samsi

NSF ● Duke ● NCSU ● UNC ● NISS

CMSS Transition Workshop
May 5-7, 2014

SPEAKER TITLES/ABSTRACTS

Jake Bowers
University of Illinois

"How Should we Design Experiments on Social Networks?"

We present some general principles to guide the design of randomized experiments when the goal is to learn about how an intervention might propagate across the network. We use the potential outcomes framework to help formalize causal inference in this case, and we assess characteristics of the design with regards to the very recent approaches proposed to allow valid statistical inference for counterfactual causal quantities in the context of networks.

Qixuan Chen
Columbia University

"Modifying Weights to Improve Survey Estimates using Regression Models"

We discuss two approaches to modifying survey weights based on regression models. We first review methods that modify weights arising from models for the survey variable, with the survey weights treated as covariates. Such methods include "weight pooling" methods (Elliott and Little, 2000; Elliott, 2008; 2009), "weight smoothing" methods (Elliott and Little, 2000; Elliott, 2007), and the penalized spline predictive methods (Zheng and Little, 2003; 2005; Chen, Elliott and Little, 2010; 2012). We then review weight modeling methods that replace the survey weights by the predictions obtained by modeling the survey weight conditionally on the survey variables (Beaumont, 2008; Kim and Skinner, 2013). We discuss the properties of the methods in both approaches.

Gelonia Dent
CUNY

“Agent-Based Modeling and Associated Statistical Aspects - An Overview of ABM Applications”

In the 1980’s, one of the earliest evolutions of agent-based models (ABM) grew into the field of Computation and Mathematical Organization Theory, which branched into two special interest groups, The Institute of Management Sciences (TIMS) and the Operations Research Society of America (ORSA). Agent-based models use information about individuals (agents) and their interactions to track their behavior within a large complex system. The presentation will provide a motivation for the application and use of ABMs in such areas as economics, geography/ecology and biology/epidemiology to access micro-level behavioral effects. It will cover the Overview, Design concepts and Details (ODD) of ABMs, and some key characteristics of the method that may lead to better understanding of the dynamics of these complex processes. An examination of issues such as data access, computational costs and validation, related to the implementation of ABMs in modeling real-work scenarios, will also be discussed.

Daniel Heard, David Banks, Gelonia Dent & Tracy Schifeling, “*Agent-based Models and Micro-Simulation*”, Annual Review of Statistics and Its Application (submitted March 2014).

Bruce Desmarais
University of Massachusetts

“Models, Methods and Topology: Considerations in Experimental Design for the Study of Interference on Networks”

We have recently seen rapid advances in our ability to measure the connections that comprise networks of people and organizations. As a result, there is increasing interest, across many disciplines, in studying the ways in which attributes of units in networks (e.g., behaviors, attitudes, health conditions) depend upon each other. Randomized controlled experiments have been used in several recent studies to draw causal inferences regarding how the treatment assigned to a vertex effects the other vertices via the network. The process of interest in these studies -- interference -- is one that the classic experimental toolkit is designed to avert, not illuminate. There are pressing challenges in the design of randomized controlled experiments for inference regarding interference. Many of these arise through the intersection of two conditions: (1) the functional forms of interference under consideration and of interest to the researcher, and (2) the topology of the network on which the experiment is to be conducted. In this paper we compare and evaluate several methods that have been proposed for or applied in the context of experimental study of interference in networks. We also provide recommendations regarding strategies in designing experiments on networks. Our findings are illustrated with an application to a field experiment in which we study the spillover of fraudulent voter registration in response to observer placement in Ghana.

Bailey Fosdick
SAMSI

“Relaxing Conditional Independence Assumptions in Data Fusion”

Survey practitioners often try to combine information across multiple surveys by performing data fusion, whereby separate survey data sets are merged to form a single data file with no missing entries. The individual data sets usually contain information on disjoint sets of respondents and have distinct, but overlapping, variable sets. The primary task is the imputation of the variables not included in each of the surveys. Frequently numerous pairs of variables are never observed simultaneously in the data. In these cases, it is standard to use an imputation method that assumes conditional independence between the variables given the variables common to all surveys. In this talk, we consider a situation where auxiliary information, referred to as glue, is available on the dependence between variables not observed concurrently. We discuss different types of glue that may be obtained and compare their utilities when using a latent class mixture model to perform imputations on data from HarperCollins Publishers.

Neung Soo Ha
SAMSI

“Modeling and Small Area/Domain Inference for BRFSS Data”

Large government administered surveys are designed to provide reliable estimates of finite population characteristics for large geographical regions such as the entire U.S. or each of the 50 states, but not for subpopulations and small geographical regions. One objective in this talk to develop a general approach when the variables of interest are binary. For application, we use the 2010 Behavioral Risk Factor Surveillance System, BRFSS, to find inference about the population health characteristics like the health insurance rate or the obesity rate, but the methodology should be useful more widely. For our analysis, we use a Bayesian hierarchical model to combine individual survey data and area level auxiliary covariates. Our preliminary data analysis show that the use of parametric hierarchical methods may need to be extended to provide a satisfactory analysis.

David Haziza
Université de Montréal

“Weight Trimming and Weight Smoothing Methods”

It is well known that point estimators tend to be unstable when the survey weights are highly dispersed and exhibit a low correlation with the study variables. This problem was nicely illustrated by Basu (1971) with his famous example of circus elephants. To limit the impact of highly dispersed weights, a number of techniques has been proposed in the literature, including weight trimming and weight smoothing methods. Although both types of methods are different in nature, they share the same goal: modify the survey weights so that the resulting estimators have a lower mean square error than that of the usual estimators (e.g., the Horvitz-Thompson estimator). Reduction of the mean square error is generally achieved at the expense of introducing a bias. Therefore, the treatment of survey weights by either weight trimming or weight smoothing methods can be viewed as a compromise between bias and variance. In this talk, we will review several weight trimming methods and discuss their properties.

Daniel Heard
Duke University

“Statisticial Inference Using Agent-Based Models”

Agent-based models (ABMs) are computational models used to simulate the behaviors, actions and interactions of agents within a system. The individual automonous agents each have their own set of assigned attributes and rules, which determine their behavior within the ABM system, allowing us to observe how the behaviors of the individual agents impact the system as a whole and if any emergent structure develops within the system.

I begin by presenting some background and theory related to ABMs, including procedures for model validation, assessing model equivalence and measuring model complexity.

I then discuss two approaches for performing likelihood-free inference involving ABMs: Gaussian Process emulators and Approximate Bayesian Computation. I conclude by demonstrating the approaches for inference in two applications.

Michael Hudgens
University of North Carolina

“Causal Inference in the Presence of Interference”

A fundamental assumption usually made in causal inference is that of no interference between individuals (or units), i.e., the potential outcomes of one individual are assumed to be unaffected by the treatment assignment of other individuals. However, in many settings, this assumption obviously does not hold. For example, in infectious diseases, whether one person becomes infected depends on who else in the population is vaccinated. In this talk we will discuss recent approaches to assessing treatment effects in the presence of interference. Inference about different direct and indirect (or spillover) effects will be considered in a population where individuals form groups such that interference is possible between individuals within the same group but not between individuals in different groups.

Fan Li
Duke University

“Weighting Beyond Horvitz-Thompson in Causal Inference”

Balance in the covariate distributions is crucial for an unconfounded descriptive or causal comparison between different groups. However, lack of overlap in the covariates is common in observational studies. This article focuses on weighting strategies for balancing covariates. We propose a general class of weights -- the balancing weights -- that balance the expectation of the covariates in the treatment and the control groups. The framework is closely related to propensity score and includes several existing weights, such as the inverse-probability weight, as special cases. In particular, we advocate a new type of weight -- the overlap weight--that leads to a comparison for the target population with the most overlap in the covariates between two groups. We show that the overlap weight minimizes the asymptotic variances of the weighted average treatment effect among the class of balancing weights and also possesses desirable small-sample property. Simulated and real examples are presented to illustrate the method and compare with the existing approaches.

Kristian Lum
Virginia Tech

"An Agent-Based Epidemiological Model of Incarceration"

We build an agent-based model of incarceration based on the SIS model of infectious disease propagation. Our central hypothesis is that the observed racial disparities in incarceration rates between Black and White Americans can be explained as the result of differential sentencing between the two demographic groups. We demonstrate that if incarceration can be spread through a social influence network, then even relatively small differences in sentencing can result in the large disparities in incarceration rates. Controlling for effects of transmissibility, susceptibility, and influence network structure, our model reproduces the observed large disparities in incarceration rates given the differences in sentence lengths for White and Black drug offenders in the United States without extensive parameter tuning. We further establish the suitability of the SIS model as applied to incarceration, as the observed structural patterns of recidivism are an emergent property of the model. In fact, our model shows a remarkably close correspondence with California incarceration data, without requiring any parameter tuning. This work advances efforts to combine the theories and methods of epidemiology and criminology.

Elizabeth Ogburn
Johns Hopkins University

"Vaccines, Contagion, and Social Networks"

Consider the causal effect that one individual's treatment may have on another individual's outcome when the outcome is contagious, with specific application to the effect of vaccination on an infectious disease outcome. The effect of one individual's vaccination on another's outcome can be decomposed into two different causal effects, called the "infectiousness" and "contagion" effects. We present identifying assumptions and estimation or testing procedures for infectiousness and contagion effects in two different settings: (1) using data sampled from independent groups of observations, and (2) using data collected from a single interdependent social network. The methods that we propose for social network data require fitting generalized linear models (GLMs). GLMs and other statistical models that require independence across subjects have been used widely to estimate causal effects in social network data, but, because the subjects in networks are presumably not independent, the use of such models is generally invalid, resulting in inference that is expected to be anticonservative. We introduce a way to ensure that GLM residuals are uncorrelated across subjects despite the fact that outcomes are non-independent. This simultaneously demonstrates the possibility of using GLMs and related statistical models for network data and highlights their limitations.

Andrea Mercatanti
Bank of Italy

"Bayesian Inference for Randomized Experiments with Noncompliance and Nonignorable Missing Data"

A range of models for randomized experiments with noncompliance and missing data under nonignorable conditions for the missing data mechanism are proposed. We consider the basic case where the treatment and the outcome are binary for which we show the conditions for model identification can be derived from the analysis of the contingency table for the observable data. Identified models will be proposed for some special cases with and without pre-treatment variables. A Bayesian approach to draw inferences for the parameters vector is then developed, and applied to an illustrative comparative analysis based on simulations.

Mauricio Sadinle
Carnegie Mellon University

"Detecting Killings Reported Multiple Times to the United Nations Truth Commission for El Salvador"

Finding duplicates in homicide registries is an important step in keeping an accurate account of lethal violence. This task is not trivial when unique identifiers of the individuals are not available, and it is specially challenging when records are subject to errors and missing values. The task of finding duplicate records in a datafile can be postulated as partitioning the file into groups of coreferent records. Traditional approaches to duplicate detection output independent decisions on the coreference status of each pair of records, which often leads to non-transitive decisions that have to be reconciled in some ad-hoc fashion. We present an approach that targets the partition of the file as the parameter of interest, thereby ensuring transitive decisions. Our Bayesian implementation allows us to incorporate prior information into the duplicate detection process, which is especially useful when no training data are available, and also provides a proper account of the uncertainty of the duplicate detection decisions. We present a study to detect killings that were reported multiple times to the United Nations Truth Commission for El Salvador

Tracy Schifeling
Duke University

"Marginal Information for Contingency Tables"

Given a categorical survey, an analyst may have prior information on the marginal distributions or joint distributions of some of the variables. In this talk we will discuss a way to easily incorporate such prior information using a Bayesian latent class model. We will present empirical examples and show how our method can apply to handling some types of stratified sampling designs in Bayesian modeling.

Dane Taylor
SAMSI

"Complex Contagion on Noisy Geometric Networks"

The study of contagion on networks is central to our understanding of collective social processes and epidemiology. However, for networks arising from an underlying manifold such as the Earth's surface, it remains unclear the extent to which the dynamics will reflect this inherent structure, especially when long-range, "noisy" edges are present. We study the Watts threshold model (WTM) for complex contagion on noisy geometric networks – a generalization of small world networks in which nodes are embedded on a manifold. To study the extent to which contagion adheres to the manifold versus the network, we present WTM-maps that embed the network nodes for analysis as a high-dimensional point cloud.

Ye Yang
University of Michigan

"A Comparison of Weighted Estimators for the Population Mean"

We consider a survey study with known auxiliary information for the population and an outcome that is observed only for the sampled units, possibly subject to nonresponse. The goal is to estimate the population mean of the outcome through the use of weights, which are estimated as the inverse of the sampling probability multiplied by the response probability. We compare through simulations various weighted estimators for the population mean, including the Horvitz-Thompson, Hajek, and their derivatives. In addition we consider a weight pooling method and the penalized spline of propensity prediction (PSPP) model. Performances of the estimators are evaluated based on their root mean square error (RMSE) under simple random sampling (SRS) and probability proportional to size (PPS) sampling, with varying distributions of auxiliary variables, outcome, and nonresponse. Results show that PSPP and Hajek-based estimators generally outperform those of Horvitz-Thompson in our scenarios.

Justin Zhan
North Carolina A&T University

"Networks for Data Science"

The rapid increase in the amount and diversity of data collected in multiple scientific domains implies a corresponding increase in the potential of data to empower important new collaborative research. However, the sheer volume and diversity of these data sets present new challenges in locating data relevant to a particular line of research. Realizing the potential of this long-tail of science data requires investigating algorithms and designing tools specifically targeted to navigate the increasingly complex data landscape. Such a set of algorithms and tools will enable important new multidisciplinary collaborative research on scales up to and including grand challenge problems. In this talk Dr. Zhan will present a number of recent funded projects to address various Data and Network Science challenges.