NSF-Census Research Network
Newsletter
Vol.3, Issue 1

# Research Node Focus: University of Michigan

What if you could get better indicators in real time by analyzing social media? That is one of many exciting areas that the University of Michigan's research node is exploring.

"We are trying to use social media to create real time indicators of economic activity that track what's going on in the economy in a way that's correlated with official measures but brings some value-added to the official measures. These measures potentially show different patterns," said **Matthew Shapiro**, PI of the Michigan research node. "We want to be able to track existing indicators with things such as Tweets. But if the new measure only tracks existing indicators, then one hasn't added much value. We have good existing indicators, so one wants to track existing indicators to show social media can replicate patterns in the data, but then one wants to move on to novel uses of the data in order to measure things that you can't see in official data," he said.

There are many dimensions where official data can be supplemented by these new indicators. For example, they can be produced with a greater geographic or demographic granularity. A second possibility is more timeliness. Social media are available in high volume and nearly immediately whereas standard survey data are available relatively infrequently. Even large surveys are subject to sampling error. It is too costly to produce a large data set in continuous time, so there are often long lags between collecting survey data and getting them processed and ready for publication. Social media can be harnessed essentially over night. "We set up data processing where we, for example, count Tweets at the end of a weekend and by the next day we are able to have a release," explained Shapiro. It is very hard to replicate this kind of timeliness with official statistics.

"The main promise of this research is that social media may tell us something different," Shapiro said. To realize this potential, the Michigan NCRN node is tracking job loss by examining Tweets, such as those that say, "I lost my job," or "I got fired." Up until about 2013, Tweets about job loss and official statistics of insured unemployment claims, tracked one another well. In 2014, the two series came apart; while weekly fluctuations were similar, new claims in unemployment insurance measured by the official statistics fell consistently, while Tweeting about job loss did not. Shapiro said, "So the question was, 'What's going on?' The first thing we did was investigate if there was something happening in the Tweet process. Maybe there was something going on in the way people Tweeted. We ruled that out. We found instead that it reflected changes in the labor market." The research team found that people Tweeting about new jobs and searching for jobs was important in measuring the unemployment level, in a way that it had not been when the job market was in the doldrums. The language in the messages changed in the way that is consistent with a healing job market and better job prospects. "We were able to show that the gap between our social media index and the official index was explained by increased search activity and rates of job postings that were evident in the Tweets," Shapiro explained.

The Michigan NCRN node is also developing economic indicators such as spending and income by using transactional data from checking and credit card accounts that are linked using a financial app. The challenge is to take transactional data – which are collected not for the purpose of measurement – and inferring estimates of spending or income using a combination of economic and statistical modeling. "We take raw accounting data and transform it into something that looks more like the economic concept of spending," said Shapiro.

The high frequency and disaggregation of these data allow for analysis that is not possible using traditional economic measures. For example, it is puzzling that spending is very sensitive to the receipt of income, even among

*Continued on page 2*

## In This Issue

# Research Node - Michigan (Continued)

*From page 1*

people who have significant savings. People seem to spend more on the day they receive a pay check than on other days. Shapiro explained what the Michigan team is doing to explain this pattern, "It turns out if you distinguish between recurring spending and unique spending, we can explain this pattern. Much of the "excess" sensitivity of spending to the receipt of income is simply the timing of recurring spending. It becomes quite clear that people are scheduling regular bill payments, such as rent or mortgage, utility bills, cell phone payments and more." These automatic payments are scheduled to occur just after the person receives his/her paycheck.

*Matthew Shapiro, PI of the Michigan research node.*

"With lower resolution data, it would look like the increase in income led to a relatively large increase in consumption. But in the higher resolution data, when you drill down to the daily frequency, it becomes clear that this is basically sensible cash management," Shapiro said.

The Michigan node's examination of non-survey data allows it to address questions about how to improve survey design. In order to decrease respondent burden, it is important to balance what is asked of respondents and what is obtained from administrative records. The Michigan team links survey respondents' descriptions of their employers to administrative data about those business. It is often difficult for survey respondents to provide accurate, fairly basic facts about their employer, such as the number of employees or the industry, much less more complex information such as co-worker characteristics, or if early retirement occurs in clusters at a specific company or not.

It is possible to create such measures by linking survey data to administrative data. "Say you have name and address information of an employer from a respondent. How do you actually find the employer in a large set of administrative data? How do you model that linkage process? How do you take into account that kind of linkage is subject to error and is inherently probabalistic? We have a project we are doing in collaboration with the Census Bureau to develop new tools for matching information from survey respondents about their employers to administrative data about the business in order to create a probabilistic record linkage."

Michigan researchers are also using satellite imagery to crease new statistical indicators. "Collaborators in the School of Natural Resources have used satellite imagery to measure changes in urbanization, focusing on the Detroit Metropolitan area and Eastern Michigan during the Great Recession. They looked at changes in land cover, what is paved over and what's covered by greenery. They looked at the housing boom and the subsequent collapse of the housing market and the retrenchment that occurred. They can actually see at quite high frequency the changes in urbanization related to the Great Recession," Shapiro noted, "These kinds of measurements should be quite helpful to agencies like the U.S. Census Bureau in identifying where they should put effort to track down new addresses, for example."

Another part of the Michigan node, led by **Luke Shaefer** of the University of Michigan's School of Social Work, is examining new ways to measure poverty. There has been extensive work in this area at the Census Bureau looking at various measures of income, including transfers and non-cash income. The Michigan node has taken this a bit further, both replicating Census approaches and developing new ones using the University of Michigan's Panel Study of Income Dynamics. The team has created measures of poverty and food insecurity that are comparable across different data sources and allow one to compare data from these various sources. The U.S. Census Bureau's Survey of Income and Program Participation (SIPP) is its primary dataset for the study of poverty. SIPP is data-rich, but it is very difficult to use because of its complex design and the extensiveness of the survey. Michigan's team has conducted several workshops for graduate students and junior faculty members including both an introductory course and one that introduces researchers to the SIPP Synthetic Beta (SSB). The SSB is synthetic data maintained by the Census Bureau, and accessible at the Cornell NCRN node. These training sessions have been very successful and have been offered at Duke University and in Washington, DC using the University of Michigan's curriculum. The SSB course will be offered again in Ann Arbor in August 2016. A course on the newly-redesigned SIPP will be held in Ann Arbor in October 2016. See http://ebp-projects.isr.umich.edu/NCRN/training.html for updates on these course offerings.

# Node News

Michigan NCRN graduate students have accepted the following positions: **Isaac Sorkin** has accepted a position as an Assistant Professor in the economics department at Stanford University, **Aaron Flaaen** accepted a position as an economist at the Board of Governors of the Federal Reserve, and **Peter Hudomiet** accepted a position as an economist at the Rand Corporation.

**Adam Eck,** University of Nebraska postdoc, will be an Assistant Professor in the Computer Science Department at Oberlin College starting this fall.

**Antje Kirchner**, also a University of Nebraska postdoc, will be starting at RTI International as Research Survey Methodologist (RTP location) on August 1.

**Jonathan R. Bradley**, University of Missouri postdoc, will begin at Florida State University's Department of Statistics in the fall as an Assistant Professor.

**Noel Cressie**, University of Missouri, won the 2016 Barnett Award for environmental statistics. The award is presented by the Royal Statistical Society. Cressie will give the Barnett lecture in September 2016 at the Royal Statistical Meeting in Manchester, UK.

NCRN is deeply saddened to share the news that University of Nebraska's Co-PI **Allan McCutcheon** passed away May 3. McCutcheon was a distinguished professor who retired from the University of Nebraska-Lincoln. He has been spending time between Lincoln, Nebraska and his retirement home in Puerto Rico. He is survived by his wife, Lisa Crockett and his daughter, Jennifer K. Holm.

## Levenstein named director of ISR Inter-university Consortium for Political and Social Research

**Margaret Levenstein**, University of Michigan NCRN's Co-PI, has been appointed the director of the Inter-University Consortium for Political and Social Research at the

University of Michigan Institute for Social Research. ICPSR, founded in 1962, is the largest archive of digital social science data in the world, with over 500,000 data files.

The announcement follows a national search by a joint committee composed of ISR faculty and members of the ICPSR Governing Council, who represent the consortium's 760 members worldwide. Levenstein will be the ICPSR's first female director.

"Maggie's vision for ICPSR's future is exciting," said ICPSR Governing Council Chair Chandra Muller. "She has a strong academic research background, understands the potential of data science for current and future social science researchers, and has a stellar reputation both nationally and internationally. We were impressed by her skill in bringing together experts from diverse fields for the advancement of social science research goals. The Council is looking forward to working with her."

Levenstein will begin her five-year term on July 1.

# Changes at the NCRN Coordinating Office Executive Committee

As announced in December 2015, **John Abowd** will lead the Census Bureau's Research and Methodology Directorate, and has withdrawn from the Coordinating Office grant. In addition, **Alan Karr** has also expressed an interest in withdrawing from the grant. We will miss their expertise, humor, and other contributions on this grant, and will continue to occasionally badger them with questions.

We are happy to announce that **Dan Weinberg** and **Maggie Levenstein** have agreed to serve on our Coordinating Office management team, providing us with advice and suggestions. Dan is formerly of the Census Bureau, with extensive experience across multiple directorates (and a past NCRN coordinator at the Bureau), and is now Principal at DHW Consulting. Maggie is the co-PI of the University of Michigan NCRN node and incoming director of the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan's Institute for Social Research, (see article above). We think that both are an excellent addition to the team!

*(Continued on page 5)*

# John Abowd to Receive 2016 Julius Shiskin Award

**John Abowd**, Edmund Ezra Day Professor at Cornell University and currently Associate Director for Research and Methodology and Chief Scientist at the Census Bureau, has been selected to receive the 2016 Julius Shiskin Memorial Award for Economic Statistics. The award recognizes original and unusually important contributions in the development of economic statistics or in the use of statistics in interpreting the economy.

Professor Abowd is recognized for designing and implementing disclosure avoidance techniques that enable federal statistical agencies to greatly expand the availability of their data while preserving respondents' confidentiality and for his leadership at Cornell providing access to these data over the Internet. He is also recognized for developing econometric and statistical techniques to conduct labor market analysis. Professor Abowd is the 44nd recipient of the Award; he will be honored at events hosted by the three award sponsors: the Washington Statistical Society, the National Association for Business Economics, and the Business and Economics Statistics Section of the American Statistical Association (ASA).

Abowd's most important contribution to economic statistics was his pioneering efforts to design and implement statistical disclosure avoidance techniques that have enabled statistical agencies to produce and release detailed tabulations and micro-data that both preserve the statistical properties of the original data and their confidentiality. After receiving a PhD in economics from the University of Chicago, Abowd worked on measurement issues in labor economics and on estimating gross labor force flows. A 1989 *Econometrica* article (with David Card) focused on identifying statistical models for dynamic wage processes and investigated the covariance structure of changes in earnings and hours. Subsequently, Abowd used linked data on employers and employees in several European countries to research the joint role of workers and firms in determining labor market outcomes. In conjunction with various collaborators, he developed innovative new econometric methods to analyze these linked employer-employee data. His most notable contribution in this area was the model developed in a 1999 *Econometrica* article (with Francis Kramarz and David Margolis) that used a matched sample of French employees and employers to decompose compensation into components related to employee characteristics, firm heterogeneity, and residual variation. Its econometric approach laid the groundwork for a large body of subsequent research using employee-employer linked data to understand topics such as the role of human capital in wage determination, the measurement and interpretation of wage differentials, and the dynamics of employment and wages.

Following his work with French linked data, in 1998 he joined the team of senior research fellows at the Census Bureau that developed the Longitudinal Employer-Household Dynamics (LEHD) program, which provides public-use data by integrating demographic and economic surveys and administrative data. During the course of developing the LEHD with Professors Julia Lane and John Haltiwanger, it became clear that to make the detailed data from this program available to the public, it would be necessary to develop new methods of statistical disclosure avoidance, because the existing methods were not adequate.

Abowd led the development of these new methods, the first of which was dynamic noise-infusion. This method introduced noise at the microdata level, and used the noise-infused microdata to create aggregate statistics that did not distort critical properties of the underlying data, like trends, while still protecting confidentiality. The second method was the application of synthetic data techniques. Although this was not a new concept, Abowd was one of the first to put the idea into practice as described in a 2001 paper (with Simon Woodcock). He further stimulated research in this area as a founding editor of the online *Journal of Privacy and Confidentiality* and as a major contributor to the literature on privacy and confidentiality.

Abowd's methods have been adopted by many Census Bureau programs -- initially the Quarterly Workforce Indicators and OnTheMap, and more recently Job-to-Job Flows, County Business Patterns, the Survey of Business Owners, Statistics on Businesses, Non-employer Statistics, the Survey of Income and Program Participation, and the Economic Census of Outlying Areas. As a result, the amount of detailed industry and geographic detail accessible to researchers and policy analysts has substantially increased – in the case of Non-employer Statistics by almost double.

# Node News (Continued)

*(From page 3)*

# Melissa Colbeth Joins Cornell NCRN Node

*Melissa Colbeth.*

**Melissa Colbeth** joined the Labor Dynamics Institute at the Cornell NCRN Research Node as Research Project Administrative Assistant. She has been with Cornell since March 28th. Previously, Melissa worked at GrammaTech Inc. where she had several years of experience as an Administrative Assistant. She also has prior experience as an Office Coordinator. Melissa has a B.A. in English with a concentration in Professional Communication and Design from Nazareth College.

# NCRN Spring 2016 Meetings

*By Lars Vilhuber, Principal Investigator, NCRN Coordinating Office and Cornell University node*

The NCRN Principal Investigators and Senior Researchers, as well as a large number of the nodes' graduate students and post-doctoral researchers, met at the U.S. Census Bureau in Washington D.C. on May 9-10, 2016 for a lively and varied meeting. About 40 participants attended the scientific meetings (see following article summarizing the presentations) on the mornings of May 9 and 10, followed by meetings with National Science Foundation Program Officer Cheryl Eavey and Census Bureau leadership. Cornell NCRN PI Lars Vilhuber and Cornell Professor John M. Abowd, who is also the incoming Associate Director for Research and Methodology and Chief Scientist at the Census Bureau and the former PI of the Cornell node, wrapped up the NCRN-sponsored videoconference class INFO7470 on "Understanding Social and Economic Data" on the afternoon of May 9. The nodes' researchers, including nearly a quarter of the over 40 graduate students and post-docs active in the network, Census Bureau staff working with the nodes and some guests from the federal statistical system, then continued lively discussion at the now traditional dinner at the Lebanese Taverna on Connecticut Avenue.

The next NCRN Meetings are currently being planned for the Fall of 2016. We look forward to welcoming node members and the interested public for further discussions and presentations on shared topics.

The next few pages detail the talks that were given at the Spring meeting.

*Renee Ellis, Nancy Bates and Joanne Pascale, all from the U.S. Census Bureau.*

*Enjoying dinner at the Spring 2016 Meeting in Washington DC.*

# NCRN Spring 2016 Meeting

by Melissa Colbeth

The following is a synopsis of the talks at the NCRN Spring 2016 Meeting. Several graduate students and post-docs wrote the summaries. Thank you for taking good notes!

## Session I

by Zachary H. Seeskin

### Attitudes towards geolocation-enabled Census forms

**Laura Brandimarte, University of Arizona & CMU**

Laura Brandimarte of the University of Arizona presented her work with collaborators at the Carnegie Mellon node to study respondents' reactions to different types of geolocation in online surveys. The research project is a partnership with the Census Bureau to help plan for an online response option for the 2020 Census. Geolocation automatically identifies the physical location of the internet user by the IP address. However, geolocation may also raise respondents' concerns about their privacy, leading to either nonresponse or untruthful response.

To study the effects of geolocation on respondents, Brandimarte and her coauthors designed an online census form taken by participants at Carnegie Mellon and Arizona. Respondents were randomly assigned to receive different types of geolocation. The experimental groups included a control, an imprecise identification of the area where the respondent is within a 1- or 2-mile radius and a precise determination of the respondent's location. Brandimarte mostly found that nonresponse to survey items was not affected by the type of geolocation. However, respondents assigned to the precise geolocation group were significantly more likely to admit not responding to the census form truthfully, with some respondents expressing concerns about their privacy. Brandimarte's work will help the Census Bureau consider how to use geolocation in future online surveys to maintain the trust of respondents.

### The ATUS and SIPP-EHC: Recent Developments

**Robert F. Belli, University of Nebraska**

Robert Belli discussed the Nebraska node's use of paradata to study the data quality of the American Time Use Survey (ATUS) and the Survey of Income and Program Participation's event history calendar (SIPPEHC). ATUS and SIPP-EHC are two examples of Census Bureau surveys requiring telephone interviewers to lead respondents to recall events temporally, either for activities from a recent day for ATUS or periods of employment, education, program receipt, etc. from a recent time period for SIPP-EHC. These interviews are typically conducted either via sequential retrieval, where the interviewer asks the respondent to recall events through time, or via parallel retrieval, where the interviewer guides the respondent through themes of activities or events. The Nebraska node is studying how to use paradata, particularly the keystrokes and mouse clicks of the interviewer, to evaluate the data quality of these surveys. For example, certain behaviors of the interviewer may be related to different kinds of recall error for the respondent. Belli concluded by discussing a proposed Enriched Time Diary Instrument, through which a computer conducts the time diary interview and uses insight from data mining to make suggestions to the respondent and better lead the respondent through the interview.

**Itemwise Missing at Random Modeling for Incomplete Multivariate Data**

**Mauricio Sadinle and Jerry Reiter, Duke University and NISS**

Jerry Reiter discussed new research at the Duke node with Mauricio Sadinle to handle survey nonresponse for multivariate data. Often, surveys must make assumptions about the missing data that are untestable and possibly too strong in some cases. Jerry introduced the concept of data being itemwise missing at random (IMAR), meaning that each random variable is conditionally independent of its missingness indicator given the observed values of other variables and other variables' missingness indicators. This assumption is not as strong as the missing at random assumption that is often used in Census Bureau surveys. IMAR can be used with multivariate modeling to study either categorical or continuous variables.

Jerry presented an example studying the relationship between reported and measured height in National Health and Nutrition Examination Survey. Using the IMAR assumption led to different results than estimates using other missingness mechanisms. He also discussed how the IMAR assumption can be used with marginal information from auxiliary data sources to improve estimates.

*Jerry Reiter, Duke University*

*(Continued on page 5)*

# NCRN Spring 2016 Meeting (Continued)

## Session II

by Matthew Simpson

### A 2016 view of 2020 Census Quality, Costs, and Benefits

Bruce Spencer, Northwestern University

"Cost-benefit analysis is always hard. It's incredibly hard for information. And it's even harder to do ahead of time." Bruce began his talk with this particularly apt quote. A major goal of the Census Bureau is to increase the effectiveness of the Census from 2010 to 2020 while also decreasing the cost. The information obtained from the Census is used for a number of things. For example, state population estimates are used to determine a number of things: the number of House seats each state receives is proportional to its share of the U.S. population; the allocation of some sources of Federal funds is determined in the same way; other sources are allocated using different formulas that depend on state population estimates; and finally many surveys, government agencies, and private sector entities use state population estimates and other information from the census in order to make decisions.

In order to estimate the costs and benefits of the 2020 Census, first many key parameters of the Census itself must be forecast, then the costs and benefits must be forecast conditional on these parameters. Currently most of these forecasts are only made marginally. Bruce suggested that some of the challenges of joint probabilistic forecasts could be overcome by using techniques from small area estimation - especially for state-level forecasts. Bruce's Student, Zach Seeskin, has been working on methods to predict error in apportionment of House seats and misallocation of Federal funds due to mis-estimating state population shares, using a response surface model. His method allows for quantifying the distribution of misapportioned seats or misallocated funds given a level of accuracy in terms of RMSE of state population shares. Combined with methods to predict the accuracy of Census state population share estimates, this method looks like a promising way to predict the benefits of increasing the accuracy of these estimates, specifically due to more equitable apportionment of House seats and more equitable allocation of federal funds when those funds are allocated on the basis of state population shares.

### Data Quality in Time Diary Surveys

**Ana Lucía Córdova Cazar, University of Nebraska-Lincoln**

Ana's basic question is to determine whether there is a relationship between the complexity of an interview and data quality in calendar and time use surveys. Previous research indicates this, but a major problem in calendar and time use surveys is that there is no gold standard to compare the data to, so there is no easy way to check the quality of the data. Ana proposes using paradata, e.g. keystrokes and mouse clicks, as indicators of interview complexity and thus data quality. For interviews administered by an interviewer with a computer, much of this data is easily captured. The challenge comes from trying to extract useful information from it.

One such survey is the American Time Use Survey (ATUS). Not only does it capture every keystroke and mouse click of the interviewer, but it also has several error codes for outcome variables. Ana focuses on two in particular - insufficient detail error, and memory gap error. She constructs a structural equation model to predict the occurrence of these errors with a latent predictor composed of fourteen indicators that seem related to interview complexity.

*Ana Lucia Cordova Cazar, U. Nebraska-Lincoln*

Five of the indicators were significant, including interview length and the total number of activities reported. The interview complexity latent factor does predict both types of errors, but the variance explained is not large, at 6% and 2.2% respectively.

Ana also found that some demographic variables predict an error, and this depends on the type of error. For example, increasing age predicts a lower rate of insufficient detail errors but a higher rate of memory gap errors.

From the audience, John Abowd suggested that this information could be used to tailor interviewing styles to the demographics of the interviewee. For example, the interviewer could spend more time helping older respondents recall events or spend more time encouraging younger respondents to give more detail. Ana concluded by noting that her indicators for interview complexity are not perfect and the search for more and better indicators in available paradata continues.

*(Continued on page 6)*

# NCRN Spring 2016 Meeting (Continued)

## Session III

by Sylverie Herbert

### The Advantages and Disadvantages of Statistical Disclosure Limitation for Program Evaluation

**Ian Schmutte, Cornell University**

The project aims at answering two main questions: When is statistical disclosure limitation (SDL) ignorable? When it is not ignorable, is SDL known? Can SDL-aware analysis be conducted?

SDL is ignorable if the analysis is the same with published and observed data. When non-ignorable, SDL is known if the analysis of published data can be corrected for data alterations introduced by SDL.

Having these definitions in mind, two examples in which such issues of direct disclosure of attributes of inferential disclosure can arise have been presented: randomized response (randomly assign yes/no question, one is about violent crime, one is innocuous), top coding (published income is top coded). Possible mechanisms to prevent false discovery include: 1) develop the model design on synthetic data, 2) validate against confidential data, 3) prevent ex-post adjustment of the evaluation.

Finally, a model of SDL-aware analysis was presented with an application on regression discontinuity to look at local average treatment effect.

### Developing job linkages for the Health and Retirement Study

**Maggie Levenstein, University of Michigan**

The presentation presented a methodology, and some preliminary results of the efforts, for enhancing the Health and Retirement Study (HRS) with Census data.

The first bulk of the data is the retirement survey from Michigan University. The goal of the project is to include information on HRS respondents in the context of employers and co-workers, with the ultimate goal of enhancing the HRS public-use dataset.

It consists of:

- Developing a new data infrastructure
- Creating a Health and Retirement Study-Business Register (BR) crosswalk, which will enable getting new employer characteristics, such as productivity.

The starting point of the project is the 1992 HRS private sector jobs subset, in which the 10-digit phone number was used for the crosswalk. The HRS jobs were paired with BR establishments, generating 18.3 million pairs.

The second step consists of a "training set": at least 2 employees reviewed the pairs and assigned a score to them (match, likely match, no match). An interesting feature highlighted in the presentation was the higher disagreement in the "no match" category.

Finally, a model was developed, featuring a propensity for records from HRS to match records from the BR, through a logistic model.

The challenges pertaining to this analysis are as follows: what to do when the block does not include any high probability matches? What are the reasons for non-matches? Does the blocking strategy exclude the correct match or is this a feature of the model?

## Session IV

by Flavio Stanchi

### Crowdsourcing Metadata - Challenges and Outlook

**Lars Vilhuber, Cornell University**

The last presentation of the meetings was given by Lars Vilhuber on a project involving multiple researchers at Cornell University. The project focuses on the crowdsourcing of metadata — information about data — and its connections to the replicability of scientific articles.

*Lars Vilhuber, Cornell University*

Various articles, including a working paper by Kingi, Stanchi and Vilhuber, show how the majority of research in (but not limited to) Economics is still hard or impossible to reproduce, given the information and the data available to the replicators. Validation of scientific research is a crucial step in the production of knowledge and requires information about datasets, programs and procedures used by the researchers. To this end, the Cornell NCRN node has been working on CED2AR, a user-friendly metadata curation software that will allow researchers to create and augment the information available on data they routinely make use of.

*(Continued on page 9)*

# Spring 2016 (Continued)

*(from page 8)*

The software relies on open standards, namely the Data Documentation Initiative (DDI) schema, and allows for math formulas using the LaTeX syntax. It is web based, and makes use of the academic network ORCID for user authentication.

The crowdsourced metadata will be part of a cycle that also features internal metadata — which may include information that cannot be disclosed to the public — and official metadata — which contains all the information that can be disclosed and that has been verified by the metadata curator. Version 2 of the software is authorized in the Census RDC, and in the future the creators are planning to add UTF-8 support for other languages, to increase the number of fields that describe the data, and to make the live server scalable in order to accommodate a multitude of users.

# Publications

***The following are the most recent additions to publications produced by the research nodes within NCRN. A comprehensive list can be found here. (http://www.ncrn.info/documents/bibliographies)***

Abowd, John M., Kevin L. McKinney, and Ian M. Schmutte. ***Modeling Endogenous Mobility in Wage Determination***. Cornell University Preprint 1813:40306, 2015, available at http://hdl.handle.net/1813/40306.

Abowd, John, and Ian Schmutte. ***Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods***. Cornell University Preprint 1813:39081, 2015, available at http://hdl.handle.net/1813/39081.

Manski, Charles. ***Communicating Uncertainty in Official Economic Statistics***. Northwestern University Preprint 1813:36323, 2014, available at http://hdl.handle.net/1813/36323.

Perry, Benjamin, Venkata Kambhampaty, Kyle Brumsted, Lars Vilhuber, and William Block. ***Collaborative Editing of DDI Metadata: The Latest from the CED2AR Project***. Cornell University Preprint 1813:38200, 2014, available at http://hdl.handle.net/1813/38200.

Perry, Benjamin, Venkata Kambhampaty, Kyle Brumsted, Lars Vilhuber, and William Block. ***Presentation: NADDI 2015: Crowdsourcing DDI Development: New Features from the CED2AR Project***. Cornell University Preprint 1813:40172, 2015, available at http://hdl.handle.net/1813/40172.

Sadosky, Peter, Anshumali Shrivastava, Megan Price, and Rebecca Steorts. ***Blocking Methods Applied to Casualty Records from the Syrian Conflict***, *ArXiv*. 1510.07714, 2015, available at http://arxiv.org/abs/1510.07714.

H. Shaefer, Luke. ***Introduction to The Survey of Income and Program Participation (SIPP)***. University of Michigan Preprint 1813:40169, 2015, available at http://hdl.handle.net/1813/40169.

Spielman, Seth, and David Folch. ***Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization***. University of Colorado at Boulder / University of Tennessee Preprint 1813:38121, 2014, available at http://hdl.handle.net/1813/38121.

Steorts, R., S. Ventura, M. Sadinle, S. E. Fienberg, and J. Domingo-Ferrer. "***A Comparison of Blocking Methods for Record Linkage***." In *Privacy in Statistical Databases*, 253-268. Vol. 8744. Springer, 2014, available at http://link.springer.com/chapter/10.1007/978-3-319-11257-2_20.

Wasi, Nada, and Aaron Flaaen. ***"Record Linkage using STATA: Pre-processing, Linking and Reviewing Utilities."*** *The Stata Journal* 15, no. 3 (2015): 1-15, available at http://www.stata-journal.com/article.html?article=dm0082.

*(Continued on page 10)*

# Presentations

**Noel Cressie**, University of Missouri node, co-presented a one-day short course in December (with A. Zammit Mangion) for National Institute for Applied Statistics Research, Sydney, Australia; "Spatio-temporal statistical modelling."

**Christopher Wikle**, University of Missouri, presented "Recent developments in nonlinear dynamic spatio-temporal models (an overview)" for the Department of Statistics at the University of Georgia in Athens, Georgia on January 14. He also gave a talk entitled, "Hierarchical spatio-temporal statistical methods for environmental, agricultural and federal statistics applications" at the ESRI Spatio-Temporal Statistics Summit in Redlands, California on February 18.

**Rebecca Steorts** and **Jerry Reiter**, both of Duke University, made presentations at the ENAR 2016 Spring Meeting in Austin, Texas, which was held March 6-9, 2016. Steorts' talk was "Bayesian Analysis of Complex Survey Data" and Reiter gave a talk entitled, "Dissecting Multiple Imputation from a Multi-phase Inference Perspective."

**Christopher Wikle**, University of Missouri, gave a short course entitled "An introduction to dynamical spatio-temporal models" at the Workshop on Bayesian Environmetrics at The Ohio State University on March 31.

**Lars Vilhuber**, Cornell University, presented at CASD Conference on April 6, "Vos données au coeur de la datascience" on the topic of "Quelques développements en cours aux USA et au Canada". The conference was held at the Muséum national d'histoire naturelle in Paris, France.

**J. Lee, B. Seloske, Ana Lucia Cordova Cazar, Adam Eck** and **Robert Belli**, all from University of Nebraska, presented a paper at the annual FedCasic Workshop in Suitland, Maryland on May 3. The paper was "Data Management and Analytic Use of Paradata: SIPP-EHC Audit Trails."

Several University of Nebraska researchers presented papers at the American Association for Public Opinion Research annual meeting in Austin, Texas May 11-15. **Jerry Timbrook, Jolene Smyth**, and **Kristen Olson**, presented a paper, " Why do Mobile Interviews Take Longer? A Behavior Coding Perspective." **Beth Cochran, Kristen Olson**, and **Jolene Smyth**, presented the paper "Interviewer Influence on Interviewer-Respondent Interaction During Battery Questions." **Amanda Ganshert, Kristen Olson**, and **Jolene Smyth**, presented a paper entitled," The Effects of Respondent and Question Characteristics on Respondent Behaviors," and **Jolene Smyth** and **Kristen Olson** also presented a paper, "Mismatches."

# Upcoming Events

## August 8-11, 2016 at U. Michigan

### Conducting Research Using The Survey Of Income And Program Participation (SIPP)

Lead Instructors: Lori Reeder & Holly Monti, US Census Bureau; Co-Organizers: Gary Benedetto, U.S. Census Bureau & H. Luke Shaefer, University of Michigan. With support from Lars Vilhuber, Cornell University

More info here.

## 2016 RDC Annual Research Conference

The 2016 RDC Conference will be held at the Texas A&M Campus, primarily at the Memorial Student Center.

The registration deadline is Wednesday, August 17, 2016. More info here.

# Recent NCRN Virtual Seminars

## March 2, 2016

## "The effect of question and questionnaire characteristics on interviewer and respondent behaviors in CATI surveys."

**Kristen Olsen** from University of Nebraska gave the following talk.

**Abstract:** In this paper, we evaluate the joint effects of question, respondent and interviewer characteristics on two proxy indicators of data quality - response time and question misreading - in a telephone survey. We include question features traditionally examined, such as the length of the question and format of response options, and features that are related to the layout and format of interviewer-administered questions. First, we examine how these question features affect the time to ask and answer survey questions and how different interviewers vary in their administration of these questions. Second, we investigate how choices in visual design features in particular, that is design features that require interviewer decisions, contribute to interviewer question misreading. These two measures of question time and question misreading are both proxies for the risk of measurement error in responses to survey questions.

To examine these questions, we use paradata and behavior codes from the Work and Leisure Today (n=450, AAPOR RR3=6.3%) survey and use cross-classified random effects models. Overall, more of the variation in both response time and question misreading is due to question characteristics compared to respondent or interviewer attributes. Additionally, we find that question characteristics related to necessary survey design features and respondent confusion are the primary predictors of response time, with little effect of visual design features of the question. Our results for question misreading show a different pattern. Characteristics related to task complexity and visual design significantly affect question misreading, with little contribution of necessary survey design features. We conclude with implications for survey practice.

## April 6, 2016

## Microclustering: When the Cluster Sizes Grow Sublinearly with the Data Set

**Beka Steorts**, Duke University, gave the following talk.

**Abstract:** Most generative models for clustering implicitly assume that the number of data points in each cluster grows linearly with the total number of data points. Finite mixture models, Dirichlet process mixture models, and Pitman--Yor process mixture models make this assumption, as do all other infinitely exchangeable clustering models. However, for some tasks, this assumption is undesirable. For example, when performing entity resolution, the size of each cluster is often unrelated to the size of the data set. Consequently, each cluster contains a negligible fraction of the total number of data points. Such tasks therefore require models that yield clusters whose sizes grow sublinearly with the size of the data set. We address this requirement by defining the *microclustering property* and introducing a new model that exhibits this property. We compare this model to several commonly used clustering models by checking model fit using real and simulated data sets.

The NCRN Newsletter is published quarterly by the NCRN Coordinating Office. You may reach us at:
275 Ives Hall,
Cornell University, Ithaca, NY 14853
(607) 255-2744
info@ncrn.info
www.ncrn.info
Jamie Nunnelly - Managing Editor
Mellisa Dora Colbeth - Assistant Editor