes in ability and other complications, such as local dependence and randomized item difficulty. The simulation of our models illustrates that by jointly modeling response time with item responses for a series of tests, the precision and reduction of bias for the estimates of individual latent ability can be largely improved. In applying the models to a large collection of reading test data from MetaMetrics company, we further investigated two competitive relationship in modeling response times with the distance of ability and item difficulty (i.e., monotone or inverted U-shape relationship). The empirical results of model comparison support that inverted U-shape relationship is more suitable to exemplify student's behaviors and psychology in the exam.

Fully-Specified Subdistribution Model Using Weibull Distribution

Fatemeh Sadat Hosseini-Baharanchi, *Tarbiat Modares University*
Ebrahim Hajizadeh, *Tarbiat Modares University*
Ahmad Reza Baghestani, *Shahid Beheshti University of Medical Sciences*
Jing Wu, *University of Connecticut*

Competing risks setting, which patients may fail due to different events, frequently arises in survival area. Recently, Ge and Chen (2012) introduced the fully-specified subdistribution model with subdistribution hazard for the primary event of interest and conditional hazards for the competing risks. In the present study, a modification of the fully-specified subdistribution model is proposed. We consider a piecewise function for the primary event of interest in which first piece is followed by a Weibull model and the second piece is followed by an exponential form such that $H(\infty) < \infty$, and a weibull model for the competing risk. Bayesian approach is applied for parameter estimation. In addition, the proposed model is fitted to a real clinical data set.

Application of Diagnostics for Respondent-Driven Sampling

Dongah Kim, *University of Massachusetts Amherst*
Krista J. Gile, *University of Massachusetts Amherst*

Respondent-Driven Sampling (RDS) is a sampling method designed to collect data from hard-to-reach populations; injected drug users, sex workers, and man who have sex with man. Data are collected through a peer-referral process in which members of an initial convenience sample recruit additional members connected through their social networks. Unfortunately, inference from RDS data requires many strong assumptions that may not hold in practice. Therefore we need to check several diagnostics during the data collection of an RDS study. The target population of this study consists of New York City young adult prescription opioid or heroin drug users. I apply diagnostic tools to explore features of the data and collection including homophily, seed bias, and differential recruitment.

A Statistical Analysis of Information Spread Across Social Networks

Michael Piserchia, *University of Rhode Island*
Natallia Katenka, *University of Rhode Island*

In recent years, online social networks have become a very popular and effective forum for information exchange. These large, highly interconnected networks span the globe and have the ability to disseminate information in a fraction of the time it would take other communication networks. Given the myriad ways in which online social networks can be used, creating accurate, predictive models for the spread of information across them is very valuable. With that, modeling processes on large networks is a difficult task. It is computationally expensive, and usually prohibitive, to model a process on the entirety of a very large network. Given these complexities, creating smaller network graphs that are characteristically similar to the original networks graphs enable researchers to run models that are otherwise not feasible. This project aims to create prototypic networks and model the spread of information across them using traditional and network-based epidemiological models to better understand how information spreads across an online social network. More specifically, the focus will be on the spread of the news of a scientific discovery, i.e. The Higgs-Boson particle, on Twitter.

Predictive Network Modeling of the 2014-2015 Ebola Epidemic in Sierra Leone

Daven Amin, *University of Rhode Island*
Adrian Flowers, *University of Rhode Island*

The 2014-2015 Ebola outbreak in West Africa has affected an unprecedented number of lives. The high mortality rate (¿35

For this project, we set out to create a network representation of the Sierra Leone population and model the spread of EVD as a network process. Citizens were represented as network nodes and we represented citizen-citizen interactions using network edges. As EVD is spread through contact with bodily fluids from infected individuals, we proposed to model the disease spread using a modified snowball sampling method on the network edges.

Initially, we created a number of nodes proportional to the civilian population of Sierra Leone in fourteen geographic regions and attached vertex properties to each network node denoting the geographic region of the node. Network edges were then induced using a modified Social Stochastic Block Model (SSBM) based on geographic region and freely available demographic information. By modeling the disease process using a simple linear relationship between adjacent nodes, we were able to obtain the rapid non-linear spreading characteristic of an epidemic.

Model validation was performed using aggregated EVD counts in Sierra Leone collected by Caitlin Rivers, a PhD candidate at Virginia Polytechnic Institute. We compared the output of our disease model and a standard Susceptible-Infected-Recovered deterministic compartmental model with the actual spread of EVD through Sierra Leone. Our model shows a higher level of agreement with the empirical data. Creating the SSBM population network directly from relevant geographic and demographic data has allowed us to give a meaningful interpretation of the model predictions, especially in terms of high risk geographic regions.

A Terminal Trend Model for Longitudinal Medical Cost Data and Survival

Qian Yang, *Dartmouth College*
Tor Tosteson, *Dartmouth College*
Zhigang Li, *Dartmouth College*

Many longitudinal health care studies involve repeated measurements of medical cost and time to event data. It is always complicated to estimate the average total medical cost for a specific group or population due to (1) the presence of censoring during follow-up time, (2) the semicontinuous nature of medical costs data, i.e. the data has a spike of zero followed by a right-skewed distribution of positive values, and (3) the rapid increase of medical costs prior to death, caused by the intensive treatment for patients who are dying. A joint modeling approach on survival and longitudinal data has proven to be valuable in end of life applications, especially when there is high mortality rate. Also, two-part models are commonly used in analyzing semicontinous data. Here we propose to develop and implement a novel joint modeling approach, a terminal trend model, based on the terminal decline model (by Li et al.) to produce individual level trajectories of the censored longitudinal semicontinuous data. The model is capable of capturing the terminal increasing/decreasing trend of the data. This approach includes two sub-models: one is a two-part spline regression model involving retrospective analysis of the odds of being positive and the distribution of positive values for medical costs, and the other is a piecewise exponential survival model. The model will be applied to Medicare data from a longitudinal study of secondary fracture prevention intervention in elder people. Osteoporotic fractures are recognized as a significant public health problem, especially for the elderly population, because of the associated substantial morbidity, mortality and treatment costs. The aging of the U.S. population in the following decades will make it a more critical issue. Also, elders with osteoporosis-related fracture have a dramatically increased risk of subsequent fracture within the first year following the initial fracture. The terminal trend model will provide estimates on costs and life expectancy of fracture patients in different intervention groups. These estimates will be applied in cost effectiveness analyses, which could help policy makers maximize health benefits. A simulation program has been developed using R to build a realistic dataset. The program simulates covariates such as age, gender, race and time to secondary fracture. Two outcomes, time to death and monthly costs are also simulated in the program with censoring. In the simulation, both time to secondary fracture and time to death are simulated using piecewise exponential distributions. The simulation data will be used to evaluate the feasibility of the terminal trend model and validate the estimated parameter.

Network Analysis Applied to Stock Market Data

Gregory Breard, *University of Rhode Island*

Understanding the interrelation between corporations and the industrial sectors that they comprise can be useful in the study of financial markets and economies. It would be advantageous if such relationships could be inferred without any preconceptions about a possible connection between two companies or industries. We selected a data set that consists of the adjusted closing prices for the companies that make up the Standard and Poor's 500 stock market index, gathered each day over a one year period. These were selected for two reasons: they are a well-known and widely accepted indicator of stock market performance, and they represent companies with large market capitalization in a variety of industrial sectors. We propose that the correlation coefficient between the price data of stocks may be one indicator of a relationship between the companies. More specifically, we construct a partial correlation association network over the time series data set. This accounts for the influence that outside attributes (in this case the other companies) have on each pairing. Since calculating the successive correlations creates a multiple testing problem, we apply a Benjamini-Hochberg adjustment in order to restrict the false discovery rate to under five percent. We then visualize the resulting graph and use community detection to gain insights into the meaning of the links it represents. We explore how the correlation between stock market price data demonstrates a relationship between certain market sectors and evaluate how accurately the industry membership of a company can be predicted based on the companies to which it is correlated.

Methods of Adjusting for Misclassification in Respondent-Driven Sampling Data

Isabelle Beaudry, *University of Massachusetts Amherst*
Krista J. Gile, *University of Massachusetts Amherst*
Shruti H. Mehta, *Johns Hopkins University*

Respondent-driven sampling (RDS) is a sampling method designed to study hard-to-reach populations. Beginning with a convenience sample, each participant receives a small number of coupons, which they distribute to their contacts who become eligible. RDS studies ask participants to report on the number of contacts they share with the studied population. Also, a set of characteristics is observed for each participant. The accuracy of these attributes is not considered in current prevalence estimators. However, ignoring misclassification may lead to biased estimates. The main contribution of this study is to propose an analytical correction method and to apply the Misclassification Simulation Extrapolation (SIMEX MC) procedure to correct for the bias introduced by misclassification. These two methods are assessed under varying levels of misclassification across simulated social networks of varying features. Also, an application from an RDS study is presented. Finally, we conclude with our propose extension of two bootstrap procedures to estimate the variability of the adjusted estimators.

A New Monte Carlo Method for Computing Marginal Likelihoods

Yu-Bo Wang, *University of Connecticut*
Ming-Hui Chen, *University of Connecticut*
Lynn Kuo, *University of Connecticut*
Paul O. Lewis, *University of Connecticut*

Evaluating the marginal likelihood in Bayesian analysis is essential in model selection. There are existing estimators based on a single MCMC sample from the posterior distribution, including the harmonic mean (HM) estimator proposed by Newton and Raftery (1994) and the inflated density ratio (IDR) estimator proposed by Petris and Tardella (2003). The HM estimator has been criticized for having large variance. The IDR estimator requires reparameterization and a careful selection of radius. We propose a new class of Monte Carlo estimators just based on this single MCMC sample. This class can be thought of as an HM estimator using an adaptively weighted kernel (likelihood times prior). We discuss how HM and IDR can be thought of as special cases of our proposed HAWK estimator. We also show that our estimator is consistent and has better theoretical properties than the HM and IDR estimators. In addition, we provide guidelines and an adaptive procedure on choosing the optimal weights. An extensive simulation study is conducted to examine the empirical performance of the proposed estimator and the methodology is applied to an analysis of a real data set from a six-cities child wheezing study.

Statistical Modeling of High-throughput RNase Footprinting for Genome-wide RNA Structure Inference

Chenchen Zou, *The Jackson Lab*
Zhengqing Ouyang, *The Jackson Lab*

High-throughput sequencing (HTS) technologies have emerged to dissect RNA structures at the genome scale. HTS profiles of RNase footprinting by structure-specific enzymes provide complementary information on structural features of thousands of RNAs simultaneously. Integrative analysis of sequencing read data from these assays remains challenging because of the issues of data sparsity, signal variability, and correlation as well as contradiction among the profiles. We present a novel statistical framework to integrate high-throughput footprinting profiles of the double-strand specific RNase V1 and single-strand specific nuclease S1. Our modeling approach utilizes the joint Poisson-Gamma mixtures (JPGM) coupled with the hidden Markov model (HMM), allowing genome-wide inference of RNA structure at single-nucleotide resolution. We test the JPGM+HMM on simulated datasets and genome-wide V1 and S1 footprinting data of yeast. By comparing to approaches analyzing the V1 or S1 profile separately or simply combining them, we demonstrate that our joint modeling approach probes nearly half times more nucleotides without compromising accuracy, and resolves the ambiguity of strandedness of 300,000 nucleotides with overlapping V1 and S1 peaks. Furthermore, using a shared latent variable for modeling accessibility, our model reveals the prevalent influence of three-dimensional (3D) conformation of RNA on RNase footprinting.

Induction and Priors for Finite Populations

Sudip Bose, *The George Washington University*

The problem of induction is an interesting problem in science. Laplace, Jeffreys and Good are some of the famous names who have used probabilistic or statistical approaches. We shall discuss a Bayesian approach and in particular consider the selection of priors. We present classes of priors for which the Bayes factor does not depend on the population size.

What Can We Tell about a Consumer's Political Affiliation Based on Co-Purchases on Amazon?

Gabriel De Pace, *University of Rhode Island*
Benjamin Ott, *University of Rhode Island*

Knowing a book a consumer is interested in buying, we can recommend some other titles likely to be found interesting. This is a common feature of online retailers we are all used to seeing, but how could these recommendations be made? Using a data set of political books co-purchased on Amazon in the early 2000's, we performed statistical network analysis to explore the possibilities.

For a descriptive analysis of the resulting network graph characteristics we used simple node and edge counts, node degree analysis, and more sophisticated measures of centrality. This gave a clearer picture of the books themselves, and which ones in particular were not only popular, but connected to the others. This led us to questions about the cohesion of the graph. How closely related are these books? Are there definite purchasing patterns based on consumer politics? Could we detect these communities without knowing the affiliations in advance? We conducted local density tests, and computed joint node and marginal degree distribution analysis to try and formulate answers to these questions. The communities of political leanings could indeed be detected from the purchasing patterns found in the data. In fact, we were able to predict the political point of view of a theoretical unknown title from the known potential co-purchases. Though this data set was manageable, we experimented with two data sampling techniques, induced and snowball. The sampled data showed similar results as expected.

All of the analysis was automated using R, a statistical software environment. We concluded that there are definite relationships between book co-purchases made on Amazon. These data can be organized as a network, visualized and the patterns found within exposed and studied. We confirmed that political affiliation can be determined based on buying habits, and even future purchases can be predicted.

Analysis and Comparison of Air pollutants in Wuhan, China

Zihao Zhang, *Brown University*
Cici Bauer Bauer, *Bronw University*
Zhijin Wu, *Bronw University*

Air pollution is a major risk to public health worldwide. The situation is particularly dire in China where PM2.5 index in major cities is often four or five times higher than that in New York City. Recently, the Chinese government has increased monitoring and become more transparent about pollution data. We took advantage of this trend and applied a web crawler tool to automatically collect daily air pollution report from a government agency website. To analyze the trend and spread of air pollutants, we collected the air pollution data from Government Environment website of Wuhan, China. The data include Air Quality Index (AQI) of five major pollutants (PM2.5, SO2, NO2, CO and O3) in 11 stations from 2013-01-03 to 2015-01-22. We developed comprehensive data visualization that enhances the interpretation by the public. Our analysis confirms that all five pollutants have apparent seasonal trend and high daily variance. PM2.5, NO2, SO2 and CO remain at higher level and larger variance in the winter, while O3 has higher mean and larger variance in the summer. SO2, NO2 and CO remain significantly lower level in rural areas. As contrast, PM2.5 concentration did not appear to vary significantly between stations, which indicates that PM2.5 spreads easily in a wide geological field. To assist the visualization of the pollution trend over time and space, we also use animation to show the change of different pollutants at all stations in the map over time. Our immediate plan is to disseminate the tool for data visualization to the general public, in the form of dynamic website based on the web crawler, to further increase the awareness of the air pollution risk. Our long term goal is to investigate the relationship between the concentration of specific air pollutant with human behavior and health risks.

NESS 2015 Participants

Abidemi Adeniji, Boehringer Ingelheim Pharmaceuticals, Inc.
Edoardo M. Airoldi, Harvard University
Fahad Aldosry, University of Massachusetts - Amherst
Mahasin Alnumani, West Virginia University
Yasuo Amemiya, IBM Thomas J. Watson Research Center
Daven Amin, University of Rhode Island
Claudio Antonini, AlixPartners LLP
Taylor Arnold, AT&T Labs
Brien Aronov, Travelers Insurance
Brian Bader, University of Connecticut
Khandoker Shuvo Bakar, Yale University
Seojin Bang, University of Connecticut
Yiqi Bao, UFSCAR/University of Connecticut
Sudeep Bapat, University of Connecticut
Yaakov Bar-Shalom, University of Connecticut
Guillaume Basse, Harvard University
Cici Bauer, Brown University
Iddo Ben-ar, University of Connecticut
Nate Bennett, Boehringer Ingelheim
Abhishek Bishoyi, University of Connecticut
Iavor Bojinov, Harvard University
Luke Bornn, Harvard University
Sergiy Borodachov, Towson University
Sudip Bose, The George Washington University
Emily Bowers, University of Connecticut
Zach Branson, Harvard University
Gregory Breard, University of Rhode Island
Alexandria Brown, University of Massachusetts - Amherst
Luis Campos, Harvard University
Joseph Cappelleri, Pfizer, Inc.
Luis Carvalho, Boston University
Yu-Ling Chang, Vertex
Xiaohui Chang, Oregon State University
Gordon Chavez, New York University
Kun Chen, University of Connecticut
Ming-Hui Chen, University of Connecticut
Yang Chen, Harvard University
Yiyun Chen, University of Connecticut
Jie Chen, University of Massachusetts - Boston
Zhiyi Chi, University of Connecticut
Sy Han (Steven) Chiou, University of Minnesota, Dulut
Thomas Chung, CSL Behring
Greg Cicconetti, GlaxoSmithKline
Erin Conlon, University of Massachusetts - Amherst
Lane Coonrod, The Hartford
Forrest Crawford, Yale School of Public Health
Debraj Das, North Carolina State University
Tirthankar Dasgupta, Harvard University
Gabriel De Pace, University of Rhode Island
Ved Deshpande, University of Connecticut
Michelle Deveaux, Yale University
Dipak Dey, University of Connecticut
Peng Ding, Harvard University
Richard M. Dudley, Massachusetts Institute of Technology
Sharon Elder, Johnson & Johnson
John Emerson, Yale University
Birol Emir, Pfizer, Inc.
Beatriz Etchegaray Garcia, IBM Thomas J. Watson Research Center
Jose Augusto Fioruci, UFSCAR/University of Connecticut
Jean Fitzmaurice, Arlington, Massachusetts
Patrick Flaherty, University of Massachusetts - Amherst
Laura Forastiere, Harvard University
Zhixuan Fu, Yale University
Siragan Gailus, Boston University
Jan Galkowski, Akamai Technologies
Guojun Gan, University of Connecticut
Theresa Gebert, Harvard University
Mamikon Ginovyan, Boston University
Joseph Glaz, University of Connecticut
Georg Goerg, Google, Inc.
Gyuhyeong Goh, University of Connecticut
Ruobin Gong, Harvard University
James Grady, UConn Health Center
Jiang Gui, Dartmouth College
Jian Guo, Western New England University
Yeongjin Gwon, University of Connecticut
Molly Hahn, Tufts University
Janelle Hajjar, University of Connecticut
Qiuyi Han, Harvard University
Ofer Harel, University of Connecticut
Brian Hartman, University of Connecticut
Karl Heiner, SUNY New Paltz
Moonseong Heo, Albert Einstein College of Medicine
Fatemeh Sadat Hosseini Baharanchi, Tarbiat Modares University/University of Connecticut
William Hu, University of Massachusetts - Amherst
Jun Hu, University of Connecticut
Yiming Hu, Yale university
Weilong Hu, University of Massachusetts - Amherst

Lu Huang, University of Connecticut
Lulu Ji, Travelers Insurance
John Jiang, Vertex Pharmaceuticals
Yujing Jiang, University of Connecticut
Hongbing Jin, University of Connecticut
Hee-koung Joeng, University of Connecticut
Curtis Johnston, Metrum Research Group
Michael Kane, Yale University
Edward Kao, Harvard University
Ban Kawas, IBM Thomas J. Watson Research Center
Daeyoung Kim, University of Massachusetts - Amherst
Eunhee Kim, Brown University
Dongah Kim, University of Massachusetts - Amherst
Mark Kon, Boston University
Viktoriya Krakovna, Harvard University
Srikanth Krishnamurthy, Babson College
Lynn Kuo, University of Connecticut
Chia-Ling Kuo, UConn Health Center
John Labarga, University of Connecticut
Nathan Lally, University of Connecticut
Ross Lally, Cigna Healthcare
James Landgrebe, Travelers Insurance
David Landy, Travelers Insurance
Chantal Larose, University of Connecticut
Michael Lavine, University of Massachusetts - Amherst
Yang Li, Harvard University
Ta-Hsin Li, IBM Thomas J. Watson Research Center
Alex Li, Travelers Insurance
Hao Li, University of Connecticut
Jianan Lin, University of Connecticut
Henry Linder, University of Connecticut
Yang Liu, University of Connecticut
Jin Lu, University of Connecticut
Lu Lu, Colby College
Yu Lu, Yale University
Qiongshi Lu, Yale University
Yuefeng Lu, Sanofi
Chongliang Luo, University of Connecticut
Shuangge Ma, Yale University
Li Ma, Duke University
Suman Majumdar, University of Connecticut
Solaiappan Manimaran, Boston University
Emily Martin, EMD Serono
Avery Mcintosh, Boston University
Fabrizia Mealli, Harvard University
Tom Meyer, University of Utah
Daniel Meyer, Pfizer, Inc.
Aditya Mishra, University of Connecticut
Shariq Mohammed, University of Connecticut
Peter Müller, University of Texas at Austin
Nitis Mukhopadhyay, University of Connecticut
Balgobin Nandram, Worcester Polytechnic Institute
Benjamin Neale, Massachusetts General Hospital
Matey Neykov, Harvard University
David Ohlssen, Novartis
Benjamin Ott, University of Rhode Island
Zhengqing Ouyang, The Jackson Laboratory
Chris Parks, Travelers Insurance
Lijun Peng, Boston University
Michael Piserchia, University of Rhode Island
Daniel Polhamus, Metrum Research Group
Aleksey Polunchenko, Binghamton University
Vladimir Pozdnyakov, University of Connecticut
Danielle Prado, UFLA/University of Connecticut
Jing Qian, University of Massachusetts - Amherst
Eugene Quinn, Stonehill College
Adrian E. Raftery, University of Washington
Vasanthan Raghavan, Qualcomm Flarion Technologies
Nalini Ravishanker, University of Connecticut
Nicholas Reich, University of Massachusetts - Amherst
Alexandria Rhoads, Dalhousie Univ-Halifax, NS
Maxime Rischard, Harvard University
James Rogers, Metrum Research Group
Dooti Roy, University of Connecticut
Arindam Roychoudhury, Columbia University
Denis Rybin, Boston University
Abhisek Saha, University of Connecticut
Sujit Sahu, University of Southampton
Elizabeth Schifano, University of Connecticut
Stephane Shao, Harvard University
Yuanyuan Shen, Harvard University
Daoyuan Shi, University of Connecticut
Danny Silitonga, Hunter College
Kyaw Sint, Yale University
Grigory Sokolov, Binghamton University
Matthias Steinruecken, University of Massachusetts - Amherst
Ian Stevenson, University of Connecticut
Torey Strauser, Valesta Clinical Research
Boqin Sun, University of Massachusetts - Amherst
Zhe Sun, University of Connecticut
Jiehuan Sun, Yale University
Naitee Ting, Boehringer Ingelheim Pharmaceuticals, Inc.
Panos Toulis, Harvard University
Jeffrey Townsend, Yale University
Dustin Tran, Harvard University

Gregory Vaughan, University of Connecticut
Steffen Ventz, Harvard School of Public Health
Rick Vitale, University of Connecticut
Yubing Wan, The Jackson Laboratory
Wenjie Wang, University of Connecticut
Xiaojing Wang, Google, Inc.
Chun Wang, University of Connecticut
Yu-Bo Wang, University of Connecticut
Tong Wang, Massachusetts Institute of Technology
Fulton Wang, Massachusetts Institute of Technology
Joshua Warren, Yale University
Michael Williams, University of New Hampshire
Zheyang Wu, Worcester Polytechnic Institute
Jiacheng Wu, Yale University
Qianzhu Wu, University of Connecticut
Jing Wu, University of Connecticut
Jingyang Wu, Sanofi
Shuang Wu, University of New Hampshire
Jane Wu, Sanofi
Yihui Xie, RStudio, Inc.
Yuqing Xing, University of Massachusetts - Amherst
Jing Xu, Takeda Pharmaceuticals, Inc.
Yifan Xu, Case Western Reserve University
Jun Yan, University of Connecticut
Donghui Yan, University of Massachusetts - Dartmouth
Heng Yang, City University of New York
Huyuan Yang, Takeda Pharmaceuticals, Inc.
Justin Yang, Harvard University
Qian Yang, Dartmouth College
Shihao Yang, Harvard University
Emmanuel Yashchin, IBM Thomas J. Watson Research Center
Jiani Yin, Worcester Polytechnic Institute
Jessica Young, Harvard University
Alice Zelman, University of Connecticut
Li Zeng, Yale University
Yanqiong Zhang, Vertex
Yuping Zhang, University of Connecticut
Yaohua Zhang, University of Connecticut
Anderson Y. Zhang, Yale University
Chen Zhang, University of Connecticut
Fan Zhang, University of Connecticut
Ting Zhang, Boston University
Bo Zhao, University of Connecticut
Lihui Zhao, Northwestern University
Meng Zhao, University of New Hampshire
Yada Zhu, IBM Thomas J. Watson Research Center
Beth Ziniti, University of New Hampshire
Chenchen Zou, The Jackson Lab

Additional Participants

Lu Chen, Worcester Polytechnic Institute
Dan Liu, University of Connecticut
Randy Paffenroth, Worcester Polytechnic Institute
Thelge Peiris, Worcester Polytechnic Institute
Volodymyr Serhiyenko, University of Connecticut
Veronika Shabanova, Yale University
Ellis Shaffer, University of Connecticut
Steven Wasik, The Hartford
Ruochen Zha, University of Connecticut
Xiaoran Li, University of Connecticut
Marcel Carcea, Western New England University
Anna Liu, University of Massachusetts - Amherst
Tiran Chen, Travellers Insurance
Raji Balasubramanian, University of Massachusetts - Amherst
Alicia Davis, University of Connecticut
Paul Lewis, University of Connecticut
Suman Neupane, University of Connecticut
Louise Lewis, University of Connecticut
Cunshan Wang, Pfizer, Inc.
Qian Meng, University of Connecticut