Cornell Node of the NSF-Census Research Network - Annual Report to NSF for 2016

Lars Vilhuber and William Block

March 6, 2019

Preview of Award 1131848 - Annual Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1131848
Project Title:	NCRN-MN: Cornell Census-NSF Research Node: Integrated Research Support, Training and Data Documentation
PD/PI Name:	Lars Vilhuber, Principal Investigator William C Block, Co-Principal Investigator
Recipient Organization:	Cornell University
Project/Grant Period:	10/01/2011 - 09/30/2017
Reporting Period:	10/01/2015 - 09/30/2016
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	N/A

Accomplishments

*** What are the major goals of the project?**

As part of the Cornell node's activities, we are building a Comprehensive Extensible Data Documentation and Access Repository (CED²AR) designed to improve the documentation and discoverability of both public and restricted data from the federal statistical system. The CED²AR will be based upon leading metadata standards such as the Data Documentation Initiative (DDI) and Statistical Data and Metadata eXchange (SDMX) and be flexibly designed to ingest documentation from a variety of source files.

We are also developing High Performance Logistic Regression Methods for Data Edits and Imputation for (a) multiple response variables (Census example: race/ethnicity coding) as well as (b) incompletely coded links (Census example: unit-to-worker imputation).

Filename	Description	Uploaded By	Uploaded On

More recently, we have tackled the problem of efficient trade-offs between data quality and confidentiality (privacy loss) using techniques from economics, i.e., a formal production possibilities frontier (PPF). We consider situations where data quality will be inefficiently under-supplied. Results show that government data custodians should publish more accurate statistics with weaker privacy guarantees than would occur with purely private data publishing.

Finally, we are teaching a multi-site distance learning class on "Social and Economic Data" (INFO 7470). The course is designed to teach students basic and advanced techniques for acquiring and transforming raw information into social and economic data. The course is particularly aimed at American Ph.D. students from multiple fields (economics, political science, demography, sociology, etc.) who are interested in using confidential U.S. Census Bureau data, and the confidential data of other American statistical agencies that cooperate with the Census Bureau. We cover the legal, statistical, computing, and social science aspects of the data "production" process. More information is available at the course website http://www.vrdc.cornell.edu/info7470/.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities: INFO7470 was taught to over 100 students at 11 participating sites (including Cornell and the U.S. Census Bureau). CED2AR was used to instruct users of synthetic data, and additional datasets were created and are hosted at our website. It is also in discussions with the Census Bureau's SIPP branch to support user-contributions to early-release public-use files, and a nascent collaboration with ICPSR, in additionl to existing collaborations with the Roper Center. Research into the use of confidential data continues, with presentations and publications into the use of synthetic data (one way of protecting confidentiality) and implications for other uses.

Specific Objectives:

Significant Results:

Key outcomes or Other achievements:

*** What opportunities for training and professional development has the project provided?**

6 undergraduates participated in the Spring term, and an additional 4 during the summer, to support replicability research. 2 graduate students have participated this year on research into metadata. Over 100 students participated in the INFO7470 class.

*** How have the results been disseminated to communities of interest?**

The CED2AR software is available for download as binary software for both servers and desktops. Source code is posted on Github. Publications are listed elsewhere in this report. Several presentations of the work at scientific conferences have been given. INFO7470 materials were broadcast using a combination of EdX online tools and web posting. The recorded sessions will be made available in the future on Youtube. All papers and INFO7470 materials (presentations, videos), as well as other presentations, are made available on properly curated document archives at http://ecommons.cornell.edu.

*** What do you plan to do during the next reporting period to accomplish the goals?**

We continue software development on CED2AR, making it robust to re-use in other contexts, such as ICPSR, Roper, or other communities. We will be teaching INFO7470 in a shortened version. We will be working with the Census Bureau to (a) create an infrastructure to support crowd-sourcing of metadata (b) to migrate such infrastructure to a Census Bureau-maintained infrastructure.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
NCRN-Cornell-Presentations-2015-2016.pdf	List of presentations made between Oct 1, 2015 and Sept 30, 2016	Lars Vilhuber	10/07/2016

Products

Books

Book Chapters

Abowd, John M. and Vilhuber, Lars and Block, William (2012). A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs. *Privacy in Statistical Databases* 7556. Domingo-Ferrer, Josep and Tinnirello, Ilenia. Springer Berlin Heidelberg. 216-225. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/978-3-642-33627-0_17.

Drechsler, J\"org and Vilhuber, Lars (2014). Synthetic Longitudinal Business Databases for International Comparisons. *Privacy in Statistical Databases* 8744. Domingo-Ferrer, Josep. Springer International Publishing. 243-252. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/978-3-319-11257-2_19.

Miranda, Javier and Vilhuber, Lars (2014). Using Partially Synthetic Data to Replace Suppression in the Business Dynamics Statistics: Early Results. *Privacy in Statistical Databases* 8744. Domingo-Ferrer, Josep. Springer International Publishing. 232-242. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/978-3-319-11257-2_18.

Ping Li and Art Owen and Cun-Hui Zhang (2012). One Permutation Hashing. *Advances in Neural Information Processing Systems 25* P. Bartlett and F.C.N. Pereira and C.J.C. Burges and L. Bottou and K.Q. Weinberger. 3122--3130. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://papers.nips.cc/paper/4778-one-permutation-hashing.

Ping Li and Cun-Hui Zhang (2012). Entropy Estimations Using Correlated Symmetric Stable Random Projections. *Advances in Neural Information Processing Systems 25* P. Bartlett and F.C.N. Pereira and C.J.C. Burges and L. Bottou and K.Q. Weinberger. 3185--3193. Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://papers.nips.cc/paper/4667-entropy-estimations-using-correlated-symmetric-stable-random-projections.

Inventions

Journals or Juried Conference Papers

Anshumali Shrivastava and Ping Li (2014). Graph Kernels via Functional Embedding. *CoRR*. abs/140 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://arxiv.org/abs/1404.5214

Carl Lagoze and William C. Block and Jeremy Williams and John M. Abowd and Lars Vilhuber (2013). Data Management of Confidential Data. *International Journal of Digital Curation*. 8 (1), 265-278. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.2218/ijdc.v8i1.259

Javier Miranda and Lars Vilhuber (2016). Using partially synthetic microdata to protect sensitive cells in business statistics. *Statistical Journal of the IAOS*. 32 (1), 69--80. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/SJI-160963

John M. Abowd and Ian Schmutte (2015). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*. Spring . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 00072303

John M. Abowd and Kevin L. McKinney (2016). Noise infusion as a confidentiality protection measure for graph-based statistics. *Statistical Journal of the IAOS*. 32 (1), 127--135. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/SJI-160958

Lars Vilhuber and John M. Abowd and Jerome P. Reiter (2016). Synthetic establishment microdata around the world. *Statistical Journal of the IAOS*. 32 (1), 65--68. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.3233/SJI-160964

Radhendushka Srivastava and Ping Li and Debasis Sengupta (2012). Testing for Membership to the IFRA and the NBU Classes of Distributions. *Journal of Machine Learning Research - Proceedings Track for the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*. 22 1099-1107. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://www.jmlr.org/proceedings/papers/v22/srivastava12.html

Schneider, Matthew J. and Abowd, John M. (2015). A new method for protecting interrelated time series with Bayesian prior distributions and synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. n/a--n/a. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1111/rssa.12100

Licenses

Other Conference Presentations / Papers

John Abowd and Kevin McKinney and Nellie Zhao (2015). *Analyzing Earnings Inequality in the United States: Trends from Longitudinally Linked Employer-Employee Data (Presentation)*. Federal Statistical Research Data Center Annual Conference. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Miranda, Javier and Vilhuber, Lars (2015). *Assessing the Data Quality of Public Use Tabulations Produced from Synthetic Data: Synthetic Business Dynamics Statistics (Presentation)*. Joint Statistical Meetings (JSM). . Status = OTHER; Acknowledgement of Federal Support = Yes

Shrivastava, Anshumali and Li, Ping (2013). *Beyond Pairwise: Provably Fast Algorithms for Approximate k-Way Similarity Search*. Advances in Neural Information Processing Systems 26. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Ping Li and John Abowd (2014). *Boosting Algorithms for Edit and Imputation of Multiple-response Variables (Presentation only)*. Federal Committee on Statistical Methodology Research Conference. . Status = OTHER; Acknowledgement of Federal Support = Yes

Carl Lagoze and Lars Vilhuber and Jeremy Williams and Benjamin Perry and William C. Block (2014). *CED  AR: The Comprehensive Extensible Data Documentation and Access Repository*. ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014). London, United Kingdom. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Benjamin Perry and Venkata Kambhampaty and Kyle Brumsted and Lars Vilhuber and William C. Block (2014). *Collaborative Editing and Versioning of DDI Metadata: The Latest from Cornell's NCRN CED²AR Software (Presentation only)*. 6th Annual European DDI User Conference (EDDI). . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Benjamin Perry and Venkata Kambhampaty and Kyle Brumsted and Lars Vilhuber and William C. Block (2015). *Crowdsourcing DDI Development: New Features from the CED2AR Project (Presentation only)*. North American Data Documentation Initiative Conference (NADDI). . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Lagoze, Carl and Willliams, Jeremy and Vilhuber, Lars (2013). *Encoding Provenance Metadata for Social Science Datasets*. Metadata and Semantics Research. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Carl Lagoze and William C. Block and Jeremy Williams and Lars Vilhuber (2013). *Encoding Provenance of Social Science Data: Integrating PROV with {DDI}*. 5th Annual European DDI User Conference. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Li, Ping and Zhang, Cun-Hui (2013). *Exact Sparse Recovery with L0 Projections*. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Xu Sun and Anshumali Shrivastava and Ping Li (2012). *Fast Multi-task Learning for Query Spelling Correction*. The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012) . . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Anshumali Shrivastava and Ping Li (2012). *Fast Near Neighbor Search in High-Dimensional Binary Data*. The European Conference on Machine Learning (ECML 2012). . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Ping Li and Anshumali Shrivastava and Arnd Christian K{\"o}nig (2012). *GPU-based minwise hashing: GPU-based minwise hashing*. Proceedings of the 21st World Wide Web Conference (WWW 2012) (Companion Volume). . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Anshumali Shrivastava and Ping Li (2014). *In Defense of MinHash Over SimHash*. Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS). Reykjavik, Iceland. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Benjamin Perry and Venkata Kambhampaty and Lars Vilhuber and William C. Block (2015). *Linking DDI to the Semantic Web (Poster)*. International Association for Social Science Information Services and Technology (IASSIST). . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Xu Sun and Anshumali Shrivastava and Ping Li (2012). *Query spelling correction using multi-task learning*. Proceedings of the 21st World Wide Web Conference (WWW 2012)(Companion Volume). . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Drechsler, Jörg and Vilhuber, Lars (2015). *Synthetic Longitudinal Business Databases for International Comparisons (Presentation)*. Joint Statistical Meetings (JSM). . Status = OTHER; Acknowledgement of Federal Support = No

John Abowd and Andrew Green and Kevin McKinney and Lars Vilhuber (2015). *Total Variability Measures for Selected Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in OnTheMap(Presentation)*. Federal Statistical Research Data Center Annual Conference. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Ping Li and Anshumali Shrivastava and K\"onig, Arnd Christian (2013). *b-Bit Minwise Hashing in Practice*. Internetware 2013. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Other Products

Other Publications

Patents

Technologies or Techniques

Thesis/Dissertations
Shrivastava, Anshumali. *Probabilistic Hashing Techniques for Big Data*. (2015). Cornell University. Acknowledgement of Federal Support = Yes

Websites
INFO7470 website
https://www.vrdc.cornell.edu/info747x/

The website serves as a first-point of contact for students of the class. The actual videos and exercises are hosted on the edX platform (login required, no cost).

“Understanding Social and Economic Data” (aka INFO7470) is designed to provide students a detailed overview of the US federal statistical system, where data comes from and how it can be used for research. The course also aims to teach students basic and advanced techniques for acquiring and transforming raw information into social and economic data. More information can be found in the Summary. The course is taught as a mixture of self-guided online videos (MOOC-style) together with in-classroom discussions of the material. Students from multiple universities, scattered across the country and participating via videoconference, and from multiple domains (economics, demography, geography, statistics) contribute to the discussion.

Main information site about project
https://www.ncrn.cornell.edu/

All the information about the node is catalogued on this website. It links to other websites maintained by the node, where appropriate.

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Vilhuber, Lars	PD/PI	3

Name	Most Senior Project Role	Nearest Person Month Worked
Block, William	Co PD/PI	1
Lagoze, Carl	Faculty	2
Barker, Brandon	Other Professional	3
Brumsted, Kyle	Other Professional	3
Kambhampaty, Venkata	Other Professional	4
Perry, Benjamin	Other Professional	3
Brown, Warren	Staff Scientist (doctoral level)	2
Edwards, Anne	Staff Scientist (doctoral level)	0
Herbert, Sylverie	Graduate Student (research assistant)	5
Sexton, William	Graduate Student (research assistant)	1
Shrivastava, Anshumali	Graduate Student (research assistant)	0
Stanchi, Flavio	Graduate Student (research assistant)	5

Full details of individuals who have worked on the project:

Lars Vilhuber
Email: lars.vilhuber@cornell.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 3

Contribution to the Project: Managing PI, contributed to confidentiality work, metadata work, supervision of graduate students, software development oversight, conference presentations.

Funding Support: This grant

International Collaboration: No
International Travel: No

William C Block
Email: block@cornell.edu
Most Senior Project Role: Co PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Co-PI, metadata work and direction, supervision of software developers, conference presentations and outreach.

Funding Support: This grant and Cornell University.

International Collaboration: No
International Travel: Yes, Norway - 0 years, 0 months, 5 days; Canada - 0 years, 0 months, 2 days

Carl Lagoze
Email: clagoze@umich.edu
Most Senior Project Role: Faculty
Nearest Person Month Worked: 2

Contribution to the Project: Metadata, Provenance expertise

Funding Support: This grant.

International Collaboration: No
International Travel: No

Brandon Barker
Email: beb82@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 3

Contribution to the Project: Working on CED2AR software

Funding Support: This grant

International Collaboration: No
International Travel: No

Kyle Brumsted
Email: kjb245@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 3

Contribution to the Project: Software development

Funding Support: This grant.

International Collaboration: No
International Travel: No

Venkata Kambhampaty
Email: vkambhampaty@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 4

Contribution to the Project: Software development

Funding Support: This grant

International Collaboration: No
International Travel: No

Benjamin Perry
Email: bap63@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 3

Contribution to the Project: Programming

Funding Support: This grant

International Collaboration: No
International Travel: No

Warren Brown
Email: warren.brown@cornell.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 2

Contribution to the Project: Expertise on ACS

Funding Support: NSF (this grant)

International Collaboration: No
International Travel: No

Anne Michelle Edwards
Email: ame87@cornell.edu
Most Senior Project Role: Staff Scientist (doctoral level)
Nearest Person Month Worked: 0

Contribution to the Project: metadata expertise

Funding Support: Cornell University

International Collaboration: No
International Travel: No

Sylverie Herbert
Email: sh2258@cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 5

Contribution to the Project: Assistance in creating/editing/improving metadata based on available data outside the Census firewall, assistance in preparing INFO7470

Funding Support: This grant.

International Collaboration: No
International Travel: No

William Sexton
Email: wns32@cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 1

Contribution to the Project: Assistance on confidentiality research

Funding Support: This grant.

International Collaboration: No
International Travel: No

Name	Type of Partner Organization	Location

Anshumali Shrivastava
Email: ansh@cs.cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 0

Contribution to the Project: He worked on developing a min search engine for metadata based on modern natural language processing techniques. He also worked on developing novel algorithms for graph data representations as well as large-scale statistics computations

Funding Support: NSF-EAGER 1249316 NSF-DMS 0808864 This grant.

International Collaboration: No
International Travel: No

Flavio Stanchi
Email: fs379@cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 5

Contribution to the Project: Assistance in creating/editing/improving metadata based on available data outside the Census firewall

Funding Support: No other.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
ICPSR	Other Nonprofits	Ann Arbor, MI
Roper Center	Academic Institution	Ithaca, NY
US Census Bureau	Other Organizations (foreign or domestic)	Washington, DC
University of Michigan	Academic Institution	Ann Arbor, Michigan

Full details of organizations that have been involved as partners:

ICPSR

Organization Type: Other Nonprofits
Organization Location: Ann Arbor, MI

Partner's Contribution to the Project:
In-Kind Support

More Detail on Partner and Contribution: We have had metadata contributions and discussions with ICPSR on the CED2AR project.

Roper Center

Organization Type: Academic Institution
Organization Location: Ithaca, NY

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Contribution to the development of metadata infrastructure/software.

US Census Bureau

Organization Type: Other Organizations (foreign or domestic)
Organization Location: Washington, DC

Partner's Contribution to the Project:
In-Kind Support
Facilities
Collaborative Research

More Detail on Partner and Contribution: Use of the Cornell Census Research Data implies a substantial Census Bureau participation since the Bureau pays substantially all of that RDC's operating expenses (unlike all the others, which bear these expenses themselves). The Census Bureau participated in the INFO7470 class, and we interact with the Census Bureau on the CED2AR project.

University of Michigan

Organization Type: Academic Institution
Organization Location: Ann Arbor, Michigan

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: Training course provided by Michigan NCRN node, supported by this grant's CED²AR for the purpose of training new users of the SIPP Synthetic Beta.

What other collaborators or contacts have been involved?
Nothing to report

Impacts

What is the impact on the development of the principal discipline(s) of the project?

CED2AR has contributed by posing the problem of confidentiality of metadata, and providing a solution. It also has highlighted the feasibility of crowdsourcing such information, while maintaining control over the quality of the resulting documentation at the data curator level. Work on Privacy and Confidentiality has contributed by highlighting the need to think about privacy in the context of both data providers (who desire privacy) and data users (who desire accuracy), and to provide a framework to make optimal choices. INFO7470 has contributed to making future and current researchers aware of the source of the data they are using, of the constraints in constructing such data, including confidentiality constraints, and novel methods of accessing the data.

What is the impact on other disciplines?
Nothing to report.

What is the impact on the development of human resources?

The availability of improved metadata, and of better privacyprotected publicuse data products, will enable more researchers to discover and use data, leading to new discoveries in the social sciences. INFO7470 trains new researchers in a variety of fields to use the resources of the statistical system effectively and appropriately.

What is the impact on physical resources that form infrastructure?
Nothing to report.

What is the impact on institutional resources that form infrastructure?

The availability of new metadata curation tools allows for institutions to adopt better, more transparent methods.

What is the impact on information resources that form infrastructure?
Nothing to report.

What is the impact on technology transfer?
Nothing to report.

What is the impact on society beyond science and technology?
Nothing to report.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

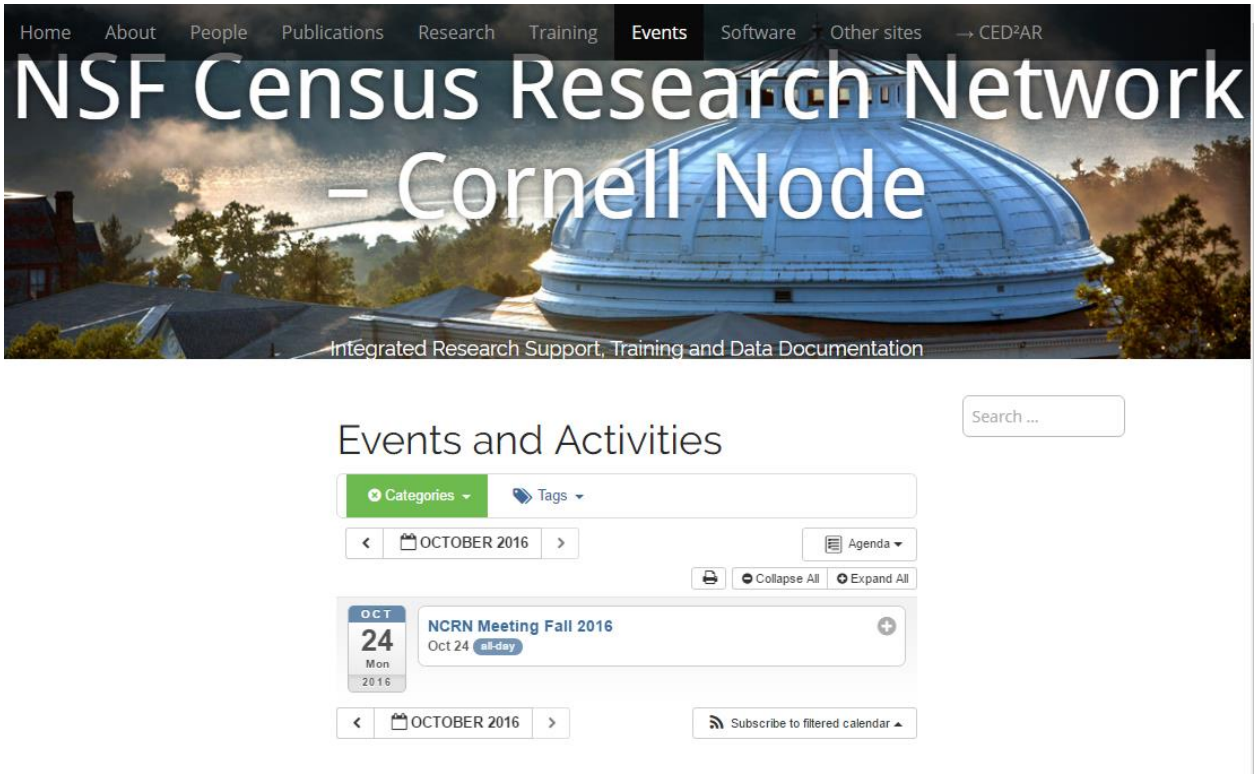
Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.



This document lists all presentations made by members of the NCRN Cornell node between October 1, 2015 and September 30, 2016.

All presentation and events are listed in a searchable form at https://www.ncrn.cornell.edu/events-and-activities/ .

OCT
6
Tue
2015

Vilhuber @ UNECE 2015: Using partially synthetic micro data to protect sensitive cells in business statistics @ UNECE Statistical Data Confidentiality Work Session

Oct 6 all-day

"*Using partially synthetic microdata to protect sensitive cells in business statistics,*" Lars Vilhuber **(NCRN, Cornell University)**, Javier Miranda (U.S. Census Bureau). This is an updated version of the presentation made at JSM 2015.

OCT
19
Mon
2015

Abowd @ Cornell Microeconomic Theory and Computer Science Workshop: "Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Good" @ Cornell University

Oct 19 @ 16:15

"Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Good", John Abowd **(Cornell University and U.S. Census Bureau),** Ian Schmutte (University of Georgia)

Abstract

We consider the problem of the public release of statistical information about a population–explicitly accounting for the public-good properties of both data accuracy and privacy loss. We first consider the implications of adding the public-good component to recently published models of private data publication under differential privacy guarantees using a Vickery-Clark-Groves mechanism and a Lindahl mechanism. We show that data quality will be inefficiently under-supplied. Next, we develop a standard social planner's problem using the technology set implied by (ε, δ)-differential privacy with (α, β)-accuracy for the Private Multiplicative Weights query release mechanism to study the properties of optimal provision of data accuracy and privacy loss when both are public goods. Using the production possibilities frontier implied by this technology, explicitly parameterized interdependent preferences, and the social welfare function, we display properties of the solution to the social planner's problem. Our results directly quantify the optimal choice of data accuracy and privacy loss as functions of the technology and preference parameters. Some of these properties can be quantified using population statistics on marginal preferences and correlations between income, data accuracy preferences, and privacy loss preferences that are available from survey data. Our results show that government data custodians should publish more accurate statistics with weaker privacy guarantees than would occur with purely private data publishing. Our statistical results using the General Social Survey and the Cornell National Social Survey indicate that the welfare losses from under-providing data accuracy while over-providing privacy protection can be substantial.

Abowd presents Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods @ 498 Uris Hall

Oct 19 @ 16:15 – 17:45

Joint Microeconomics & Computer Science Workshop: John M. Abowd

Abstract: We consider the problem of the public release of statistical information about a population–explicitly accounting for the public-good properties of both data accuracy and privacy loss. We first consider the implications of adding the public-good component to recently published models of private data publication under differential privacy guarantees using a Vickery-Clark-Groves mechanism and a Lindahl mechanism. We show that data quality will be inefficiently under-supplied. Next, we develop a standard social planner's problem using the technology set implied by (ε, δ)-differential privacy with (α, β)-accuracy for the Private Multiplicative Weights query release mechanism to study the properties of optimal provision of data accuracy and privacy loss when both are public goods. Using the production possibilities frontier implied by this technology, explicitly parameterized interdependent preferences, and the social welfare function, we display properties of the solution to the social planner's problem. Our results directly quantify the optimal choice of data accuracy and privacy loss as functions of the technology and preference parameters. Some of these properties can be quantified using population statistics on marginal preferences and correlations between income, data accuracy preferences, and privacy loss preferences that are available from survey data. Our results show that government data custodians should publish more accurate statistics with weaker privacy guarantees than would occur with purely private data publishing. Our statistical results using the General Social Survey and the Cornell National Social Survey indicate that the welfare losses from under-providing data accuracy while over-providing privacy protection can be substantial.

Paper: https://ecommons.cornell.edu/handle/1813/40581

OCT
23
Fri
2015

Vilhuber @ CAED 2015: "Usage and outcomes of the Synthetic Data Server" @ Comparative Analysis of Enterprise Data (CAED) 2015 Conference

Oct 23 @ 08:30 – Oct 25 @ 14:15

"Usage and outcomes of the Synthetic Data Server," Lars Vilhuber **(NCRN, Cornell University)** and John Abowd **(NCRN, Cornell University)**

The Synthetic Data Server (SDS) at Cornell University was set up to provide early access to new synthetic data products by the U.S. Census Bureau. These datasets are made available to interested researchers in a controlled environment, prior to a more generalized release. Over the past 5 years, 4 synthetic datasets were made available on the server, and over 100 users have accessed the server over that time period. This paper reports on interim outcomes of the activity: results of validation requests from a user perspective, functioning of the feedback loop due to validation and user input, and the role of the SDS as a access gateway to and educational tool for other mechanisms of accessing detailed person, household, establishment, and firm statistics.

OCT
30
Fri
2015

Abowd @ Temple University Economics Department Workshop: "Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Good" @ Temple University RA580

Oct 30 @ 14:30 – 16:00

"Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Good", John Abowd **(Cornell University and U.S. Census Bureau),** Ian Schmutte (University of Georgia)

Abstract

We consider the problem of the public release of statistical information about a population–explicitly accounting for the public-good properties of both data accuracy and privacy loss. We first consider the implications of adding the public-good component to recently published models of private data publication under differential privacy guarantees using a Vickery-Clark-Groves mechanism and a Lindahl mechanism. We show that data quality will be inefficiently under-supplied. Next, we develop a standard social planner's problem using the technology set implied by (ε, δ)-differential privacy with (α, β)-accuracy for the Private Multiplicative Weights query release mechanism to study the properties of optimal provision of data accuracy and privacy loss when both are public goods. Using the production possibilities frontier implied by this technology, explicitly parameterized interdependent preferences, and the social welfare function, we display properties of the solution to the social planner's problem. Our results directly quantify the optimal choice of data accuracy and privacy loss as functions of the technology and preference parameters. Some of these properties can be quantified using population statistics on marginal preferences and correlations between income, data accuracy preferences, and privacy loss preferences that are available from survey data. Our results show that government data custodians should publish more accurate statistics with weaker privacy guarantees than would occur with purely private data publishing. Our statistical results using the General Social Survey and the Cornell National Social Survey indicate that the welfare losses from under-providing data accuracy while over-providing privacy protection can be substantial.

NOV
11
Wed
2015

Vilhuber @ Université du Québec à Montréal (UQAM): Economics and the economics of privacy: new methods of accessing new data (in French) @ Université du Québec à Montréal

Nov 11 @ 15:00 – 16:30

NOV
13
Fri
2015

Abowd @ NBER Conference on Firm Heterogeneity and Income Inequality: "Earnings Inequality Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data" @ Stanford University

Nov 13 – Nov 14 all-day

"Earnings Inequality Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data", John Abowd **(Cornell University and U.S. Census Bureau),** Kevin McKinney **(U.S. Census Bureau),** Nellie Zhao **(Cornell University)**

Extended Abstract

We track sources of earnings inequality using the statistical technique introduced to the labor economics literature in 1999 (Abowd, Kramarz and Margolis, Econometrica 1999). When this technique has been used in Europe (Card, Heining and Kline QJE 2013 for Germany, in particular), the biggest contributor to the increase in earnings inequality appears to be increased employer-level heterogeneity (called the firm effect in AKM). Using the Census Bureau's Longitudinal Employer-Household Dynamics Infrastructure data for 1990-2013, we show that with respect to the U.S. data, the CHK result does not hold. There

has been very little change in employer-level earnings heterogeneity in the U.S. when one compares wage measures similar to the ones used to analyze the European data. European administrative databases allow one to construct something akin to a wage rate (usually, the amount that would be earned if an individual worked full-time full-year). The American data does not directly allow that. We develop a statistical approximation to the full-year full-time wage rate, using integrated Current Population Survey, Census 2000, and American Community Survey data. Using that measure, the earnings inequality trends in the U.S. look more similar to the European analyses.

But, for the purposes of studying earnings inequality, considering only the wage rate, and not the amount of time a person actually works, is seriously incomplete—especially in the U.S. where there is very little statutory employment security except in the public sector. The most important determinant of increased earnings inequality in our analyses is changes in labor force attachment (weeks worked in the year, hours worked per week).

In attempting to estimate how important the labor-force attachment component is, we reconstruct the work-eligible population (18-70) for each year from 1990-2013. The administrative records database developed at the Census Bureau uses an encrypted SSN to track individuals. The researcher can tell if the number that was encrypted is a valid SSN, and can also access the demographic details and employment history associated with the underlying SSN. In our model, there are two kinds of SSNs that are suspect: ones that are not valid (this means that the employer reported earnings in a state's UI system for an SSN that was never issued) and ones associated with demographic characteristics that mean it is unlikely that the owner of the SSN used it (leading case: the SSN was issued to a person who was less than 10 years old in the year during which the SSN was used to report UI eligible earnings). Our working hypotheses are: (1) the use of an invalid SSN reflects the work of a single undocumented immigrant, so we add that person to both the eligible population and the working population and (2) the use of a valid SSN issued to someone who appears to be too young (or too old) to work legally represents one person in the population (not working, not immigrant; i.e., eligible to get an SSN by virtue of birth in the U.S.) and at least one other person both working and in the work-eligible population, who is an undocumented immigrant.

Getting the non-working work-eligible population as accurate as possible is important because, especially during the Great Recession, many persons had no income from work for a full calendar year. We have no trouble finding these people for properly documented native-born and immigrant subpopulations, but we have to estimate how many work-eligible non-documented immigrants are still in the U.S. looking for work in any given year.

We also link data from the 1992-2012 Economic Censuses. These data are used to construct a measure of surplus per worker (revenue minus factor opportunity costs) for every private establishment in the censuses. These data show similar results for the population of working persons employed in the private sector. In particular, they show that there has not been an increase in overall earnings variability for this population.

NOV
30
Mon
2015

Abowd @ MIT/Census Big Data Meeting: Invited Speaker @ MIT/Census Big Data Meeting

Nov 30 all-day

Second Workshop: Location Confidentiality and Official Surveys

DEC
1

FCSM 2015: Total Variability Measures for Selected Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in OnTheMap @ FCSM 2015 Research Conference

Tue
2015

Dec 1 @ 13:15 – 15:00

"Total Variability Measures for Selected Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in OnTheMap", Kevin McKinney **(U.S. Census Bureau)**, Lars Vilhuber **(Cornell University and U.S. Census Bureau),** John Abowd **(Cornell University and U.S. Census Bureau),** Andrew Green **(Cornell University)**

Abstract

We report results from the first comprehensive total quality evaluation of three major indicators in the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program Quarterly Workforce Indicators (QWI): beginning-of-quarter employment, full-quarter employment, and average monthly earnings of full-quarter employees. Beginning-of-quarter employment is also the main tabulation variable in the LEHD Origin-Destination Employment Statistics workplace reports as displayed in OnTheMap (OTM). The evaluation is conducted using the multiple threads generated by the edit and imputation models used in the LEHD Infrastructure File System. These threads conform to the Rubin (1987) multiple imputation model. Each implicate is the output of formal probability models that address coverage, edit and imputation errors. Design-based sampling variability and finite population corrections are also included in the evaluation. We derive special formulas for the Rubin total variability and its components that are consistent with the disclosure avoidance system used for QWI and LODES/OTM workplace reports. These formulas allow us to publish the complete set of detailed total quality measures for QWI and LODES. The analysis reveals that the three publication variables under study are estimated very accurately for tabulations involving at least 10 jobs. Tabulations involving three to nine jobs have acceptable quality. Tabulations involving one or two jobs, which are generally suppressed in the QWI, have substantial total variability but their publication in LODES allows the formation of larger custom aggregations, which will in general have the accuracy estimated for tabulations in the QWI of similar magnitude.

FCSM 2015: Two Perspectives on Commuting and Workplace: A Microdata Comparison of Home to Work Flows Across Linked Survey and Administrative Files @ Federal Committee on Statistical Methodology (FCSM) 2015 Research Conference

Dec 1 @ 15:15 – 17:00

"Two Perspectives on Commuting and Workplace: A Microdata Comparison of Home to Work Flows Across Linked Survey and Administrative Files," Andrew Green **(U.S. Census Bureau, Cornell University)**, Mark Kutzbach **(U.S. Census Bureau)**, Lars Vilhuber **(U.S. Census Bureau, Cornell University)**

DEC
2
Wed
2015

Block and Edwards present on “What comes first? Metadata or Data Access?” at EDDI 2015

Dec 2 all-day

Michelle Edwards and William Block, Presentation at EDDI 2015

Brown presents @ EDDI 2015: Improving Access and Data Security to Confidential Labor Market Data @ Royal School of Library and Information Sciences

Dec 2 – Dec 3 all-day

“Improving Access and Data Security to Confidential Labor Market Data”, Warren Brown (Cornell University), Stephanie Jacobs (Cornell University), David Schiller (German Institute for Employment Research), Jörg Heining (German Institute for Employment Research)

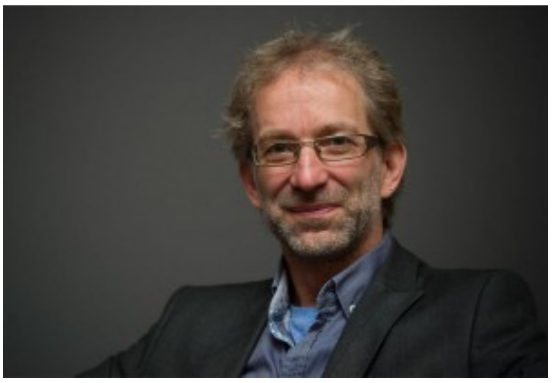
Abstract: The Cornell Institute for Social and Economic Research (CISER), Cornell University and the Institute for Employment Research (IAB), German Federal Employment Agency are collaborating to expand use of IAB’s confidential Sample of Integrated Labour Market Biographies (SIAB). DDI 2.5 is used to enable researchers to discover the files by means of variable level searching in a repository of metadata on U.S. and German labor market related data files. The repository is the Comprehensive Extensible Data Documentation and Access Repository (CED2AR) being developed by researchers at Cornell University with funding from the U.S. National Science Foundation. CED2AR provides researchers access to machine-readable codebooks with variable characteristics thus enabling researchers to develop detailed proposals for access to these data that are submitted to IAB. Researchers with approved projects are able to access and analyze the data using the Cornell Restricted Access Data Center (CRADC), a remote access virtual data enclave using remote desktop protocol. In the initial testing phase several researchers located in Europe and North America are successfully accessing and analyzing the Scientific Use Files of the SIAB. The project is well on its way to realizing the goal of wider access to researchers while improving secure management of confidential data.

The presentation can be found at http://hdl.handle.net/1813/44707

Vilhuber presents at FCSM 2015: Crowdsourcing Codebook Enhancements: A DDI-based Approach @ Federal Committee on Statistical Methodology (FCSM) 2015 Research Conference

Dec 2 @ 08:30 – 10:15

“Crowdsourcing Codebook Enhancements: A DDI-based Approach”
Benjamin Perry **(Cornell University)**, Venkata Kambhampaty **(Cornell University)**, Kyle Brumsted **(McGill University)**, Lars Vilhuber **(Cornell University),** William Block **(Cornell University)**



DEC
3
Thu
2015

FCSM 2015: "Formal Privacy Protection for Data Products Combining Individual and Employer Frames" @ Federal Committee on Statistical Methodology (FCSM) 2015 Research Conference
Dec 3 @ 10:30 – 12:15

"Formal Privacy Protection for Data Products Combining Individual and Employer Frames", Ashwin Machanavajjhala **(Duke University)**, Samuel Haney **(Duke University)**, Matthew Graham **(U.S. Census Bureau)**, Mark Kutzbach **(U.S. Census Bureau)**, Lars Vilhuber **(Cornell University and U.S. Census Bureau)**, John Abowd **(Cornell University and U.S. Census Bureau)**

DEC
18
Fri
2015

Abowd at Shanghai University of Finance and Economics: Invited Seminar @ School of Economics, Shanghai University of Finance and Economics
Dec 18 all-day

Invited Seminar, John Abowd **(Cornell University and U.S. Census Bureau)**

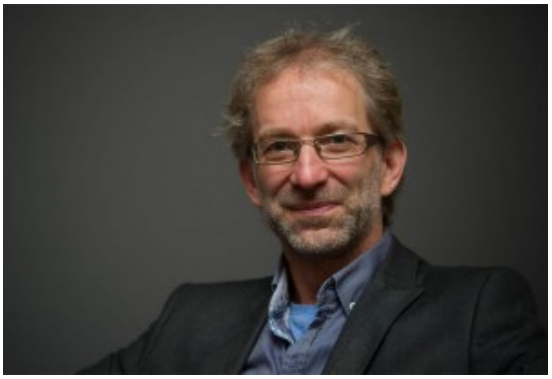
DEC
19
Sat
2015

Abowd presents at 2015 WISE International Symposium on Labor Economics: Keynote speaker @ 2015 WISE International Symposium on Labor Economics
Dec 19 – Dec 20 all-day

FEB
19
Fri
2016

Lars Vilhuber presents at Statistics Canada Socio-Economic Workshop @ Statistics Canada

Feb 19 all-day



Lars Vilhuber presents at Statistics Canada Socio-Economic Workshop “Using Business Microdata for Economic Research” on the topic of “Synthetic Establishment Microdata”.

APR
6
Wed
2016

Lars Vilhuber presents at “CASD Conference: Vos données au coeur de la datascience” @ Muséum national d'histoire naturelle

Apr 6 all-day

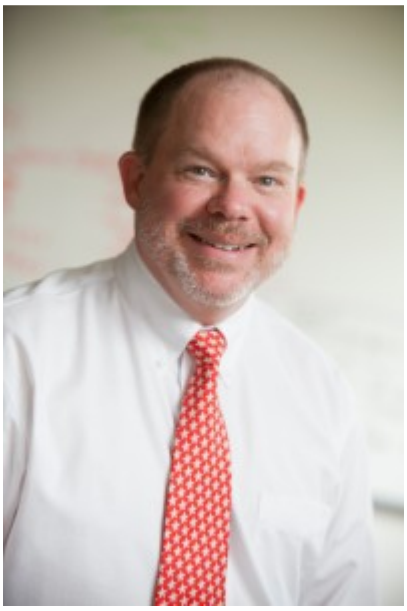
Title of the presentation: “Quelques développements en cours aux USA et au Canada”

APR
7
Thu
2016

William Block presents on “Incorporating W3C’s DQV and PROV in CISER’s Data Quality Review and Reproduction of Results Service” at NADDI

Apr 7 all-day

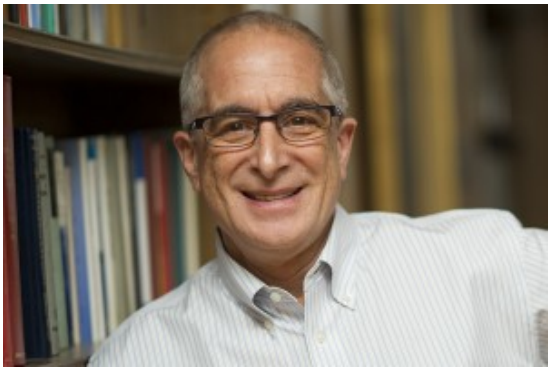
Poster presentation at North American DDI (NADDI) Conference, held in Edmonton, Alberta, CA on April 7, 2016. Download the poster from http://hdl.handle.net/1813/44704.



APR
22
Fri
2016

John Abowd: Social Science Research in the Era of Restricted-Access Data @ University of Nebraska-Lincoln

Apr 22 all-day



John Abowd will be giving two talks at the University of Nebraska-Lincoln, at the opening of the Central Plains Federal Statistical Research Data Center. The first talk is titled “Social Science Research in the Era of Restricted-Access Data”

APR
29
Fri
2016

Vilhuber gives Presentation at Annual Workshop of the Canadian Data Liberation Initiative @ McLennan Library, McGill University

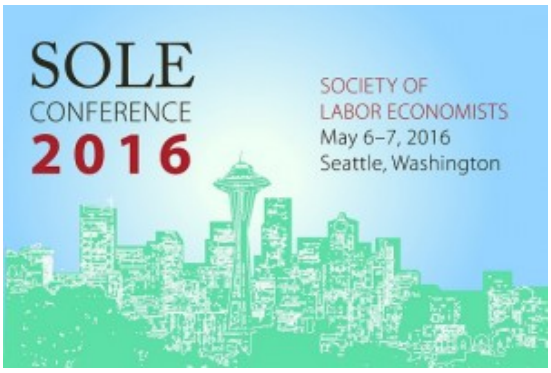
Apr 29 @ 11:00 – 12:00

Lars Vilhuber will present on “Crowdsourcing et métadonnées : défis et perspectives” (in French). Presentation is available at http://hdl.handle.net/1813/43875

MAY
6
Fri
2016

Abowd organizes Session at Society of Labor Economists Annual Meetings: A6: Issues in Data Privacy @ The Westin Seattle

May 6 @ 08:00 – 09:30



Session organized by John M. Abowd at the SOLE Conference 2016

Preliminary program:

A6: Issues in Data Privacy

- **Kobi Nissim**: “TBD”
- John M. Abowd and **Ian M. Schmutte:** “The Advantages and Disadvantages of Statistical Disclosure Limitation for Program Evaluation“
- Lars Vilhuber and John M. Abowd: “Usage and Outcomes of the Synthetic Data Server“

MAY
9
Mon
2016

Vilhuber gives Opening remarks at NCRN Spring 2016 Meeting @ U.S. Census Bureau

May 9 @ 10:00 – 10:15

Opening Remarks on behalf of the NCRN Coordinating Office at the NCRN Spring 2016 meetings – Lars Vilhuber

MAY
10
Tue
2016

Schmutte presents on The Advantages and Disadvantages of Statistical Disclosure Limitation for Program Evaluation @ U.S. Census Bureau

May 10 @ 10:15 – 10:45

John M. Abowd and **Ian M. Schmutte** : "*The Advantages And Disadvantages Of Statistical Disclosure Limitation For Program Evaluation*"

Abstract: This paper formalizes the manner in which statistical disclosure limitation (SDL) hinders empirical research in economics. We also highlight a hitherto unappreciated advantage of SDL, formal privacy models, and synthetic data systems: they can serve as a defense against model overfitting and falsediscovery bias. More specifically, a synthetic data validation system can – and we argue should – be used in conjunction with systems in which researchers register their research design ahead of analysis. The key insight is that privacyprotected data can be used for model development while minimizing risk of model overfitting. To demonstrate these points, we develop a model in which the statistical agency collects data from a population, but publishes a version in which the data that have been intentionally distorted by some SDL process. We say the SDL process is ignorable if inferences based on the published data are indistinguishable from inferences based on the unprotected data. SDL is rarely ignorable. If the researcher has knowledge of the SDL model, she can conduct an SDLaware analysis that explicitly corrects for the effects of SDL. If, as is often the case, if the SDL model is unknown, we describe circumstances under which SDL can still be learned.

Vilhuber presents on Crowdsourcing Codebook Development and Enhancements in CED²AR @ U.S. Census Bureau

May 10 @ 11:30 – 12:00

Benjamin Perry, Venkata Kambhampaty, Kyle Brumsted, **Lars Vilhuber, & William C. Block**: "*Crowdsourcing Codebook Development and Enhancements in CED²AR"*

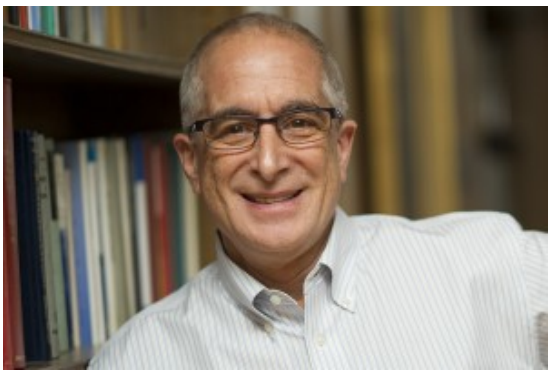
Abstract: Recent years have shown the power of usersourced information evidenced by the success of Wikipedia and its many emulators. This sort of unstructured discussion is currently not feasible as a part of the otherwise successful metadata repositories. Creating and augmenting metadata is a laborintensive endeavor. Harnessing collective knowledge from actual data users can supplement officially generated metadata. As part of our Comprehensive Extensible Data Documentation and Access Repository (CED²AR) infrastructure, we demonstrate a prototype of crowdsourced DDI on actual codebooks. While the system itself is more general, the demonstrated implementation relies on a set of linked deployments of the basic software on web servers. The backend transparently handles changes, and frontend has the ability to separate official edits (by designated curators of the data and the metadata) from crowdsourced content. The implementation allows a data curator, such as a statistical agency, to collect and incorporate improvements suggested by knowledgeable users in a structured way.

Available: https://ecommons.cornell.edu/handle/1813/43887

MAY
13
Fri
2016

John Abowd presents on "The Fate of Empirical Economics When All Data are Private"

May 13 all-day

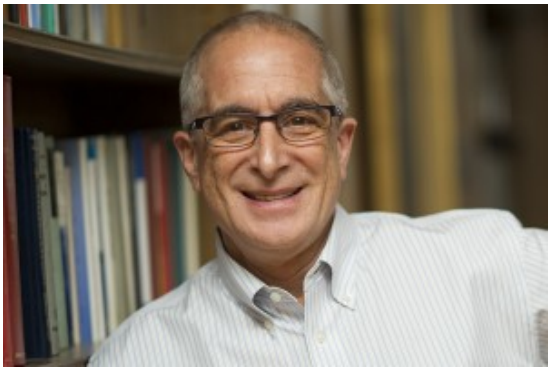


John M. Abowd (Cornell University and U.S. Census Bureau) presents at the Society of Government Economists on "The Fate of Empirical Economics When All Data are Private".

MAY
16
Mon
2016

Abowd presents on Four Challenges for Statistical Agencies @ Bureau of Labor Statistics

May 16 all-day



John M. Abowd (Cornell University and U.S Census Bureau) presents at the BLS Commissioner's Invited Seminar on "Four Challenges for Statistical Agencies".

The presentation can be found on the NCRN Presentation archive at http://hdl.handle.net/1813/44639.

MAY
18
Wed
2016

Vilhuber on "Big data, a supplementary tool for informed public decision making" @ Centre de Congrès de Québec

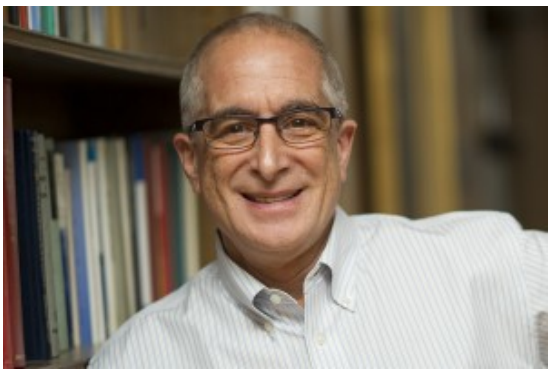
May 18 @ 08:30 – 11:00

Lars Vilhuber will present at the Association of Québecois Economists (Association des économistes québecois) in a session on "Les données massives (big data), un outil supplémentaire pour une décision publique éclairée" (Big data, a supplementary tool for informed public decision making)

MAY
20
Fri
2016

John Abowd on "An Integrated Approach to Statistical Agency Modernization"

May 20 all-day



John Abowd (Cornell University and U.S Census Bureau) presents "An Integrated Approach to Statistical Agency Modernization" at the Missouri-hosted workshop on "Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics".

JUN
2
Thu
2016

Florio Arguillas presents on "Towards Fully Replicable Data Analysis in an Increasingly Connected World"

Jun 2 all-day

For the presentation, see http://hdl.handle.net/1813/44703


CORNELL UNIVERSITY
FOUNDED A.D. 1865

JUN
3
Fri
2016

Warren Brown presents on “Online Tools and Training for Access and Analysis of Restricted Government Data Files”

Jun 3 all-day

Develop tools and training modules for online access enabling researchers to work more effectively with official restricted access statistical files. For the presentation, see http://hdl.handle.net/1813/44705

JUL
6
Wed
2016

Lagoze presents on “Reproducible research at Census” at NIST

Jul 6 all-day

The presentation can be found at http://hdl.handle.net/1813/44708

JUL
31
Sun
2016

JSM Session: Employer List Linking: Methods, Implementation, and Usage of Probabilistic Matches for Enhancing Workforce Statistics @ McCormick Conference Center

Jul 31 @ 14:00 – 15:30

Lars Vilhuber chairs session at JSM which includes multiple papers with NCRN contribution (presenter **bolded**, NCRN participants in *red italics*):

2:05 PM Robustness of Employer List Linking to Methodological Variation — **Mark J. Kutzbach, U.S. Census Bureau** ; Graton Gathright, U.S. Census Bureau ; Andrew Green, U.S. Census Bureau/Cornell University ; Kristin McCue, U.S. Census Bureau ; Holly Monti, U.S. Census Bureau ; *Ann Rodgers, University of Michigan ; Lars Vilhuber, Cornell University ; Nada Wasi, University of Michigan* ; Christopher Wignall, Amazon.com

2:25 PM	Two Perspectives on Commuting and Workplace: A Microdata Comparison of Home-to-Work Flows Across Linked Survey and Administrative Files— **Andrew Green, Cornell University/U.S. Census Bureau** ; Mark J. Kutzbach, U.S. Census Bureau ; *Lars Vilhuber, Cornell University*
2:45 PM	Developing Job Linkages for the Health and Retirement Study — **Kristin McCue, U.S. Census Bureau** ; *John M. Abowd*, U.S. Census Bureau/Cornell University ; *Margaret Levenstein, University of Michigan ; Matthew Shapiro, University of Michigan ; Ann Rodgers, University of Michigan ; Nada Wasi, University of Michigan* ; Dhiren Patki, University of Michigan

AUG
1
Mon
2016

JSM Session: Advances in Statistical Methods for Dissemination and Analysis of Official Statistics @ McCormick Conference Center

Aug 1 @ 14:00 – 15:50

John M. Abowd (former NCRN-Cornell PI) and other NCRN PIs present:

2:05 PM	An Integrated Approach to Providing Access to Confidential Social Science Data — **Jerome Reiter, Duke University**
2:30 PM	The Challenge of Reproducible Science and Privacy Protection for Statistical Agencies — **John M. Abowd, U.S. Census Bureau/Cornell University**
2:55 PM	Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates — **Scott H. Holan, University of Missouri** ; Jonathan R. Bradley, University of Missouri ; Christopher Wikle, University of Missouri

JSM Session: Statistical Foundations of Data Privacy @ McCormick Conference Center

Aug 1 @ 14:00 – 15:50

Lars Vilhuber chairs a session organized by the **Committee on Privacy and Confidentiality** and *Aleksandra Slavkovic, Penn State University:*

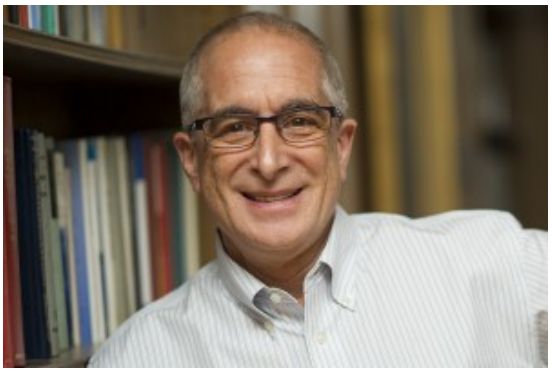
2:05 PM	Connections Between Privacy Definitions and Arbitrage-Free Pricing Functions — **Daniel Kifer, Penn State University**
2:25 PM	Differentially Private Statistical Inference and Hypothesis Testing — **Vishesh Karwa, Carnegie Mellon University**
2:45 PM	Learning with Differential Privacy: Stability, Learnability, and the Sufficiency and Necessity of ERM Principle — **Yu-Xiang Wang, Carnegie Mellon University** ; Jing Lei, Carnegie Mellon University ; Stephen E. Fienberg, Carnegie Mellon University

3:05 PM	Performance Bounds for Graphical Record Linkage: Can record linkage bounds provide guidance for private synthetic data release? — **Rebecca Steorts, Duke University** ; Matt Barnes, Carnegie Mellon University ; Willie Neisweigner, Carnegie Mellon University
3:25 PM	**Discussant:** Adam Smith, Penn State University

SEP
6
Tue
2016

John M. Abowd, WSS JULIUS SHISKIN MEMORIAL AWARD SEMINAR, "How Will Statistical Agencies Operate When All Data are Private?" @ U.S. Census Bureau Auditorium

Sep 6 @ 13:00 – 15:00



John M. Abowd, NCRN Cornell and now Associate Director for Research and Methodology and Chief Scientist at the U.S. Census Bureau is the2016 Recipient of Julius Shiskin Memorial Award for Economic Statistics. He will speak on **September 6, 2016** at the WSS JULIUS SHISKIN MEMORIAL AWARD SEMINAR on "*How Will Statistical Agencies Operate When All Data are Private?"*

Time: 1 – 3 p.m.

Location: Auditorium, U.S. Census Bureau, 4600 Silver Hill Road, Suitland, Maryland, available through Webex.

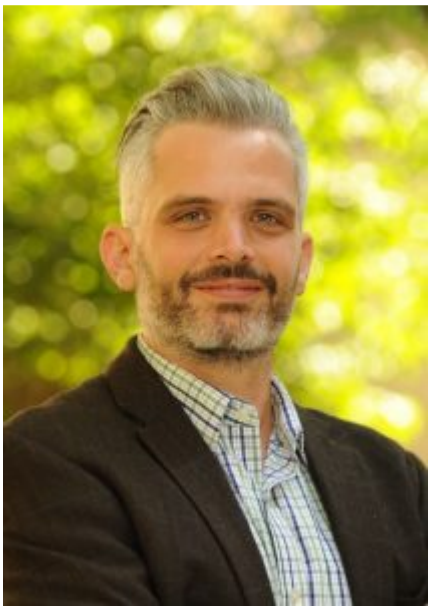
Abstract: The dual problems of respecting citizen privacy and protecting the confidentiality of their data—Ken Prewitt's famous "don't ask/don't tell" dictum—have become hopelessly conflated in the "Big Data" era. There are orders of magnitude more data outside an agency's firewall than inside it—compromising the integrity of traditional statistical disclosure limitation methods. And increasingly the information processed by the agency was "asked" in a context wholly outside the agency's operations—blurring the distinction between what was asked and what is published. Already private businesses like Microsoft, Google and Apple recognize that cybersecurity (safeguarding the integrity and access controls for internal data) and privacy protection (ensuring that what is published does not reveal too much about any person or business) are two sides of the same coin. This is a paradigm-shifting moment for statistical agencies. This talk will examine how statistical agencies can respond in manner consistent with their missions.

SEP
17
Sat
2016

Schmutte presents "Estimating Compensating Wage Differentials with Endogenous Job Mobility"

Sep 17 all-day

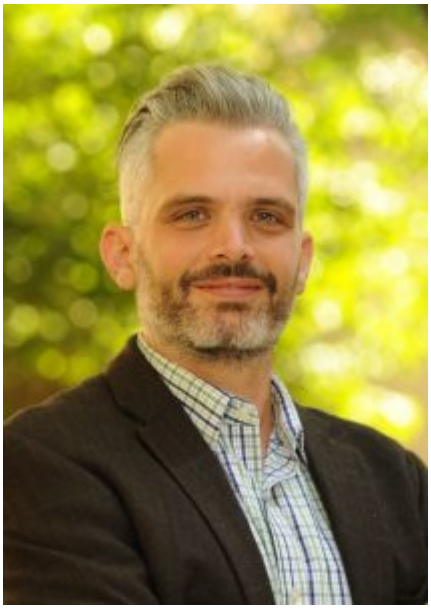
Joint work with Kurt Lavetti. Presentation can be found at http://hdl.handle.net/1813/44709



OCT
28
Fri
2016

Schmutte presents “Revisiting the Economics of Privacy: Population Statistics and Privacy as Public Goods” @ Issac Newton Institute, Cambridge University

Oct 28 all-day



Workshop on “Privacy: Recent Developments at the Interface Between Economics and Computer Science” at the Issac Newton Institute, Cambridge University. Joint work with John M. Abowd.