Title: Big Data in Public Sector

Date and Time: March 10, 2016
12:30pm to 3:30pm

Sponsor: WSS Methodology Section

Location: Bureau of Labor Statistics Conference Center 9 and 10

Guest List: "To be placed on the seminar attendance list at the Bureau of Labor Statistics, or to let us know that you will attend online, you need to pre-register (free) at http://www.eventbrite.com/e/big-data-in-public-sector-registration-21116863106 by noon at least two days in advance of the seminar."

Schedule

Time	Speaker	Affiliation	Point of Contact
12:30	Donsig Jang	Mathematica Policy Research	djang@mathematica-mpr.com
12:40	Frauke Kreuter	University of Maryland	fkreuter@umd.edu
1:05	David Banks	Duke University	banks@stat.duke.edu
1:30	Harlan Harris	The Education Advisory Board	harlan@harris.name
1:55	Intermission		
2:10	Ravi Goyal	Mathematica Policy Research	RGoyal@mathematica-mpr.com
2:35	John Eltinge	Bureau of Labor Statistics	Eltinge.John@bls.gov
3:00	Floor Discussion		

Title: Data Generating Processes and Research Goals: How to think about coverage, measurement, and inference

Frauke Kreuter

Abstract:

The increased digitalization of our economy and society as a whole, spurred the interest of statistical agencies and other producers of statistics, to expand the set of data used to include alternative data sources. Collected through administrative or social processes, these alternative data sources can differ from more traditional ones in size or by the speed with which they can be obtained; however, the most important difference is the lack of research design prior to data collection. Instead, data rise organically, are found by researchers, and need to be retrofitted to match the research question. Also the seemingly lower costs compared to surveys add to that increased interest.

Despite the potential, many arguments have been made for why these alternative data sources are not sufficient to serve all research needs, neither in official statistics, nor in social science research. A few prominent ones are the lack of control over the measurement itself, issues with coverage, and instability of the data sources. One of the biggest sticking points for survey researchers and survey methodologists is the fact that these alternative data sources are not based on random samples from the population of interest, that elements in the data do not have known selection probabilities, and those cannot serve as a basis for inference. As a consequence, data that lack these two features: positive and known selection probabilities, are often dismissed as a basis for solid social science research.

However, looking closely at the nature of the research problems social scientists tackle, this presentation will make three points. First, despite the novelty of the data sources, there is no new inferential issue. Instead we are still faced with the same challenges and responsibilities as we were before in the survey and small data collection environment. Second, given all the other data sources, there are now more opportunities than ever to put our theories out for falsification, which we should embrace. Third, survey methodologist and statisticians have something to offer, to a (data) world that is heavily looking at computer scientists to provide answers on how to deal with Big Data.

Title: Text Mining a Blog Network

David Banks

Abstract:

The last decade has seen substantial progress in topic modeling, and considerable progress in the study of dynamic networks. This research combines these threads, so that the network structure informs topic discovery and the identified topics predict network behavior. The data consist of text and links from all U.S. political blogs curated by Technorati during the calendar year 2012. A particular advantage of the model used in this research is that it naturally enforces cluster structure in the topics, through a block model for the bloggers.

Title: Big but Noisy Data in Higher Education Administration

Harlan Harris

Abstract:

Colleges and universities have a wide array of data sources, from admissions records to transcripts to web app logs to ID card swipes at the library -- and an even wider array of challenges in improving student success outcomes. In this talk, I'll review some of the problems that higher-ed struggles with outside the classroom, including helping advisors target and aid struggling students, helping admissions officers recruit and offer financial aid to best-fit candidates, and helping administrators and faculty design curricula and course schedules. All of these areas have been impacted by the availability of novel data sets and the ability to build analyses and predictive and prescriptive models on top of that data. However, data quality, variety, and sparsity are all major challenges, along with the perennial challenges with building decision-support tools used by nontechnical domain experts. I'll provide some thoughts about statistical techniques that can reduce those challenges and help institutions and vendors build useful tools that improve graduation rates and other outcomes.

Title: Forecasting Network Evolve using Large Temporal Relational Data

Ravi Goyal

Abstract:

The simultaneous advances in network research demonstrating the influence of social networks on our lives and technology, such as cellphones, the Internet, and RFID wireless sensors, to collect detailed information on temporal networks have led to an interest of going beyond passively analyzing to actively intervening on these networks in order to mediate epidemics, dismantle terrorist networks, and increase the effectiveness of marketing. However, current statistical network methods are overwhelmed by vast amounts of fine-scale temporal network data even on moderate populations. We present a statistical method to generate a series of predicted networks based on the historical evolution of social relations in a given population. The rationale for developing a new network generation method was to allow greater flexibility in capturing uncertainty in our estimates of dynamic network properties. The method, which is based on a novel and flexible procedure to sample dynamic networks given a probability distribution on evolving network properties, permits the use of a broad class of approaches in order to model trends, seasonal variability, uncertainty, and changes in population compositions.

We demonstrate the proposed method on two existing dynamic networks; the first represents the sponsor/co-sponsor relationships between senators indicated from bills introduced in the US Senate from 2003-2012. The second is temporal network data-sampled every 20 seconds-representing interactions between participants of the ACM Hypertext 2009 conference. The proposed method enables investigators to make rapid and informed decisions regarding network interventions such as mechanisms to either encourage information diffusion or minimize the impact of a contagion; we present a network-based intervention for the ACM conference for demonstration.

Title: Characterization and Management of Risk and Cost in the Integration of Surveys with Alternative Data Sources

John Eltinge

Abstract:

Government statistical agencies have missions centered on providing the public with high-quality information on a sustainable and cost-effective basis. Historically, these agencies have addressed their goals through the use of sample surveys and some administrative record systems.

The increasing availability of alternative data sources (sometimes called “big data” or “organic data”) provides agencies with an opportunity to reconsider the ways in which they fulfill their missions. Productive responses to that opportunity will require thoroughgoing characterization and management of multiple dimensions of quality, risk and cost that are inherent in statistical production processes.

Following a brief review of recent literature on the quality of alternative data sources, this paper develops a framework for evaluation and management of risk and cost structures. The discussion of risk highlights tools for the timely detection and management of three issues: (1) interruption of standard publication schedules due to loss of a data source or disruption of processing systems; (2) “break in series” phenomena characterized by, e.g., changes in error mean and covariance structures or seasonality patterns; and (3) violation of respondent confidentiality, including both identity disclosure and attribute disclosure.

In addition, work with cost structure involves a wide range of fixed and variable cost components associated with: (a) acquisition of data through surveys or alternative sources; (b) data management, linkage, editing, imputation, curation and documentation; (c) computation, review and dissemination of estimates; and (d) development, testing, implementation and maintenance of systems for (a)-(c). Special challenges in cost assessment and management include the evaluation of approximate costs attributable to distinct parts of legacy processes and prospective alternative processes; amortization of investment costs over time and across product lines; and capture and re-investment of savings obtained through improvements in data sources, methodology or technology.