Computational Methods in Social Sciences: Final Report

David Banks (Duke University)
Tom Carsey (UNC)
Malay Ghosh (University of Florida)
Fan Li (Duke University)
Peter Mucha (UNC)
Jerry Reiter (Duke University)
Richard L. Smith (SAMSI Directorate Liaison)
Joe Sedransk (Case Western Reserve University)
Mary E. Thompson (University of Waterloo and Director of CANSSI)

September 30, 2014

1 Organization

1.1 Program Leaders

- Elena Erosheva, University of Washington
- Stephen Fienberg, Carnegie Mellon University
- Krista Gile, University of Massachusetts, Amherst
- Mark Handcock, UCLA
- Tian Zheng, Columbia University

1.2 Directorate Liaison

- Richard Smith

1.3 Local Scientific Coordinators

- Tom Carsey, Department of Political Science, UNC
- Peter Mucha, Department of Mathematics, UNC
- Jerry Reiter, Department of Statistical Science, Duke

1.4 National Advisory Committee Liaison

- Adrian Raftery, Departments of Statistics and Sociology, University of Washington

2 Personnel

2.1 Postdocs

- Bailey Fosdick, PhD, University of Washington
- Neung Soo Ha, PhD, University of Maryland
- Dane Taylor, PhD, University of Colorado, Boulder

2.2 Faculty Fellows

- Fan Li, Department of Statistical Science, Duke University
- Peter Mucha, Department of Mathematics, UNC
- Jerry Reiter, Department of Statistical Science, Duke University

2.3 Graduate Fellows

- Tracy Schifeling, Department of Statistical Science, Duke University
- Simi Wang, Department of Mathematics, UNC

3 Planning Meeting: July 29, 2012

A planning meeting was held at the Joint Statistical Meetings on Sunday, July 29, 2012, 10:00am-1:00pm in HQ Aqua 304 (Hilton San Diego Bayfront). The purpose of this meeting was to discuss the structure of the program and the tentative themes of working groups. Approximately 30 people attended.

4 Opening Workshop: August 18-22, 2013

LOCATION: Radisson Hotel RTP (August 18-21) AND SAMSI (August 22)

The SAMSI program on Computational Methods in Social Sciences is built around three major themes: Social Networks; Agent-Based Modeling; and Statistical Methodology for Censuses and Surveys. The purpose of the Opening Workshop was threefold: (a) to provide a series of tutorial lectures aiming to introduce the major research themes in the field to graduate students and other newcomers to the topic; (b) a series of focused research sessions highlighting current developments, (c) the formation and initial meetings of Working Groups which will meet weekly through the year of the program.

Sunday, August 18, featured five tutorial lectures by leading participants in the field. Topics covered the three main areas of the program and the related mathematical and statistical methodology.

From Monday, August 19, through lunch on Wednesday, August 21, there were five research sessions, each consisting of two or three talks followed by discussion.

Wednesday afternoon, August 21, was devoted to the formation and initial meetings of working groups.

Thursday, August 22, was devoted to initial meetings of the working groups at SAMSI.

4.1 Tutorial Lectures: Sunday, August 18

9:00-10:00 Adrian Raftery, University of Washington
Statistical Demography: Probabilistic Population Reconstruction and Projection.

10:30-11:30 Krista Gile, University of Massachusetts, Amherst
Proceeds of the Partnership: Statistics and Social Science

11:30-12:30 Simon Jackman, Stanford University
Data and Computation in Political Science: The State of the Discipline and Emerging Trends

2:00-3:00 Roderick Little, University of Michigan
The Analysis of Census and Survey Data: History, Current Approaches, and Research Topics

3:30-4:30 Sara Del Valle, Los Alamos National Laboratory
Agent-based Modeling Approaches for Simulating Infectious Diseases

4.2 Research Session 1: Networks. Monday, August 19

Organizer: Krista Gile, University of Massachusetts at Amherst

9:00-9:45 Edo Airoldi, Harvard
Design and Analysis of Experiments in the Presence of Network Interference

9:45-10:30 Mark Handcock, UCLA
Exponential-Family Random Network Models for Social Networks

11:00-11:45 Tom Carsey, University of Chapel Hill
Networking Network Scholars: Generating Best Practices for Archiving Network Data

11:45-12:30 Discussants: David Banks, Duke; Justin Gross, UNC; Peter Mucha, UNC

4.3 Research Session 2: Modern Computational Methods for the Analysis of Survey and Census Data. Monday, August 19

Organizer: Jerry Reiter, Duke University

2:00-2:45 Stephen Fienberg, Carnegie Mellon University
Record Linkage as a Statistical Procedure: Some History, Formal Frameworks, Applications, and Challenges

2:45-3:30 David Dunson, Duke University
Bayesian Methods for Huge Multiway Tables

4:00-4:45 Jay Breidt, Colorado State University
Model-Assisted Survey Regression Estimation with the Lasso

4:45-5:30 Discussants: Frauke Kreuter, University of Maryland; Tom Louis, Johns Hopkins University

4.4 Research Session 3: Agent-Based Models. Tuesday, August 20

Organizers: David Banks, Duke University; Sara Del Valle, LANL

9:00-9:45 Georgiy Bobashev, RTI International
Computational Ethnography and Agent-based Modeling

9:45-10:30 Ben Klemens, US Census Bureau
A Simulation of Nonresponse and Imputation

11:00-11:45 Kathleen Carley, Carnegie Mellon University
Networks and Agents: The Value of a Multi-Level Approach to Agent-Based Dynamic- Network Modeling

11:45-12:30 Discussant: Kristian Lum, Virginia Institute of Technology

4.5 Research Session 4: Weighting. Tuesday, August 20

Organizer: Joseph Sedransk, Case Western Reserve University

2:00-2:45 Roderick Little, University of Michigan
Weighting Methods in Surveys

2:45-3:30 Mary Thompson, University of Waterloo
The Use of Weights in Analysis of Survey Data

4:00-4:45 Keith Rust, Westat
Survey Weights for the Analysis of Complex Survey Data

4:45-5:30 Discussant: Joseph Sedransk, Case Western Reserve University

4.6 Research Session 5: Causal Inference. Wednesday, August 21

Organizer: Tian Zheng, Columbia University

9:00-9:45 Michael Sobel, Columbia University
Causal Inference for fMRI Time Series Data with Systematic Errors of Measurement in a Balanced On/Off Study of Social Evaluative Threat

9:45-10:30 James O'Malley, Dartmouth
Causal Estimation of Peer Effects Using Instrumental Variables

11:00-11:45 Elizabeth (Betsy) Ogburn, Johns Hopkins University
Causal Inference for Interference and Social Networks: Challenges and Tools

11:45-12:30 Discussant: Fan Li, Duke University

4.7 Working Group Formation

Wednesday, March 21, 2:00-5:00
Working Group Formation and Initial Meetings

5 Workshop on Social Network Data: Collection and Analysis: Oct. 21-23, 2013

A two-and-a-half day workshop, held at SAMSI, that involved roughly 12-15 invited talks along with formal panel discussions, a poster session, and time available for informal collaboration-building discussions.

This workshop directly interfaced with the Computational Methods in Social Science program year by focusing on pressing issues in the systematic collection, statistical analysis, and mathematical modeling of social science network data. The social world is inherently one of interacting entities. While qualitative and theoretical social science has long had free reign to study complex structures arising from the relations among multiple entities, recent advances in network statistics have begun to allow for the quantitative exploration of these more complex network structures that are central to the structure of the social world. Fundamentally, all networks consist of nodes and edges, or relations between those nodes. Perspectives on networks and the possibilities for statistical research based on such structures are myriad and varied. Additional dimensions of data may be available, including: flows over edges, dynamics over time, and static or fixed covariates on any of the above. Inferential perspectives can then aim to characterize any sub-set of these variables either jointly or conditioning on any others. Data collection, sampling, experimentation, and missing data add further levels of complexity. While the methodological questions associated with networks are broad and disparate, so are the substantive problems they are able to address. Indeed, it is these substantive problems that determine which statistical problems are addressed first. By focusing on data collection efforts (e.g. Add Health, micro-financing in Indian villages) and the relevant methodologies for their analysis, we aimed to further engage mathematical, statistical and computational approaches with social science questions.

5.1 Workshop Organizers

- Tom Carsey
- Stephen Fienberg
- Krista Gile
- Peter Mucha

5.2 Monday, October 21, 2013

9:30-9:40 Welcome Remarks (Peter Mucha, University of North Carolina)

9:40-10:15 Eric Kolaczyk, Boston University
Estimating Network Degree Distributions from Sampled Networks: An Inverse Problem

10:15-10:50 Krista Gile, University of Massachusetts
Inference from Link-Tracing Network Samples

11:20-11:55 Brendan Murphy, University College Dublin
Mixed Membership of Experts Stochastic Blockmodel

1:55-2:30 Jacob Foster, UCLA
Cultural Enrichment: Linking Structure to Culture in Network Analysis

2:30-3:05 Tyler McCormick, University of Washington
Latent Space Models for Multiview Network Data

3:35-4:10 Elena Erosheva, University of Washington
Asking Questions about Numbers: Practical Considerations in RDS Degree Measurement

4:10-5:00 Student Poster Fast Forward

5:00-7:00 Poster Session and Reception

5.3 Tuesday, October 22, 2013

9:30-10:05 Rebecca Willett, University of Wisconsin
Tracking Influence in Dynamic Social Networks

10:05-10:40 Karl Rohe, University of Wisconsin
Local Clustering and the Blessing of Transitivity

11:10-11:45 Aleksandra Slavkovic, Pennsylvania State University
Differentially Private Graphical Degree Sequences and Synthetic Graphs

1:45-2:20 A.C. Thomas, Carnegie Mellon University
Protocols for Randomized Experiments to Identify Network Contagion

2:20-2:55 Johan Ugander, Cornell University
Graph Cluster Randomization: Design and Analysis for Experiments in Networks

3:25-4:25 Panel Discussion: Tom Carsey, UNC; Steve Fienberg, CMU; Mark Handcock, UCLA

5.4 Wednesday, October 23, 2013

9:30-10:05 Bruce Desmarais, University of Massachusetts
Topic-Partitioned Multinetwork Embeddings

10:05-10:40 Blair Sullivan, N.C. State/ORNL
Is Intermediate-Scale Structure Tree-like in Social Networks?

11:10-11:45 Cosma Shalizi, Carnegie Mellon University
When Can We Learn Network Models from Samples?

11:45-11:55 Concluding Remarks

6 Workshop on Computational Methods for Censuses and Surveys: January 8-10, 2014

This two-and-a-half day workshop, held at the Bureau of Labor Statistics in Washington, DC, involved roughly 12-15 invited talks along with formal panel discussions, a poster session, and time available for informal collaboration-building discussions.

The workshop was part of SAMSI's 2013-14 program on Computational Methods in Social Sciences (CMSS). For many years, practical work with censuses and surveys has involved complex methodological issues. Standard approaches have addressed some of these issues in a reasonably comprehensive form, while leaving other issues with unsatisfactory or incomplete solutions. In addition, large-scale statistical organizations are now encountering important new methodological opportunities and challenges arising from prospective new data sources; and from changes in salient features of the datacollection environment, resource constraints and cost structures. Addressing these issues requires modern statistical methodology and novel computational approaches.

This workshop, in conjunction with other CMSS activities, brought together researchers and practitioners from academia, statistical agencies, and survey organizations to discuss recent research advances and needs related to the abovementioned challenges and opportunities.

RELATED EVENT: On January 7, 2014, Stephen Fienberg of Carnegie Mellon University was the featured speaker at the 23rd Morris Hansen Lecture that took place at the US Department of Agriculture. Further details are here.

6.1 Workshop Organizers

- John Eltinge, Bureau of Labor Statistics
- Stephen Fienberg, Carnegie Mellon University
- Jerry Reiter, Duke University

6.2 Wednesday, January 8, 2014

Session 1: Models for Longitudinal Surveys. Chair: Steve Fienberg

9:15-9:45 Mike Daniels, University of Texas
A Flexible Bayesian Approach to Longitudinal Studies with (Monotone) Nonignorable Missing Data with Extensions to Surveys

9:45-10:15 Daniel Manrique-Vallier, Indiana University
Mixed Membership Trajectory Models for Longitudinal Survey Data on Disability

10:45-11:15 Jason Fields, U.S. Census Bureau
On the Shop Floor: Issues and Questions for Computational Methodologists from the SIPP Program. - Perspectives from a Large-Scale Longitudinal Survey

11:15-11:45 Discussion

Session 2: Imputation in Complex Data. Chair: Richard Smith

1:30-2:00 Jared Murray, Duke University
Bayesian Nonparametric Models for Heterogenous Data

2:00-2:30 Hang Joon Kim, NISS and Duke University
Bayesian Automatic Editing

2:30-3:00 Thomas Mule, U.S. Census Bureau
Application of Administrative Records Usage for the Nonresponse Followup Operation in the Decennial Census

3:00-3:30 Discussion

3:30-5:00 Poster Session

6.3 Thursday, January 9, 2014

Session 3: Integrated Data from Multiple Sources. Chair: Jerry Reiter

9:15-9:45 Scott Holan, University of Missouri
Spatial Fay-Herriot Models for Small Area Estimation with Functional Covariates

9:45-10:15 Tracy Schifeling, Duke University
Combining Information from Multiple Sources in Bayesian Modeling

10:45-11:15 Nat Schenker, National Center for Health Statistics
Combining Information from Multiple Data Systems to Enhance Analyses Related to Health: Examples and Lessons Learned

11:15-11:45 Discussion

Session 4: Record Linkage. Chair: Connie Citro

1:30-2:00 Roee Gutman, Brown University
Full Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs

2:00-2:30 Rebecca C. Steorts, Carnegie Mellon University
Clustering Approaches to Human Rights Violations in Syria

3:00-3:30 Mauricio Sadinle, Carnegie Mellon University
A Bayesian Framework for Duplicate Detection, Record Linkage, and Subsequent Inference with Linked Files

3:30-4:00 William Winkler, U.S. Census Bureau
Quality and Analysis of Sets of National Files

4:00-4:30 Discussion

6.4 Friday, January 10, 2014

Group Discussions

7 Transition Workshop: May 5-7, 2014

The workshop was held at SAMSI.

The goals of the workshop were:

- To reunite the participants from all active working groups in the SAMSI CMSS Program.
- To report and to review the progress of the working groups.
- To foster continuation of the research of the working groups beyond the SAMSI CMSS Program.

There were five sessions, covering the five active working groups:

- Social Networks (session organizers: Peter Mucha and Tom Carsey)
- Causal Inference (Fan Li)
- Censuses and Surveys (Jerry Reiter)
- Weighting in Surveys (Joe Sedransk and Malay Ghosh)
- Agent-Based Models (David Banks)

7.1 Workshop Organizers

- David Banks
- Tom Carsey
- Malay Ghosh
- Peter Mucha
- Jerry Reiter
- Joe Sedransk

7.2 Monday, May 5, 2014

Session 1: Social Networks (Peter Mucha and Tom Carsey, University of North Carolina)

9:00-9:45 Dane Taylor, SAMSI
Complex Contagion on Noisy Geometric Networks

9:45-10:30 Justin Zhan, N.C. A&T University
Networks for Data Science

11:00-12:00 Jake Bowers, University of Illinois and Bruce Desmarais, University of Massachusetts
Experiments on Networks

12:00-12:30 General Discussion

Session 2: Causal Inference (Fan Li, Duke University)

1:30-2:15 Michael Hudgens, University of North Carolina
Causal Inference in the Presence of Interference

2:15-3:00 Andrea Mercatanti, Bank of Italy
Bayesian Inference for Randomized Experiments with Noncompliance and Nonignorable Missing Data

3:30-4:15 Elizabeth Ogburn, Johns Hopkins University
Vaccines, Contagion, and Social Networks

4:15-4:45 Fan Li, Duke University
Weighting Beyond Horvitz-Thompson in Causal Inference

4:45-5:00 General Discussion

5:00-6:30 Poster Session and Reception

7.3 Tuesday, May 6, 2014

Session 3: Censuses and Surveys (Jerry Reiter, Duke University)

9:00-9:45 Tracy Schifeling, Duke University
Marginal Information for Contingency Tables

9:45-10:30 Bailey Fosdick, SAMSI
Relaxing Conditional Independence Assumptions in Data Fusion

11:00-11:45 Mauricio Sadinle, Carnegie Mellon University
Detecting Killings Reported Multiple Times to the United Nations Truth Commission for El Salvador

11:45-12:15 Discussant: Jerry Reiter, Duke University

12:15-12:30 General Discussion

Session 4: Weighting in Surveys (Joe Sedransk, University of Maryland, and Malay Ghosh, University of Florida)

1:30-2:15 David Haziza, University of Montreal
Weight Trimming and Weight Smoothing Methods

2:15-3:00 Qixuan Chen, Columbia University
Modifying Weights to Improve Survey Estimates using Regression Models

3:30-4:15 Ye Yang, University of Michigan
A Comparison of Weighted Estimators for the Population Mean

4:15-5:00 Neung Soo Ha, SAMSI
Modeling and Small Area/Domain Inference for BRFSS Data

5:00-5:15 General Discussion

7.4 Wednesday, May 7, 2014

Session 5: Agent-Based Models (David Banks, Duke University)

9:00-9:45 Gelonia Dent, CUNY
Agent-Based Modeling and Associated Statistical Aspects - An Overview of ABM Applications

9:45-10:30 Kristian Lum, Virginia Tech
An Agent-Based Epidemiological Model of Incarceration

11:00-11:45 Daniel Heard, Duke University
Statistical Inference Using Agent-Based Models

11:45-12:15 Discussant: David Banks, Duke University

12:15-12:30 General Discussion

8 Computational Methods for Survey and Census Data in the Social Sciences: June 20-21, 2014

This workshop will be held at the Centre de Recherches Mathématiques, Montréal, Canada.

This workshop was one of two being held this year (the other is part of the LDHD program) as joint workshops with CANSSI, the newly formed Canadian Statistical Sciences Institute.

8.1 Workshop Organizers

Mary E. Thompson (University of Waterloo)
Louis-Paul Rivest (Université Laval)
David Haziza (Université de Montréal)
Anne-Sophie Charest (Université Laval)
Mike Hidiroglou (Statistics Canada)
Jean Poirier (Université de Montréal)

8.2 Speakers, Discussants and Invited Posters (*supported by SAMSI)

Jimmy Baulne (Institut de la statistique du Québec)
Jean-Francois Beaumont (Statistique Canada)
Raymond Chambers (University of Wollongong)
Qixuan Chen* (Columbia University)
Lisa Y. Dillon (Université de Montréal)
Claire Durand (Université de Montréal)
Michael Elliott* (University of Michigan)
Roee Gutman *(Brown University)
Neung-Soo Ha* (SAMSI)
Yan Kestens (Université de Montréal)
Phillip Kott* (Research Triangle Institute)
France Labrèche (IRSST)
Pierre Lavallée (Statistique Canada)
Roderick J. Little* (University of Michigan)
Thomas Lumley (University of Auckland)
R. Wayne Oldford (University of Waterloo)
Jean Opsomer* (Colorado State University)
Louis-Paul Rivest (Université Laval)
Mauricio Sadinle (Carnegie Mellon University)
Abdelnasser Saidi (Statistics Canada)
Joseph Sedransk* (University of Maryland)
Yajuan Si* (Columbia University)
Chris Skinner (London School of Economics)
Rebecca Steorts* (Carnegie Mellon University)
Hélène Vézina (Université du Québec à Chicoutimi)
Suojin Wang* (Texas A&M University)
Changbao Wu (University of Waterloo)

Number of participants : 46

8.3 Report on workshop (by Mary E. Thompson)

This workshop was co-sponsored by the Canadian Statistical Sciences Institute (CANSSI), the Statistical and Applied Mathematical Sciences Institute (SAMSI) and the CRM. It came about from meetings at the opening workshop of the 2013-2014 program on Computational Methods in the Social Sciences (CMSS) at SAMSI, in particular the formation of a working group on weighting in surveys, co-led by Joseph Sedransk of the University of Maryland and Malay Ghosh of the University of Florida. The workshop incidentally provided a final chance for much of the working group to meet in person.

One purpose for the workshop was to increase opportunities for statisticians and social scientists to communicate about problems and new directions. Thus the first day of the workshop emphasized methods for the social sciences. The lead-off speaker, Chris Skinner of LSE, spoke on using paradata to correct for measurement error in social and economic survey data. The second session laid some groundwork with talks by social and health scientists Claire Durand of Université de Montréal, France Labrèche of the Institut de recherche Robert-Sauvé en santé et en sécurité du travail, and Hélène Vézina of UQ Chicoutimi on various aspects of combining data from multiple sources: analysis techniques, administrative and ethical challenges, and record linkage. Louis-Paul Rivest of Université Laval led the discussion. Then the afternoon focused on the practice and theory of record linkage, beginning with an overview by Mauricio Sadinle of Carnegie Mellon. Jimmy Baulne of the Institut de la statistique du Québec and Abdelnasser Saidi of Statistics Canada described record linkage techniques used in their agencies. Roee Gutman of Brown University and Rebecca Steorts of Carnegie Mellon talked about applications of Bayesian record linkage approaches to problems in medical record mining and estimation of human rights violations, respectively. Lisa Y. Dillon of Université de Montréal spoke after dinner about the census Mining Microdata project funded under the SSHRC/NSERC Digging into Data program.

Most of the second day was devoted to survey data analysis. The morning sessions, with speakers Rod Little of the University of Michigan, Jean-Franois Beaumont of Statistics Canada, Changbao Wu of the University of Waterloo, Qixuan Chen of Columbia University and Suojin Wang of Texas A & M and discussion by Jean Opsomer of Colorado State University and Michael Elliott, University of Michigan –discussed the evolving role of weighting in survey data analysis, and had strong connections with the subject of the SAMSI CMSS working group. In the afternoon, there were talks on the use of network data in sampling and estimation, by Ray Chambers of Wollongong University and Pierre Lavallée of Statistics Canada, with discussion by Phil Kott of Research Triangle Institute. The workshop wrapped up with a session on survey data exploration, visualization, and mapping, subjects of increasing importance in the era of "big data". The speakers were statisticians Thomas Lumley of the University of Auckland and Wayne Oldford of the University of Waterloo, and epidemiologist Yan Kestens of Université de Montréal. Discussant Joseph Sedransk made the closing remarks of the event.

The discussion was very lively throughout the workshop. The topics of record linkage and of weighting in surveys generated some debate between theorists and practitioners and between proponents of Bayesian and frequentist approaches. It is felt that these kinds of discussions could lead to important syntheses in a few years.

The participants were a good mix of relatively new and more seasoned researchers. Attendees Jon Rao of Carleton University, Louis-Paul Rivest, Chris Skinner, Ray Chambers, Rod Little, Thomas Lumley, Jean Opsomer, Joseph Sedransk, and Phil Kott are among the stars of the survey methodology world, and the younger researchers were very appreciative of their presence.

It had been hoped that more postdoctoral fellows and graduate students, particularly Canadians, would have been attracted by the possibilities of some travel support and the chance to present a poster. In future workshops of this kind we will increase the offered support.

We would like to acknowledge the important contributions to this workshop by Statistics Canada. Michael Hidiroglou of Statistics Canada was a member of the organizing committee, and was able to arrange for three speakers from Statistics Canada, an unusually high number in these times of budget shortages. As well, the Institut de la statistique du Québec kindly provided an expert (Jimmy Baulne) to participate and speak on record linkage.

Jean Poirier of the Centre interuniversitaire québecois de statistiques sociales (CIQSS) was able to recruit five social scientists/epidemiologists from Québec doing path-breaking work in several areas related to the workshop themes. Danielle Gauvreau, Director of the CIQSS, spoke at the beginning to welcome the delegates to the workshop. We are very grateful to the CIQSS for this support. We were somewhat regretful that the non-statisticians attended mainly their own sessions. At least in part this was because there is still relatively little communication between social scientists and statisticians in Canada.

The full program is given as an appendix to this document.

8.4 Demographic Data

Gender
Female (16) 35%
Male (30) 65%

Profession
Academic (36) 78%
Education (1) 2%
Government (8) 18%
Industry (1) 2%

Academic Level (if known)
Professor (27) 79%
Postdoc (2) 6%
Graduate Student (5) 15%
Undergrad (0) 0%

Area of Research
Statistics (39) 83%
Health Science (3) 6%
Social Science (4) 11%

Country of Residence
Canada (27) 59%

USA (15) 33%
Australia (1) 2%
New Zealand (1) 2%
Saudi Arabia (1) 2%
United Kingdom (1) 2%

Province, if Canadian
BC (1) 4%
ON (10) 37%
QC (16) 59%

9 Working Group 1: Social Networks

9.1 Personnel

9.1.1 Working group leaders

- Peter Mucha (UNC Mathematics & Applied Physical Sciences)
- Tom Carsey (UNC Political Science & Odum Institute)
- David Banks (Duke Statistical Science)
- Bailey Fosdick (Postdoctoral Associate, SAMSI and Duke Statistical Science)

9.1.2 Postdoctoral associates affiliated with group

- Bailey Fosdick (SAMSI and Duke Statistical Science)
- Dane Taylor (SAMSI and UNC Mathematics)
- Nishant Malik (UNC Mathematics)

9.1.3 Graduate students affiliated with group

- Simi Wang (UNC Mathematics and current SAMSI graduate fellow)
- Hsuan-Wei "Wayne" Lee (UNC Mathematics)
- Joan Pharr (UNC Mathematics)

9.1.4 Other active members

- Jake Bowers (University of Illinois at Urbana-Champaign Political Science and Statistics)
- Skyler Cranmer (UNC Political Science)
- Bruce Desmarais (University of Massachusetts Amherst Political Science)
- Tyler McCormick (University of Washington Statistics and Sociology)

- Brendan Murphy (University College Dublin Mathematical Sciences)
- Betsy Ogburn (Johns Hopkins Biostatistics)
- Blair Sullivan (NCSU Computer Science)

9.2 Topics and goals

This working group grew out of varied discussions both before and at the Opening Workshop and evolved across the Fall semester. In particular, a variety of topics were discussed in our weekly meetings across the Fall, some of which then spun off into their own working groups. Notably, discussions about Respondent Driven Sampling spun off to become their own group; but we understand that working group only met a few times. Some other interests in the intersection of network geometry and spectra merged over into groups associated with the LDHD program. This report does not represent any further activities by these other groups (except insofar as they connect to the network sampling presentations at the October workshop and to the "noisy geometric networks" project described below that was presented by Dane Taylor at the May transition workshop).

The CMSS Social Networks working group was the lead organizing working group for a successful 2.5-day workshop at SAMSI in October. This workshop represented a broad array of interests from the working groups and its spinoffs (such as RDS and other network sampling methodologies).

9.3 Projects (SAMSI postdocs/graduates involved)

1. Experiments on networks (Simi Wang)
   This project was motivated by the fundamental question, if you are going to design an experiment on networks, how many nodes should you treat? The approach of this project was to take various proposals for peer effects in networks to see how they perform in the context of a simulated Ising-based model of diffusion, exploring network structures under which the known effects in the simulation can be statistically identified from the simulated data. This ongoing and successful collaboration included Bowers, Desmarais, Lee & Wang. In addition to their presentation at the transition workshop, they are presenting this work at another upcoming workshop and are working on a publication.

2. Bicycle sharing data (Bailey Fosdick)
   The goal for the bicycle sharing project was to develop methodology that explicitly models both the effect of geographic distance on the travel flow of bicycles between stations and the inherent constraints on flows through the system due to the finite number of bicycles and spaces at each station. Encouraging progress analyzing this data by Fosdick and McCormick points to possible work to be done in the future.

3. Wikipedia (Dane Taylor)
   The goal of this project was to compare and contrast network organization within the Wikipedia as identified from (i) a text analysis topic model and (ii) the hyperlinked network of pages. An Odum Institute programmer has been critically helpful in processing Wikipedia data dumps. Preliminary community detection calculations on the resulting data sets were carried out. While working on this project, we became aware of new literature on topic modeling Wikipedia data, so we did not pursue this aspect further for the time being.

We are assessing possible next steps for this project. This project included effort from Banks, Carsey, Sullivan & Taylor, among many other ideas contributed from the working group.

4. Environmental treaties (Bailey Fosdick)
   The working group has also spent multiple meetings discussing a data set on international environmental treaties between nations. During the Fall, Nishant Malik started on this project but could not continue as he prioritized other projects. Without a "leader" for this project, we then tabled it. Further discussion was rebooted in the second half of the spring semester, with a new collaboration between Fosdick, Cranmer, and Tobias Böhmelt (ETH Zurich) over the past six weeks that we expect to continue fruitfully.

5. Complex contagions on noisy geometric networks (Dane Taylor)
   This project was related to this working group, as well as some of the overlap with the LDHD program, bridged by Taylor, supporting the background necessary for this work by Taylor working with Mucha and other collaborators. The results from this project were presented by Taylor as part of the social networks session at the May transition workshop. A manuscript is in preparation by Taylor, Mucha and others.

9.4 Sources of data

- Wikipedia (`http://en.wikipedia.org/wiki/Wikipedia:Database_download`)
- Capital BikeShare (`http://api.citybik.es`, `http://capitalbikeshare.com/system-data`)
- Ghana voting data (Jake Bowers)
- Environmental Treaties (Skyler Cramner & Tobias Böhmelt)

10 Working Group 2: Causal Inference

10.1 Leaders

Fan Li, Jake Bowers, Tian Zheng

10.2 Participants

Jake Bowers, Peng Ding, Bailey Fosdick, Alan Lenarcic, Fan Li, Andrea Mercatanti (SAMSI visiting fellow), Besty Ogburn Michael Sobel, Tian Zheng.

10.3 Background and Goal

Causal inference concerns evaluating effects of treatments, interventions or actions in randomized experiments and observational studies, which is central to decision making in many disciplines such as social sciences and medicine. The causal inference working group is formed in during the CMSS opening workshop in August, 2013. The group leaders are Fan Li, Jake Bowers, Tian Zheng, and has been maintained by Fan Li and Andrea Mercatanti during 2014 Spring.

The goals of the causal inference are (1) to develop design, theory, analysis and computational tools for drawing causal inference under a wide range of challenging situations, (2) to apply the methods to investigate social science problems with important practical implications, (3) to enhance the communication between researchers in causal inference and others disciplines, and (4) to improve undergraduate and graduate education in causal inference.

10.4 Activities

10.4.1 Web Research Seminars

From September, 2013 to March, 2014, the working group had seven webex research seminars, during each a presentation of a group member is given:

1. Fan Li (Sep 10, 2013): Regression Continuity Designs: Framework and Bayesian Inference
2. Michael Sobel (Sep 24, 2013): Does Marriage Boost Mens Wages? Identification of Treatment Effects in Fixed Effects Regression Models for Panel Data
3. Alan Lenarcic (Oct 8, 2013): Heterogeneous causal inference in the diallel
4. Andrea Mercatanti (Jan 28, 2014): Do Debit Cards Decrease Cash Demand? Causal Inference and Sensitivity Analysis Using Principal Stratification
5. Tian Zheng (Feb 18, 2014): Discussions on social contagion
6. Jack Bowers (March 4, 2014): Ethnicity and Electoral Fraud in New Democracies: Modelling Political Party Agents in Ghana
7. Peng Ding (March 25, 2014): A Paradox in Randomization-Based Causal Inference

10.4.2 Research subgroups

Personal communications between the leaders determines that the most efficient format is to break the working group into smaller subgroups, each with one or more specific research topics. Active subgroups include: (1) Andrea Mercatanti and Fan Li; (2) Jack Bowers and Bruce Desmarais; (3) Tian Zheng and Michael Sobel.

10.5 Research Products and Achievements

10.5.1 Papers

The following manuscripts are direct products of the causal inference working group.

1. Mercatanti, A, and Li, F. (2014a). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. invited revision, The Annals of Applied Statistics.
2. Mercatanti, A, Li, F, and Mealli, F. (2014). Improving inference of Gaussian mixtures using auxiliary variables. under review, Scandinavian Journal of Statistics.

3. Mercatanti, A, and Li, F. (2014b). Bayesian Inference for Randomized Experiments with Noncompliance and Nonignorable Missing Data. Manuscript.

4. Li, F, and Mercatanti, A. (2014). Do Debit Cards Decrease Cash Demand? Evaluation and Sensitivity Analysis Using Principal Stratification. Manuscript.

5. Li, F, Morgan, LK, and Zaslavsky, AM. (2014). Balancing covariates via propensity score weighting. arXiv:1404.1785.

6. Jake Bowers and Bruce Desmarais et al. (2014): Design of randomized experiments for causal inference when treatment dose/exposure propagate across a network/graph. (This subgroup is overlapped with the Network working group.)

10.5.2 Presentations

Related research has been presented in several conferences and seminars, including ENAR 2014, UNC-CH Biostatistics Department Seminar, Atlantic Causal Inference Conference 2014.

10.5.3 Education and Mentorship

Several of the research topics developed in the working group have been incorporated into the undergraduate course of"Studies and Designs of Causal Studies", and the graduate course of "Statistics Case Studies" in the Duke University Department of Statistical Science. Fan Li is also acting as a research mentor to the SAMSI visiting fellow Andrea Mercatanti.

10.6 Continuation Plan

The causal inference working group has successfully brought a group of researchers in causal inference together and produced a number of high quality papers. The group members plan to continue to work on the research projects developed during the program and disseminate the research results through publishing in statistical and applied journals, as well as presentations in major statistical conferences after the CMSS program ends. The collaboration built between several group members during the program will also be carried on.

11 Working Group 3: Censuses and Surveys

As it becomes increasingly expensive and difficult to mount new data collection efforts, survey organizations need to consider combining information from multiple sources. The Censuses and Surveys working group focused on novel approaches to combining information from multiple sources. Specific topics included (i) incorporating external information when fitting models on survey data with applications to imputation of missing values, (ii) using multiple sources i n data fusion, and (iii) linking records from multiple databases. The working group also briefly discussed statistical disclosure limitation methodology, as this is an interest of several members. These particular discussions did not result in any tangible output.

11.1 Participants

- Faculty
  - Jerry Reiter, Statistical Science, Duke University (WG leader)
  - Joe Sedransk, Statistics, Case Western Reserve University
  - Elena Erosheva, Statistics, University of Washington
  - Aleksandra Slavkovic, Statistics, Penn State University
  - Rebecca Steorts, Statistics, Carnegie Mellon Universirty
  - Yong Wang, Mathematics and Statistics, Eastern Kentucky University
- Postdocs
  - Yajuan Si, Statistics, Columbia University
  - Neung Soo Ha, SAMSI (webmaster)
  - Bailey Fosdick, SAMSI
- Graduate Students
  - Nicole Dalzell, Statistical Science, Duke University
  - Monika Hu, Statistical Science, Duke University
  - Tracy Schifeling, Statistical Science, Duke University (SAMSI grad student)
  - Mauricio Sadinle, Statistics, Carnegie Mellon

11.2 Follow-up activities

Members of the WG will continue to meet over the summer and beyond.

11.3 Works in progress

1. Tracy Schifeling and Jerome Reiter. *Incorporating prior information in latent class models.* To be submitted summer 2014.
2. Tracy Schifeling and Jerome Reiter. *Imputing nonignorable nonresponse using auxiliary information.* To be submitted winter 2014.
3. Bailey Fosdick, Tracy Schifeling, Nicole Dalzell, Jungchen Hu, Elena Erosheva, and Jerome Reiter. *Utilizing marginal information in model-based data fusion.* To be submitted summer 2014.

11.4 Topics investigated by WG

1. *Incorporating prior information in latent class models* In latent class models for categorical data, each individual is a member of an unobserved group, and variables are conditionally independent within groups. Latent class models have been shown to be effective at capturing complex dependence structures, particularly in large-scale categorical data. However, it is not obvious how to incorporate prior information on marginal probabilities into the models. For example, if one knew the percentage of particular demographic proportions from census counts, it would be sensible to incorporate that information in the prior distributions for the parameters of the latent class model. To incorporate prior information, we propose to add pseudo-observations to the collected data. These pseudo-observations have "observed" values for the variables involved in the known marginal distributions and are missing values for the remainder of the variables. These "observed" values are set to match the known margins; for example, if we know the percentage of men is 50% we add a very large number of pseudo-observations comprising 50% men and 50% women. In this way, the latent class model estimates the known marginal distribution exactly but estimates the rest of the joint distribution using the information from the collected (not augmented) data. In addition to incorporating marginal information in the prior distributions of latent class models, we expect this approach to be useful for a variety of settings. For example, we showed that it can be used to adjust latent class model inferences for stratified random sampling and to perform sensitivity analysis in data fusion. We will submit a manuscript on this work in summer 2014.

2. *Using information about margins in imputation of missing data* Many surveys suffer from unit (and item) nonresponse. One approach to handling this nonresponse is multiple imputation, in which one imputes missing values multiple times. When unit nonresponse is not missing at random, typical approaches to imputation—whether based on models or hot decks—result in unreliable imputations. Often, however, one has external information about the marginal distributions of some of the missing variables, for example from census counts or reliable surveys like the American Community Survey. In such contexts, it makes sense to use this information when imputing values for the unit nonresponse. However, we are not aware of any approaches that use this information, particularly for model-based imputations. We developed an approach for imputation of unit nonresponse that utilizes known (up to standard errors) marginal distributions. The basic idea is to augment the sample with pseudo-observations with characteristics that match the known margins. We then add a vector of indicators with values equal to one for all respondents in the planned sample, equal to zero for all nonrespondents in the planned sample (rows for these cases are included in the data but are not populated with values for survey variables), and missing for the cases in the augmented data. The imputations of missing values, including the missing indicator and the missi ng survey variables, are imputed in an MCMC. The resulting distribution of the completed survey variables (imputed plus originally observed) is close to that in the external data, thereby correcting for nonignorable attrition. We will submit a manuscript on this work in winter 2014.

3. *Weakening conditional independence assumptions in data fusion* Data fusion, which is widely applied in business, is defined as follows. Consider two surveys, each with a common set of demographic variables X and a disjoint set of substantive variables Y and Z. The goal is

to create a concatenated data file of both surveys, using imputation to complete the missing items. This is a challenging imputation problem. Since no individual has data on all variables, there is no information about the partial association of Y and Z given X. As a result, most applications of data fusion assume conditional independence between Y and Z given X. However, conditional independence may not be sensible in some applications, so that it is prudent to assess sensitivity of results to violations of conditional independence. We adapt the methodology for incorporating prior information/beliefs into latent class models as a means to weaken conditional independence assumptions. In particular, we create augmented samples that reflect certain beliefs about the relationships among the variables—particularly among Y and Z—and create completed datasets by imputing missing Y, Z in the original (to-be-fused) data. Alternatively, if available, one can use information from pilot studies or other sources by appending them to the original, concatenated samples.

For illustration, we are using data from the company Harper Collins, which they provided freely and without expectation. These data are about reading habits and preferences. We also collaborated with the company CivicScience—an online, rapid-response survey outfit that gets voluntary response samples to short questionnaires—to get data on several Y, Z questions, which we intend to use to inform sensitivity analyses. CivicScience is very interested in working with members of the WG in the long term. Relevant questions to be addressed with CivicScience include, (i) how to best use the information from non-probability samples like CivicScience as external information in data fusion, (ii) how to determine what information is worth paying for, and how much that information is worth, when seeking external information f or data fusion, and, more generally, (iii) how to make inferences from non-probability samples like CivicScience.

4. Record linkage methods Record linkage involves finding common records in two (or more) datasets, and merging those records to create richer information. Most techniques from record linkage do not account for uncertainty in matches when making inferences on the linked data or when creating matched data files. Several WG members are engaged in developing record linkage methods that do account for uncertainty. These members had a series of discussions on record linkage, sharing ideas, techniques, and code. Although this work involved a relatively smaller part of the WG members, we expect these collaborations to lead to improved record linkage techniques in the future.

11.5 Anticipated future publications directly resulting from WG research

1. Bailey Fosdick, Tracy Schifeling, Jerome Reiter. *Picking the glue: what supplemental margins are best to collect for data fusion?*
2. Nicole Dalzell and Jerome Reiter. *Record linkage with uncertain blocking variables.*

12 Weighting

12.1 Personnel

12.1.1 Group Leaders

- Malay Ghosh (University of Florida)

- Joseph Sedransk (Case Western Reserve University)

12.1.2 Other Leading Participants

- Qixuan Chen (Columbia University)
- Mike Elliott (University of Michigan)
- Neung Soo Ha (SAMSI postdoc)
- David Haziza (Université de Montréal)
- Meena Khare (National Center for Health Statistics/CDC)
- Rod Little (University of Michigan)
  Yajuan Si (Columbia University)
- Mary Thompson (University of Waterloo)
- Ye Yang (University of Michigan)

12.2 Report of Activities

From September 2013, this group held regular WebEx meetings on Wednesdays from 2:30-4:30pm. The main focus of these meetings has been the discussion of weight trimming and smoothing methods that may provide improvements over the use of the standard Horvitz-Thompson estimator (with enhancements to include adjustments for nonresponse, poststratification, etc.). The group has worked on four projects during the year, whcih are summarized next.

12.2.1 Project 1: Review Paper

The first objective was to outline a review paper describing and, ideally, comparing these alternative methods. The current draft of this review paper is available. Concurrently, we are conducting simulation studies to provide comparisons of the alternative methods. When completed, the results from these simulations will comprise a section of the review paper.

12.2.2 Project 2: Galveston

This project, centered at the University of Waterloo (Thompson, Chen, Hobbs) is a study of the effects of Hurricane Ike on inhabitants of Galveston TX. A three wave survey was conducted after the hurricane, and we have data on socio-demographic variables and the perceived psychological effects of the hurricane on the respondents. Our objective is to investigate the use of visualization techniques to aid in the analysis of these data.

Survey data collected with complex probability sampling designs are special for two reasons. For one, the aims are often both descriptive and analytic. On the descriptive side, the purpose is to use the sample results to describe the population, while on the analytic side, the purpose is to use the sample results to model the dependence of certain responses or outcomes on explanatory variables. The second reason is that the variables determining the sampling design, such as the

cluster sizes, and those involved in constructing the survey weights, may be "informative". That is, the distribution of the outcomes, given the explanatory variables, may depend on which units were chosen and how. It would be useful to provide survey data analysts not only with exploratory tools for model selection, but also with methods for visualizing the extent to which the sampling design is informative. The working group has undertaken work on a case study to try to develop and illustrate strategies for visual exploration of survey data. The data for the case study come from the Galveston Bay Recovery Survey, of which the purpose was to assess the well-being of the Galveston Bay area population following the severe damage caused by Hurricane Ike in 2008. This survey was chosen because the survey has a complex sampling design and a longitudinal structure (with Waves 2 and 3 two months and one year after Wave 1, respectively), as well as research aims that require sophisticated analyses. Aspects of spatial geography are important as potential explanatory variables.

Following development of an analytic plan and ethics clearance for secondary analysis, the data were provided to some members of the working group on March 3, 2014. Explorations of the data have begun, with some preliminary results being shared with the group on April 16. The work will continue over the summer months.

12.2.3 Project 3: Small Area Estimation

This project (Ha, Sedransk) is to provide improved ways of making inferences for quantities associated with small geographical areas and subpopulations, and, in particular, to investigate the role of conventional survey weights in such analyses. This research project uses data from the 2010 BRFSS sample survey in Florida with the objective of making inferences for counties and subpopulations of the proportion of individuals without health insurance. We present a template for such analyses, thus facilitating the use of Bayesian methods. There is special emphasis on model diagnostics, model checking (including graphical methods) and the use of maps to display the results.

12.2.4 Project 4: New Priors for Random Effects Models

The final project (Ghosh, Ha, Sedransk) is also about small area estimation. The area level models that are typically used in this context are random effect models that include random area effects. However, very often these area effects are so insignificant that the assumption of homoscedastic error variances for all small areas is not very meaningful. The new methodology is to use "global-local shrinkage priors" where the variance for an individual area level random effect is a product of two components, one corresponding to that local area while the other is a common global parameter for all of the areas. The global parameter should be small to squelch insignificant area effects towards zero, while the local shrinkage parameters should be large to offset the effect of the global parameter so that direct estimates for relatively large areas have very little shrinkage effect.

13 Working Group 5: Agent-Based Models

The goals of the Agent-Based Model (ABM) WG were to (1) develop methods for statistical inference based upon ABMs and (2) to explore applications of ABMs to important problems.

The first goal reflects the broad popularity of ABMs in many fields of social science. This popularity is due to the fact that such models are often relatively easy to program and to validate,

and provide insight into how interesting ensemble behavior can flow from the relatively simple rule sets with which agents are endowed. But despite wide use, relatively little is known about their inferential properties. ABMs are models in the same sense that linear regression is a model, but he complexity of ABM parameter spaces and the lack of a tractable likelihood function prevents use of traditional tools for quantifying uncertainty, estimating parameters, testing for effects, and performing goodness-of-fit assessments.

The second goal reflects the broad relevance of ABMs in modeling complex systems, especially in the social sciences. Common applications include the study of mechanisms for change in social networks, the behavior of economic systems and how they respond to different incentive structures, the spread of disease, criminal recidivism, and flow through transportation networks. The ABM working group wanted to identify how different features of a system either enabled or impaired the use of ABM technology.

13.1 WG members

Senior Researchers:

- David Banks, Duke University
- Georgiy Bobashev, Research Triangle Institute
- Sara Del Valle, Los Alamos National Laboratory
- Gelonia Dent, City University of New York
- Brian Frizzelle, University of North Carolina, Chapel Hill
- Kristian Lum, Virginia Tech
- Alyson Wilson, North Carolina State University
- Daniel Heard, Duke University
- Jacob Norton, North Carolina State University
- Tracy Schifeling, Duke University

13.2 Publications directly resulting from ABM WG research

1. Lum, K., Price, M., and Banks, D., Applications of Multiple Systems Estimation in Human Rights Research, *The American Statistician*, **67**, 191–200 (a discussion paper).
2. David Banks and Jacob Norton, "Agent-Based Modeling and Associated Statistical Aspects," to appear in *Encyclopedia of Social and Behavioral Sciences*, ed. by James Wright, Elsevier.
3. Daniel Heard, David Banks, Gelonia Dent, and Tracy Schifeling, "Agent-Based Models and Microsimulation," to appear in the *Annual Review of Statistics and Its Application.*
4. Kristian Lum, Samarth Swarup, Stephen Eubank, and James Hawdon, "The contagious nature of imprisonment: An agent-based model to explain racial disparities in incarceration rates". Accepted by the Journal of the Royal Society Interface

5. Daniel Heard, *Statistical Inference for Agent-Based Models*, Ph.D. thesis, Duke University.

6. Subbiah, R., Lum, K., Marathe, A., Marathe, M. (2013). "A High Resolution Energy Demand Model for Commercial Buildings," in *Security in Critical Infrastructures Today, Proceedings of the International ETG-Congress 2013*.

7. Daniel Heard, Georgiy Bobashev, R. J. Morris, Reducing the complexity of an agent based local heroin market model. submitted to PLOS One

8. Georgiy Bobashev, Daniel Heard, Simulating recovery trajectories of drug users. In preparation.

9. Georgiy Bobashev, Daniel Heard, R. J. Morris, Modeling dynamic interventions and treatment strategies for drug users. In preparation.

13.3 Grant proposals related to work done at SAMSI

1. Kristian Lum, PI, and David Banks, co-PI. "Hazards SEES Type I: An Assessment of the Reliability of Convenience Data Using Agent-Based Models." Submitted to NSF, but rejected.

2. David Banks, PI. "Inference on Agent-Based Models for a Drug Market." Funded by RTI, and provided graduate support for Daniel Heard in Fall, 2013.

3. David Banks, PI. "Model Equivalence." Funded by RTI, and provided graduate support for Daniel Heard in Spring, 2014.

13.4 Presentations

1. Daniel Heard, "Network Analysis Techniques for a Hard-to-Reach Population," Duke Network Analysis Center, November 2013, Durham, NC.

2. Daniel Heard, "Bayesian Network Analysis: HIV Risk in Southern Indian Community," 10th International Conference on Health Policy Statistics, October, 2013, Chicago, IL.

3. Daniel Heard, "Bayesian Network Analysis: HIV Spread in Indian Community," 2013 Joint Statistics Meetings, August 2013, Montreal, Quebec.

4. Gelonia Dent, "Agent-Based Modeling and Associated Statistical Aspects—An Overview of ABM Applications," SAMSI Transitional Workshop, May 2014, RTP, NC. item Kristian Lum, "An Agent-Based Epidemiological Model of Incarceration," SAMSI Transitional Workshop, May 2014, RTP, NC.

5. Daniel Heard, "Statistical Inference Using Agent-Based Models," SAMSI Transitional Workshop, May 2014, RTP, NC.

6. David Banks, "Inference for Agent-Based Models," Fall Technical Conference, October 2014, Richmond, VA.

13.5 Topics investigated by the ABM WG

1. *Statistical Inference for ABMs.* (David Banks, Georgiy Bobashev, Daniel Heard)

   There are two likelihood-free strategies for statistical inference, and such an approach is needed in nearly all ABM applications. One strategy uses emulators, and the other uses Approximate Bayesian Computation. This group applied both techniques to several data sets (HIV transmission, an illegal drug market), finding that emulators generally showed superior performance.

2. *Model Equivalence.* (David Banks, Georgiy Bobashev, Daniel Heard)

   When dealing with complex ABMS (or other models), it is often unclear when two models are identical, or identical in mean, or identical in distribution, where identity is defined modulo a monotone calibration function and an offset. This group found a way to formulate this concept precisely, and obtained a set of theorems describing topological conditions under which model are identical in one of these senses.

3. *Epidemic Models for Incarceration and Recidivism.* (Sara Del Valle, Kristian Lum)

   ABMS are commonly used in epidemiology. This group explored application of such models to incarceration and subsequent criminal behavior. The goal was to model factors that affected the kind and amount of crime that a person would commit, using social network structures and covariates such as age, gender, and previous criminal history.

4. *ABMs for Drug Markets* (David Banks, Georgiy Bobashev, Daniel Heard)

   Scientists at RTI built an ABM describing an illegal drug market in Denver. The ABM included different kinds of agents, such as addicts, policemen, dealers, wholesalers, and the homeless. It was a complex model, and took a long time to run. This group used emulator methodology and concepts of model equivalence to develop a faster and simpler model that nonetheless captured all the relevant behavior of the original ABM.

5. *ABMs for HIV Spread in India* (David Banks, Daniel Heard)

   Using survey data collected by John Schneider on HIV status, sexual position preference, caste, and social network information, this group performed inference using both emulator theory and Approximate Bayesian Computation. The most important conclusions from the standpoint of public health were that there was a need to protect women married to men who had homosexual relations, that antibiotic lubricants could be effective, and that it was possible to prevent spread by identifying and protecting critical individuals. The most important theoretical conclusion was that the emulators gave better predictive accuracy in one-step look-ahead forecasting.

6. *ABMs for the STEM Pipeline* (Gelonia Dent, Daniel Heard, Kristian Lum, Alyson Wilson)

   This group worked out an ABM representation for how women and minorities get drawn into, or lost from, the STEM pipeline. The model included peer network effects, school and teacher effects, and various covariates, such as race, gender, income, and previous education history. The model was not built, due to difficulty in obtaining data that could be used to calibrate it.

7. *ABMs in Counterterrorism* (David Banks, Alyson Wilson)

   As part of the interest in the Laboratory for Analytical Science's data readiness project, this group developed specifications for an ABM that would help identify which potential targets were most likely to be attacked by terrorists. The model considered terrorist agents of many different kinds, where each type ranged from opportunistic to strategic, resourced to unresourced, religious to secular, solo to group. Lack of data prevented full development of this model.

14 Appendix: Program for June 20-21, 2014, workshop on Computational Methods for Survey and Census Data in the Social Sciences, at the Centre de Recherches Mathématiques, Montréal, Canada.

Centre de recherches mathématiques
Université de Montréal

«Méthodes de calcul des données de sondage et de recensement en sciences sociales» Un atelier pour statisticiens et chercheurs en sciences sociales
Du 20 au 21 juin 2014

"Computational Methods for Survey and Census Data in the Social Sciences" A workshop for statisticians and social scientists
June 20-21, 2014

HORAIRE / *PROGRAM*

Conférences : salle 6214 (Pavillon André-Aisenstadt)
Pauses-café : salon Maurice-L'Abbé (salle 6245, Pavillon André-Aisenstadt)

Lectures: *Room 6214 (Pavillon André-Aisenstadt)*
Coffee Breaks: *Salon Maurice-L'Abbé (Room 6245, Pavillon André-Aisenstadt)*

Le vendredi 20 juin 2014 / *Friday, June 20, 2014*

08:00 - 08:30 Inscription (salle 5345) et café-croissants (salle 6245)
Registration (Room 5345) and Coffee & Croissants (Room 6245)

08:30 - 08:45 Mots de bienvenue / *Welcoming addresses*

Session - *Analytical uses of survey data*

08:45 - 09:45 **Chris Skinner** (London School of Economics)
"Using binary paradata to correct for measurement error in survey data analysis"

09:45 - 10:15 Pause-café / *Coffee break*
Salle / *Room* 6245

Session - *Combining data from multiple source*

10:15 - 10:45 **Claire Durand** (Université de Montréal)
"Combining data: Why not dream big?"

10:45 - 11:15 **France Labrèche** (IRSST)
"Consideration of multiple sources of data: an epidemiological study of cancer in the workplace"

11:15 - 11:45 **Hélène Vézina** (Université du Québec à Chicoutimi)
"The linkage of micro census data to vital records: new perspectives for population reconstruction"

11:45 - 12:15 **Panéliste / *Discussant* :** Louis-Paul Rivest (Université Laval)

12:15 - 13:45 Pause-déjeuner / *Lunch break*
Salle / *Room* 6245

Session - *Overview of record linkage*

13:45 - 14:35 **Mauricio Sadinle** (Carnegie Mellon University)
"An overview of record linkage"

Session - *Applications of record linkage*

14:35 - 15:05 **Jimmy Baulne** (Institut de la statistique du Québec)
"Access to linked data; the ISQ approach"

15:05 - 15:35 **Abdelnasser Saidi** (Statistics Canada)
"Overview of record linkage at Statistics Canada"

15:35 - 16:05 Pause-café / *Coffee break*
Salle / *Room* 6245

Session - *Recent developments in record linkage*

16:05 - 16:35 **Roee Gutman** (Brown University)
"Full Bayesian procedure for file linking to analyze end-of-life medical costs"

16:35 - 17:05 **Rebecca Steorts** (Carnegie Mellon University)
"Entity resolution by Bayesian unsupervised clustering: Applications to human rights violations in El Salvador and Syria"

17:05 - 17:40 **Panéliste / *Discussant* :** Jae-Kwang Kim (Iowa State University)

17:40 - 18:45 Session d'affiches et réception / *Poster Session and Reception*
Salle / *Room* 6245

19:00 - 20:00 **Lisa Y. Dillon** (Université de Montréal)
"Footprints in the manuscript: Reflections on the linkage of Canadian historical census data"

Le samedi 21 juin 2014 / *Saturday, June 21, 2014*

08:00 - 08:30 Café croissants / *Coffee & Croissants*
Salle / *Room* 6245

Session - *Beyond traditional weighting*

08:30 - 09:00 **Roderick J. Little** (University of Michigan)
"Weighting for sample selection and nonresponse: a calibrated Bayesian perspective"

09:00 - 09:30 **Jean-Francois Beaumont** (Statistique Canada)
"A weight smoothing approach to improve the efficiency of design-based survey estimators"

09:30 - 10:00 **Changbao Wu** (University of Waterloo)
"Calibration weighting methods for complex surveys"

10:00 - 10:30 **Panélistes / *Discussants* :** Jean Opsomer (Colorado State University) & Michael Elliott (University of Michigan)

10:30 - 11:00 Pause-café / *Coffee break*
Salle / *Room* 6245

Session - *Weights and analysis of survey data*

11:00 - 11:30 **Qixuan Chen** (Columbia University)
"Bayesian post-stratification models using multilevel penalized spline regression"

11:30 - 12:00 **Suojin Wang** (Texas A&M University)
"Maximum likelihood logistic regression with auxiliary information for probabilistically linked data"

12:00 - 13:40 Pause-déjeuner / *Lunch break*
Salle / *Room* 6245

Session - *Networks in sampling and estimation*

13:40 - 14:10 **Raymond Chambers** (University of Wollongong)
"Using social network information for survey estimation"

14:10 - 14:40 **Pierre Lavallée** (Statistique Canada)
"Indirect sampling for hard-to-reach populations"

14:40 - 15:10 **Panéliste / *Discussant* :** Phil Kott (Research Triangle Institute)

15:10 - 15:35 Pause-café / *Coffee break*
Salle / *Room* 6245

Session - *Survey data exploration and visualization*

15:35 - 16:15 **Thomas Lumley** (University of Auckland)
"Scatterplots for complex survey data"

16:15 - 16:45 **R. Wayne Oldford** (University of Waterloo)
"Here be dragons: the challenges of visualizing data on maps"

16:45 - 17:15 **Yan Kestens** (Université de Montréal)
"Leveraging survey data using geographic information systems, or putting statistical analyses into context"

17:15 - 17:45 **Panéliste et conclusion / *Discussant and closing remarks* :**
Joe Sedransk (University of Maryland)