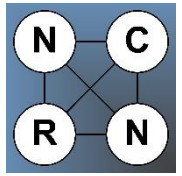
Has the NSF-Census Bureau Research Network Helped Improve the U.S. Statistical System?

JSM SPAIG Award Session

August 2018

Lars Vilhuber, Cornell University

Acknowledgements, Disclaimer



- NCRN the result of a huge collaboration
  - The National Science Foundation and the Census Bureau (the funders)
  - Many academic institutions
  - More than a score of principal and co-principal investigators

The conclusions reached in this presentation are not the responsibility of the National Science Foundation (NSF), the Census Bureau, or any of the institutions to which the authors belong.

What is the NSF-Census Bureau Research Network?

- The **NSF-Census Bureau Research Network (NCRN)** is a set of eight **research nodes**, conducting **interdisciplinary research and educational** activities on **methodological questions of interest and significance to the broader research community and the Federal Statistical System**, particularly the U.S. Census Bureau.
- The activities will be expected to **advance both fundamental and applied knowledge**.
- Nodes selected in an open competition by NSF, along with a **Coordinating Office**; awards cover 2011-2016 (some extensions).
- **http://www.ncrn.info**

NCRN Nodes

Eight nodes comprised of researchers conducting innovative, highly-interdisciplinary investigations of theory, methodology and computational tools of interest.

NCRN Coordinating Office

Carnegie-Mellon University

Cornell University

Duke University / National Institute of Statistical Sciences (NISS)

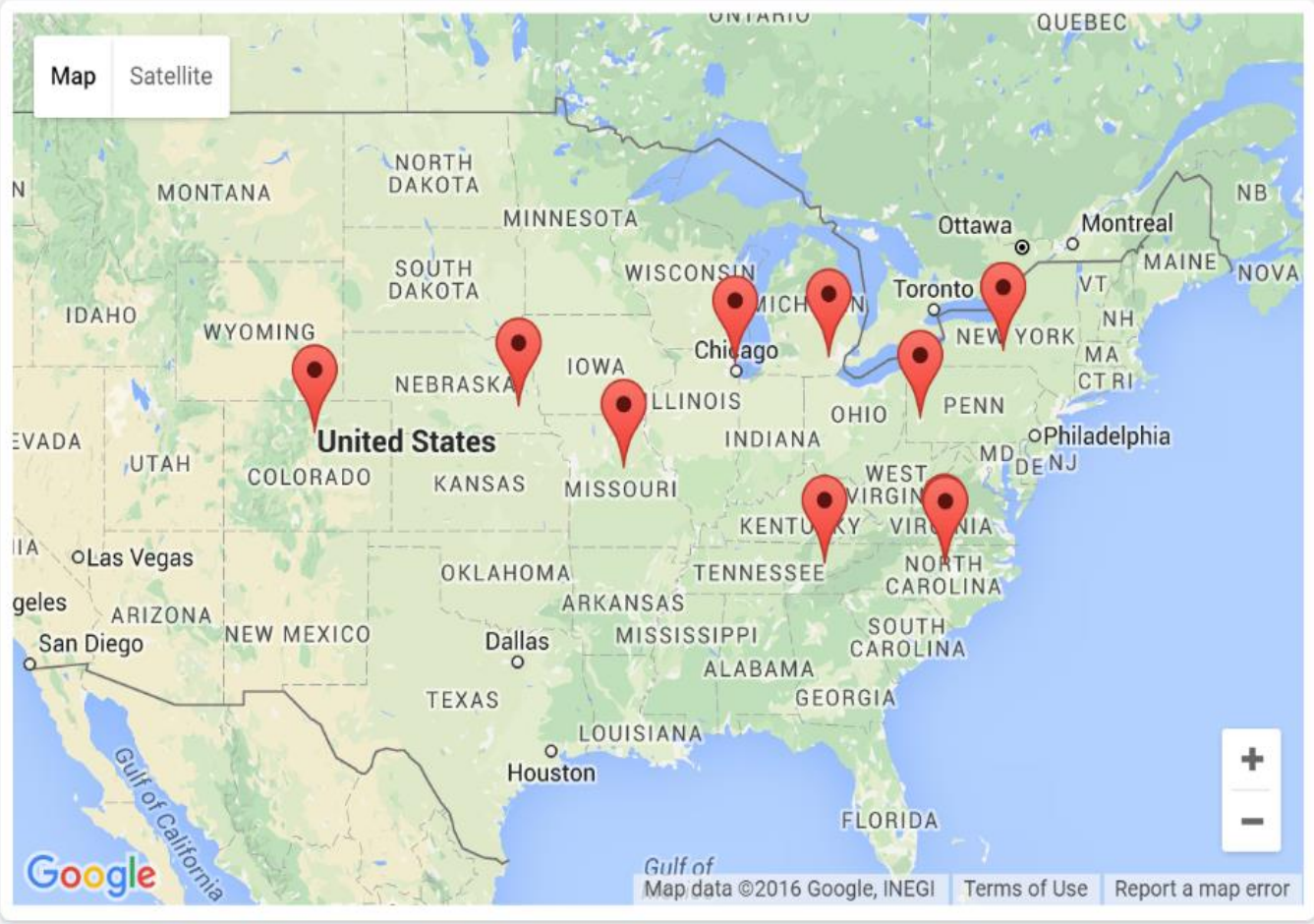
Northwestern University

University of Colorado at Boulder / University of Tennessee

University of Michigan

University of Missouri

University of Nebraska



Elements of the Presentation

- Diversity of the network: research outcomes.
- Education activities and outcomes.
- Collaborations: across nodes and with federal agencies.
- Lessons Learned.

Key Research Findings

- Based on more than 400 articles, papers.
- Organized into **six topics**:

Improving census and survey data-quality and data collection methods

Using spatial and spatio-temporal statistical modeling to improve estimates

Using alternative sources of data

Assessing data-cost and data-quality tradeoffs

Protecting privacy and confidentiality by improving disclosure avoidance

Statistically combining multiple sources

Key Research Findings

- Based on more than 400 articles, papers.
-

Improving census and survey <u>data-quality</u> and <u>data collection</u> methods

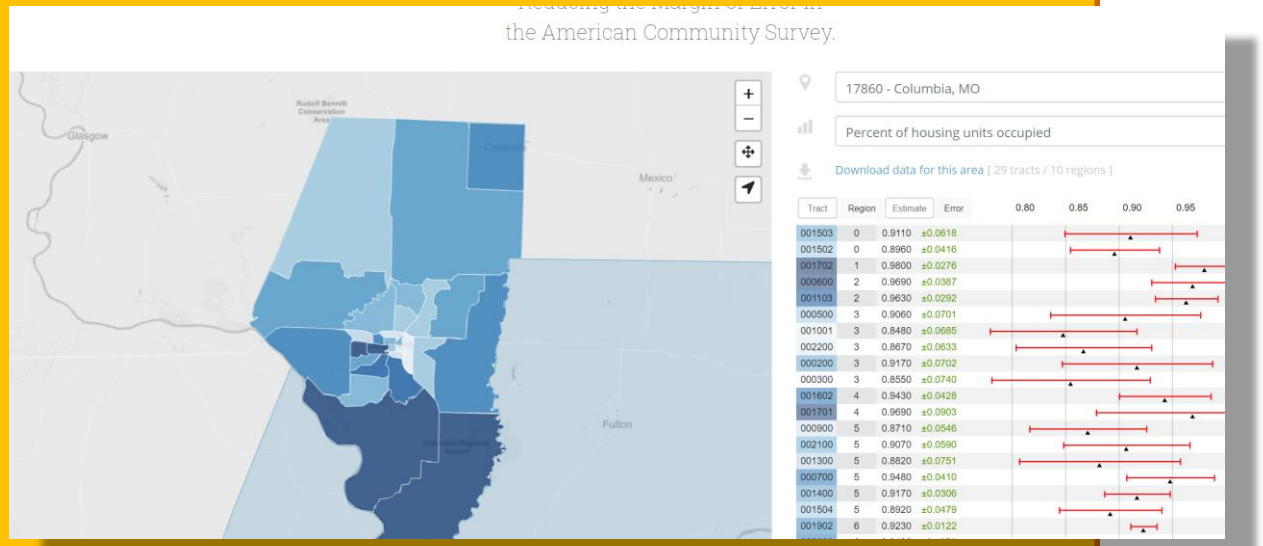
- **Design of the questions -> survey data quality indicators, participant behavior (Nebraska)** (Olson et al, Olson and Smyth 2015, Timbrook et al)
- **Flexible engine for multiple imputation for continuous and categorical variables (Duke)** (White et al 2018, Sadinle and Reiter 2017, 2018, Hu et al 2018).
- **Bayesian approaches for Fellegi-Holt stochastic editing (Duke)** (Kim et al 2015, Manrique-Vallier and Reiter 2018)

Key Research Findings

Using spatial and spatio-temporal statistical modeling to improve estimates

- **New spatial techniques for aggregating and disaggregating the basic ACS estimates geographically and over time (Colorado-Tennessee, Missouri)** (Bradley et al 2015b, Folch and Spielman 2014, Spielman and Folch 2015, Spielman and Singleton 2015)

- **Improved visualization techniques that account for uncertainty (Colorado-Tennessee, Missouri).** (Jurjevich et al 2018, Raim et al 2017)



Key Research Findings

Using alternative sources of data

- Generalizations of **Fellegi-Sunter** linkages
  (Carnegie Mellon, Duke)
  (Sadinle 2017, Sadinle and Fienberg 2013)

- Feasibility of **large-scale linkages** (decennial census, administrative data, surveys)
  (CMU, Duke, Michigan, Cornell)
  (Wasi and Flaaen 2015, Steorts et al 2016)

- Using "**non-designed" data** to create new indicators and measures at lower cost, greater frequency, more geographic detail
  (Michigan)
  (Gelman et al 2014 2015, Antenucci et al 2013, 2014)

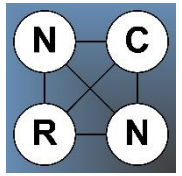
Key Research Findings

Assessing data-<u>cost</u> and data-<u>quality</u> <u>tradeoffs</u>

- understand the **<u>value of statistics</u>** produced,
- compare value to cost in order to **<u>guide rational setting of statistical priorities</u>**,
- better **<u>communicate</u>** the value of data programs to those who set their budgets.

Extensions and applications of statistical decision theory, distinguishing transitory statistical uncertainty, permanent statistical uncertainty, and conceptual uncertainty (Northwestern)
(Manski 2015, Seeskin and Spencer 2015, 2018)

Key Research Findings



Protecting <u>privacy</u> and <u>confidentiality</u> by improving disclosure avoidance

- **Trade-offs associated with the privacy** and the sharing of personal data, value of statistics; consumers understanding thereof (Carnegie Mellon, Cornell) (Abowd & Schmutte 2017, Acquisti et al 2015, 2016)
- Quantifying **disclosure risks** associated with **large-scale record linkage** (Duke) (Kim et al 2016)
- Use of **synthetic data** as a disclosure avoidance technique (Cornell, Duke) (Kinney et al. 2014; Miranda and Vilhuber 2016; Chen et al 2016)
- Improving **disclosure avoidance** methods for **spatially correlated** data (Duke, Missouri) (Quick et al 2015, 2016)

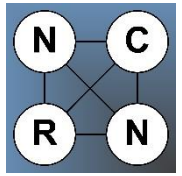
Key Research Findings

Statistically combining <u>multiple sources</u>

- hierarchical Bayesian approach using geography and/or time to enhance model estimation and prediction (Missouri)

(Bradley et al 2015b, 2016a)

Broad diversity



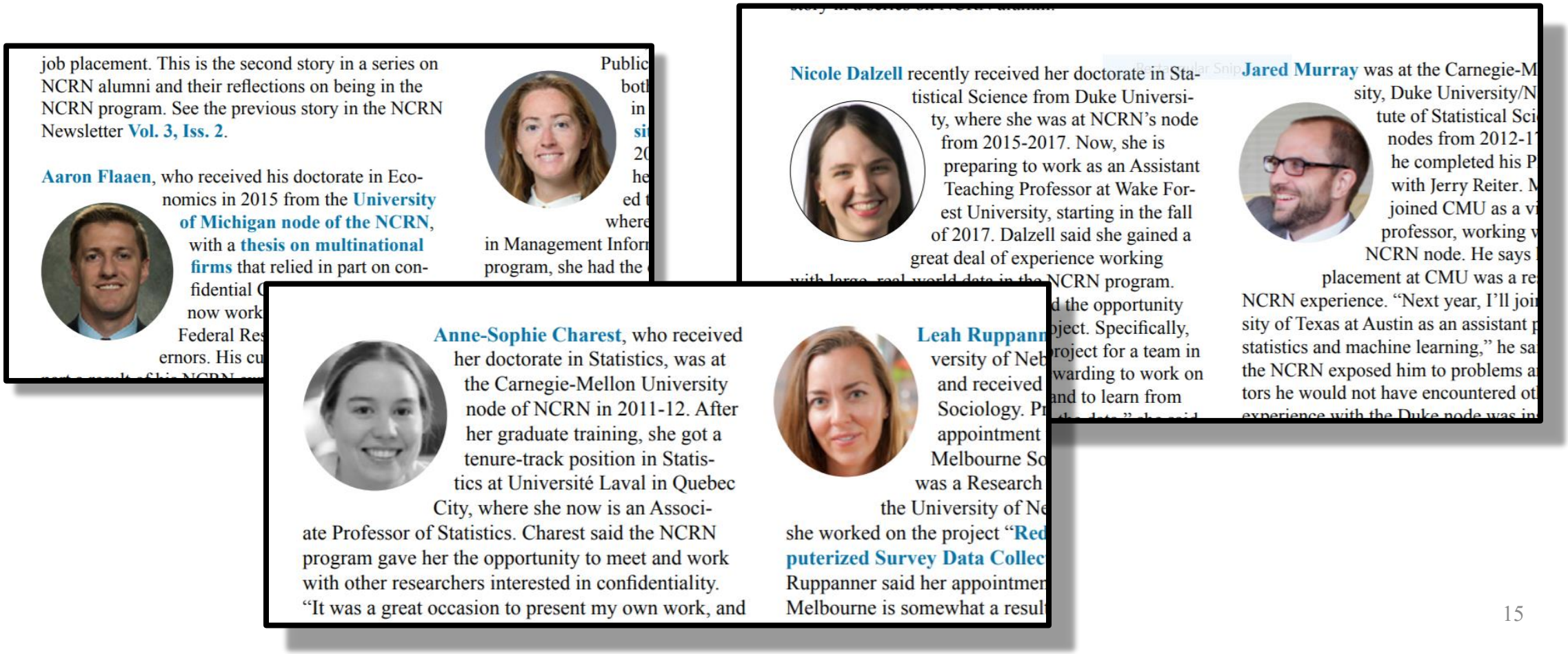
- 400+ papers in six themes
- 30+ researchers
  - Statisticians (survey, spatio-temporal)
  - Economists
  - Geographers
  - Information scientists
  - Computer scientists

How to incentivize collaboration?

Educational activities

Educational outcomes

- **Graduate students** and **post-doctoral fellows** mentored within the network.



job placement. This is the second story in a series on NCRN alumni and their reflections on being in the NCRN program. See the previous story in the NCRN Newsletter **Vol. 3, Iss. 2**.

Aaron Flaaen, who received his doctorate in Economics in 2015 from the **University of Michigan node of the NCRN**, with a **thesis on multinational firms** that relied in part on con-

Nicole Dalzell recently received her doctorate in Statistical Science from Duke University, where she was at NCRN's node from 2015-2017. Now, she is preparing to work as an Assistant Teaching Professor at Wake Forest University, starting in the fall of 2017. Dalzell said she gained a great deal of experience working

Jared Murray was at the Carnegie-M

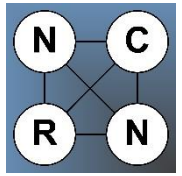
Anne-Sophie Charest, who received her doctorate in Statistics, was at the Carnegie-Mellon University node of NCRN in 2011-12. After her graduate training, she got a tenure-track position in Statistics at Université Laval in Quebec City, where she now is an Associate Professor of Statistics. Charest said the NCRN program gave her the opportunity to meet and work with other researchers interested in confidentiality. "It was a great occasion to present my own work, and

Leah Ruppann

Educational outcomes: Placement

- Aim was to "groom" students for a **career in government**
- Most graduate students and post-docs went into **academia**
- Why not government?
  - Salaries are not always competitive (but in choice set)
  - Citizenship a problem
  - Some did: Federal Reserve, Census Bureau
  - "*My work with NCRN [...] has made me more interested in the Census as a potential employer,*"

Educational outcomes

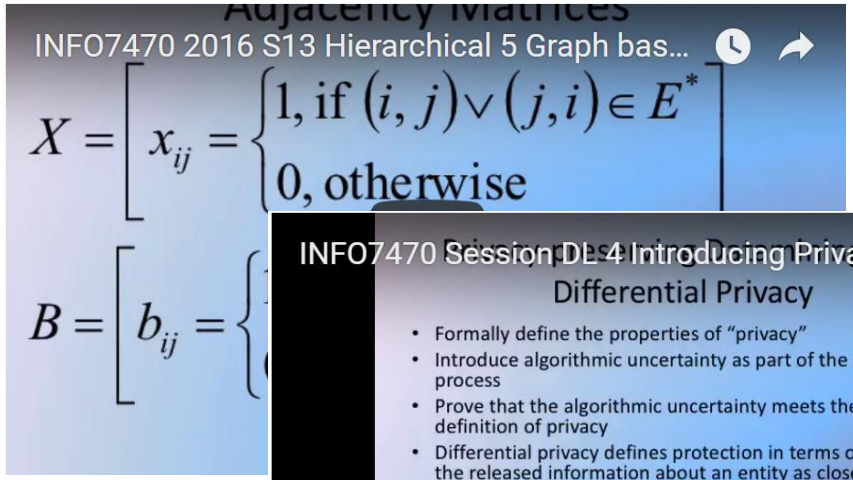


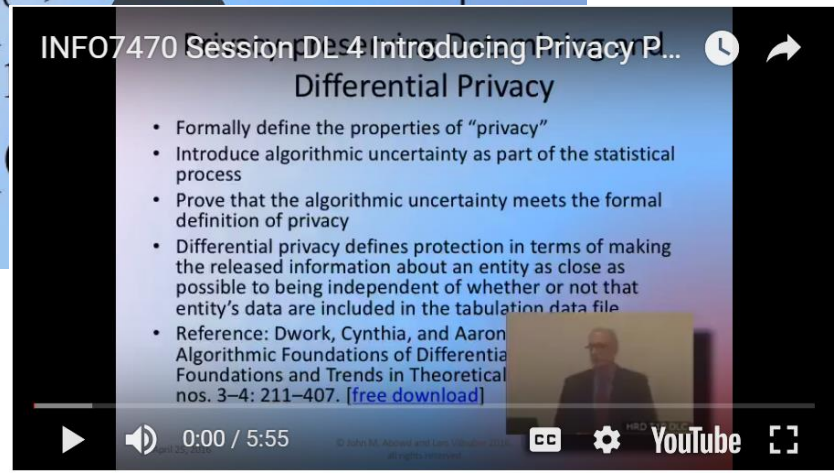
- More than **700 students** have taken undergraduate and graduate-level **classes**, as well as **short courses**

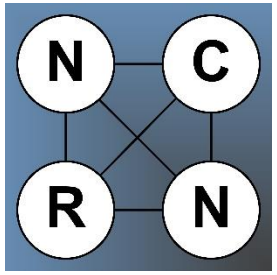
Educational outcomes

Popular: "Understanding Social and Economic Data," led by the Cornell node,

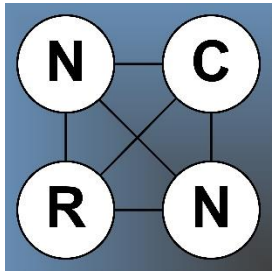
- taught as a **hybrid MOOC**/ live-distance-learning course.
- Up to a dozen remote sites each time
- Around 100 participants each time
- Government (Census, Fed, NSF), Academia (FSRDC nodes)
- Guest speakers on history, geographic concepts, former NSO heads





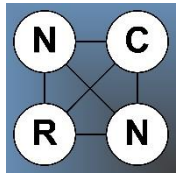


How to get all these folks to talk to each other?



It’s a network

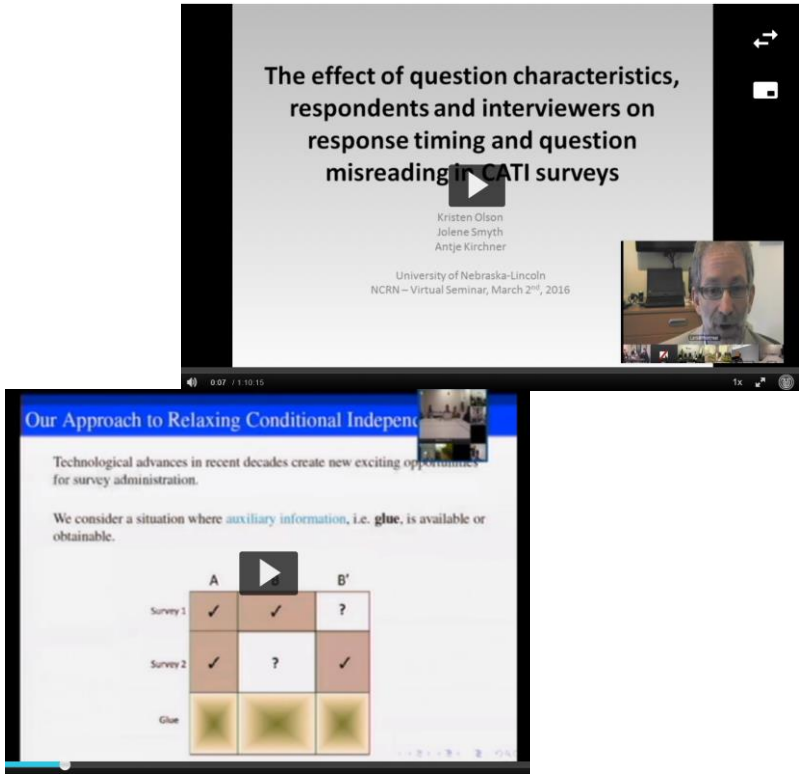
Networking opportunity



- **Meetings and Workshops**
  - Started in 2012 at Duke University
  - PI-meetings with the Census Bureau
  - Subsequent bi-annual meetings in New York, in Washington DC
  - Participants from nodes, Census Bureau, other statistical agencies
- Lessons learned
  - Location, location, location
  - Arm-twisting

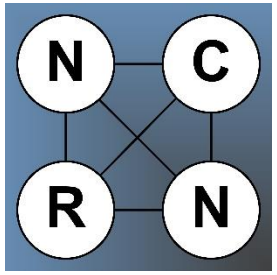
Presentation opportunities

- **Video-enabled Virtual Seminar**
  - Each node presented at least once
  - Monthly during academic semester
  - Opportunity to ask questions across the network



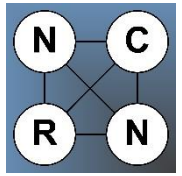
Surfacing information

- **Network website** (ncrn.info)
  - Aggregating and original news posts
  - Aggregating lists of publications
  - Announcing meetings, workshops, virtual seminars



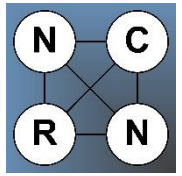
Some outcomes

Collaboration (1): Inter-nodal collaborations

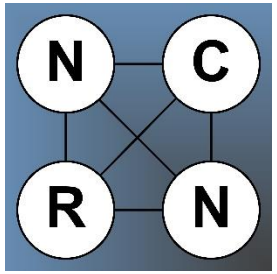


- Duke – Missouri, Duke – Carnegie Mellon
- Duke – Cornell
- Missouri and most of the other nodes at the 2016 “Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics”;
- Michigan, Carnegie Mellon, Cornell, and Duke
- Michigan – Cornell
- Michigan – Duke
- Nebraska – Carnegie Mellon
- Missouri – Cornell

Collaborations with FSS

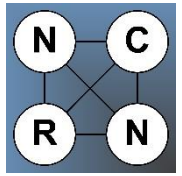


- Naturally with Census Bureau
- Less successfully with other agencies
  - Lack of funding?
  - Not for lack of trying



Lessons learned

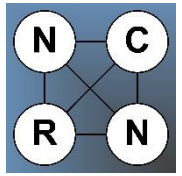
Lessons learned - 1



Coordinating a diverse set of academics and agency researchers is hard

Requires idiosyncratic skill and serendipity

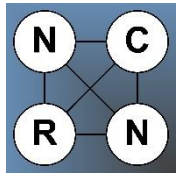
Lessons learned - 2



In order to
get (more) **useful research outcomes**,
(better) **coordination**
between **academic partners** and
interested agencies is critical

Requires **opportunities** and (frequent)
communication

Lessons learned - 3

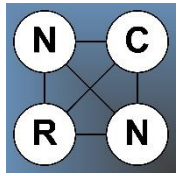


In order to
get (more) **technology transfer**,

(closer) **collaboration**
between **academics** and **government**
researchers is critical

Requires (hands-on) **opportunities** and
(active) **encouragement**

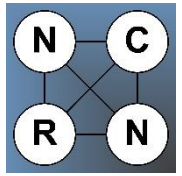
Lessons learned – 4



Access to **confidential data**
is a **productivity-enhancer**

Requires (early) **planning,**
FSRDCs are key tool

Lessons learned – 5

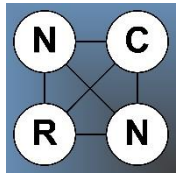


Hiring (embedding) students

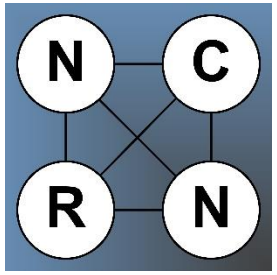
is a **technology-transfer vector**

Limited by federal **hiring rules**

Concluding Remarks



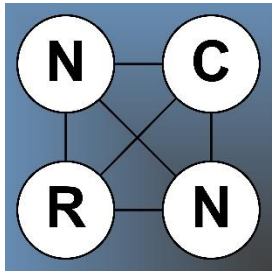
- Challenges of managing a network comprising researchers from many disciplines spread across both academia and government.
- Difficulty of bridging the gap between theory and practice, and the various gaps in expectations between academic researchers and government practitioners.
- Keep the network participants talking with one another and the sponsoring agencies; the NCRN's semi-annual meetings were more frequent than those of many other networks, and hence they may have led to a faster convergence of ideas and language.
- NSF often recognizes the long-term aspect of creating effective collaborations when creating centers of excellence, but these are not typically initiated in collaboration with a non-grant-making agency like the Census Bureau, and the budgetary intricacies of an NSF-agency collaboration are challenging.



Thank you

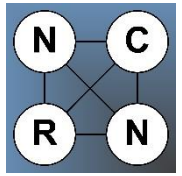
For contact: lars.vilhuber@cornell.edu

For information: https://www.ncrn.info



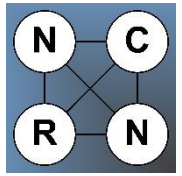
Extra slides

RF: Improving census and survey data-quality and data collection methods



- Design of the questions plays a greater role in predicting survey data quality indicators and interviewer and respondent behaviors during a survey than characteristics of interviewers or respondents (Nebraska).
- Developed a model that blends mixtures of multinomial distributions with mixtures of multivariate normal regression models to create a flexible engine for multiple imputation or missing multivariate continuous and categorical variables (Duke).
- Developed methods to improve on Fellegi-Holt by using Bayesian approaches to allow stochastic editing to create multiply imputed, plausible datasets, based on hierarchical models (Duke).

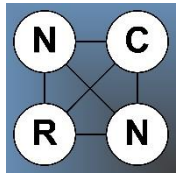
RF: Using alternative sources of data



Record linkage is a critical component of the efforts to reduce census costs and, potentially, to improve accuracy.

- Record linkage solutions provide conceptual generalizations of Fellegi-Sunter method that are computationally feasible for application at the scale of the decennial census, while acknowledging and propagating the uncertainty from the matching process into subsequent analyses (Carnegie Mellon, Duke).
- Probabilistic linkage of survey-identified employers to Census Business Register in ways that address the complexity and benefits of linking household and business data (Michigan).
- Investigations of whether "non-designed" data (e.g., account data, social-media data) can provide useful indicators and checks on traditional time series, or produce measures at lower cost, greater frequency, more geographic detail, or in conjunction with survey data to reduce respondent burden (Michigan).

RF: Protecting privacy and confidentiality by improving disclosure avoidance

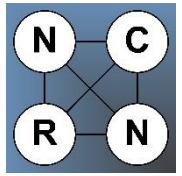


- Connecting theoretical and empirical research on the economics of privacy by focusing on the economic value and consequences of protecting and disclosing personal information, and on consumers' understanding and decisions regarding the trade-offs associated with the privacy and the sharing of personal data (Carnegie Mellon, Cornell).
- Quantifying the disclosure risks associated with large-scale record linkage (Duke).
- Extending prior work on the use of synthetic data as a disclosure avoidance technique (Cornell, Duke).
- Improving disclosure avoidance methods for spatially correlated data (Duke, Missouri).

RF: Using spatial and spatio-temporal statistical modeling to improve estimates

- New spatial techniques for aggregating and disaggregating the basic ACS estimates geographically and over time (Colorado-Tennessee).
  - Multivariate statistical clustering to group demographically similar census tracts into latent classes, along with implementing software.
- Improved visualization techniques that account for uncertainty (Missouri).

RF: Assessing data-cost and data-quality tradeoffs

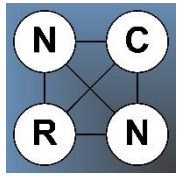


The key problem for statistical agencies is multifaceted:

1. understand the value of the statistics they produce,
2. compare value to cost in order to guide rational setting of statistical priorities,
3. increase value for given cost, and
4. better communicate the value of data programs to those who set their budgets.

- Extensions and applications of statistical decision theory, including cost-benefit analysis, to attack such questions (Northwestern).
- Distinguishing transitory statistical uncertainty, permanent statistical uncertainty, and conceptual uncertainty (Northwestern).

RF: Combining information from multiple sources



While record linkage attempts to combine data sources in a way that matches information from multiple sources, better estimates can be made by combining information from multiple sources by modeling.

- Developed a hierarchical Bayesian approach using geography and/or time to enhance model estimation and prediction, in effect creating powerful spatio-temporal mixed effects models that include Fay-Herriot models as a special case (Missouri).