

[My Desktop](#)
[Prepare & Submit Proposals](#)
[Prepare Proposals in FastLane](#)
[New! Prepare Proposals \(Limited proposal types\)](#)
[Proposal Status](#)
[Awards & Reporting](#)
[Notifications & Requests](#)
[Project Reports](#)
[Submit Images/Videos](#)
[Award Functions](#)
[Manage Financials](#)
[Program Income Reporting](#)
[Grantee Cash Management Section Contacts](#)
[Administration](#)
[Lookup NSF ID](#)

Preview of Award 1131848 - Final 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 - 09/30/2018
Reporting Period:	10/01/2017 - 09/30/2018
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).

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](http://www.vrdc.cornell.edu/info7470/)" (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: A final version of CED2AR under this grant was released (2.10). Multiple components from the backend were cleaned up, documented, and released under open source licenses (see <https://github.com/ncrnucornell>). A new codebook for the SIPP Synthetic Beta was edited, and released. A collaboration with ICPSR on incorporating CED2AR technology into the data curation workflow was advanced, but did not lead to the desired integration. Multiple articles were completed and submitted, with several published.

Specific Objectives: A new codebook for the SIPP Synthetic Beta was released on the platform, and is the primary reference for that dataset. The article on "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices" is forthcoming in the top economics journal, and will greatly contribute to the discussion of privacy and accuracy.

Significant Results: The article "Effects of a Government-Academic Partnership: ..." by the NCRN PIs was published, showcasing the effect of each node's contributions, as well as the network's contribution as a whole, on furthering research in collaboration with the federal statistical system.

Key outcomes or Other achievements:

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

A graduate student (Herbert) is working on research with confidential data.

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

Multiple papers were published in academic journals, including the top economics journal and a widely read statistics journal. 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. All papers are made available on properly curated document archives at <http://ecommons.cornell.edu> as well as <https://zenodo.org/communities/labordynamicsinstitute/>.

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers

Andrew S. Green and Mark J. Kutzbach and Lars Vilhuber (2017). {Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files}. *Center for Economic Studies, U.S. Census Bureau*,

Working Papers. (17-34), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

Chen, Beidi and Shrivastava, Anshumali and Steorts, Rebecca C. (2018). Unique entity estimation with application to the Syrian conflict. *Ann. Appl. Stat.*. 12 (2), 1039--1067. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1214/18-AOAS1163

Daniel H. Weinberg and John M. Abowd and Robert F. Belli and Noel Cressie and David C. Folch and Scott H. Holan and Margaret C. Levenstein and Kristen M. Olson and Jerome P. Reiter and Matthew D. Shapiro and Jolene Smyth (2018). {Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Improve the U.S. Statistical System?}. *Journal of Survey Statistics and Methodology*. smy023. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1093/jssam/smy023

John M. Abowd (2017). How Will Statistical Agencies Operate When All Data Are Private?. *Journal of Privacy and Confidentiality*. 7 (3), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.29012/jpc.v7i3.404

John M. Abowd and Francis Kramarz and Sebastien Perez-Duarte and Ian M. Schmutte (2018). Sorting Between and Within Industries: A Testable Model of Assortative Matching. *Annals of Economics and Statistics*. 1-32. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.15609/annaeconstat2009.129.0001

John M. Abowd and Ian M. Schmutte (2017). {Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods}. *Center for Economic Studies, U.S. Census Bureau, Working Papers*. (17-37), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

John M. Abowd and Ian M. Schmutte (2018). An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *arXiv*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

John M. Abowd and Ian M. Schmutte (2019). An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*. . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

John M. Abowd and Ian M. Schmutte and Lars Vilhuber (2018). {Disclosure Limitation and Confidentiality Protection in Linked Data}. *Center for Economic Studies, U.S. Census Bureau, Working Papers*. (18-07), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

John M. Abowd and Kevin L. McKinney and Nellie Zhao (2018). Earnings Inequality and Mobility Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data. *Journal of Labor Economics*. 36 (S1), 183-300. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1086/694104

Kevin L. McKinney and Andrew S. Green and Lars Vilhuber and John M. Abowd (2019). {Total Error and Variability Measures with Integrated Disclosure Limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in On The Map}. *Journal of Survey Statistics and Methodology*. (17-71), . Status = UNDER_REVIEW; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Lars Vilhuber and Carl Lagoze (2017). Making Confidential Data Part of Reproducible Research. *Labor Dynamics Institute Document*. (41), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

Lars Vilhuber and Ian Schmutte (2017). Proceedings from the 2017 Cornell-Census-NSF-Sloan Workshop on Practical Privacy. *Labor Dynamics Institute Document*. (43), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

Lars Vilhuber and Saki Kinney and Ian Schmutte (2017). Proceedings from the Synthetic LBD International Seminar. *Labor Dynamics Institute Document*. (44), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

Samuel Haney and Ashwin Machanavajjhala and John M. Abowd and Matthew Graham and Mark Kutzbach (2017). Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics. *Proceedings of the 2017 ACM International Conference on Management of Data*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/3035918.3035940

Vilhuber, Lars and Lagoze, Carl (2017). Making Confidential Data Part of Reproducible Research. *Chance*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Vilhuber, Lars and Schmutte, Ian (2017). Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy. *Labor Dynamics Institute Document*. (1813:46197), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No

Licenses

Other Conference Presentations / Papers

Other Products

Other Publications

Lars Vilhuber (2015). *Codebook for NBER-CES Manufacturing Industry Database*. Codebook for NBER-CES Manufacturing Industry Database (2009) [NAICS and SIC], by Randy A. Becker , Wayne B. Gray , Jordan Marvakov , and Eric J. Bartelsman Main website: <https://www.nber.org/data/nberces5809.html> (note: a newer version is available at <http://www.nber.org/data/nberces.html> - this codebook does not necessarily reflect the more recent version.) Live version of the DDI codebook at <https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/nber-ces/> <http://doi.org/10.5281/zenodo.2527908>. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Lars Vilhuber (2015). *Codebook for the National QWI [Codebook file]*. Codebook for the early research version of National QWI. <http://doi.org/10.5281/zenodo.2527906> Live version of the DDI codebook at <https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/nqwi/>. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Reeder, Lori B., Stanley, Jordan C., & Lars Vilhuber. (2018). *Codebook for the SIPP Synthetic Beta 7.0 (DDI-C and PDF)*. <http://doi.org/10.5281/zenodo.1477097>. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Reeder, Lori B., Stanley, Jordan C., & Vilhuber, Lars. (2018). *Codebook for the SIPP Synthetic Beta 7.0 (PDF version)*. <http://doi.org/10.5281/zenodo.1477099>. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Lars Vilhuber (2016). *DDI Codebook for the Synthetic LBD*. Codebook for the Synthetic LBD, a Census Bureau data product, see <https://www.census.gov/ces/dataproducts/synlbd/>. The SynLBD usage model relies on a Synthetic Data Server, maintained (as of 2018) by Cornell University, see <https://www2.vrdc.cornell.edu/news/synthetic-data-server/>. Live version of the DDI codebook at <https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/synlbd/> <http://doi.org/10.5281/zenodo.2527910>. Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Patents

Technologies or Techniques

We create and publish CED2AR, software to edit, display, and disseminate DDI-C codebooks. The source code is published on Github at <https://github.com/ncrn-cornell/ced2ar>, and releases are published and archived on Zenodo with DOI [10.5281/zenodo.597000](https://doi.org/10.5281/zenodo.597000). A sample production server can be found at <http://www2.ncrn.cornell.edu/ced2ar-web/>.

Thesis/Dissertations

Websites

Github repositories for NCRN-Cornell
<https://github.com/ncrn-cornell>

We published all code developed, as well as various other artifacts, on Github. Select releases are also archived on Zenodo.

Maven artifacts by NCRN project
<https://mvnrepository.com/artifact/edu.cornell.ncrn.ced2ar>

Maven is an online repository for Java packages. CED2AR is software developed under our project for editing, displaying, and disseminating DDI-C codebooks. We have published for re-use by the community relevant components of our work.

NCRN Cornell Node website
<https://www.ncrn.cornell.edu/>

The website displays information about various outputs from the project.

eCommons collections for NCRN Cornell Node
<https://hdl.handle.net/1813/30503>

Documents have been preserved on the Cornell eCommons. Select presentations can also be found at
<https://hdl.handle.net/1813/43872>

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Vilhuber, Lars	PD/PI	2
Block, William	Co PD/PI	1
Lagoze, Carl	Faculty	1
Barker, Brandon	Other Professional	3
Simmer, Charles	Other Professional	12
Brown, Warren	Staff Scientist (doctoral level)	1
Herbert, Sylvérie	Graduate Student (research assistant)	6
Sexton, William	Graduate Student (research assistant)	0
Stanchi, Flavio	Graduate Student (research assistant)	0
Schmutte, Ian	Consultant	1

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:** 2**Contribution to the Project:** Lead PI, work on confidentiality, metadata, CED2AR, overall management.**Funding Support:** This grant, Sloan 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: Work on metadata

Funding Support: This grant

International Collaboration: No

International Travel: No

Carl Lagoze

Email: clagoze@umich.edu

Most Senior Project Role: Faculty

Nearest Person Month Worked: 1

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

Charles C Simmer

Email: chuck.simmer@gmail.com

Most Senior Project Role: Other Professional

Nearest Person Month Worked: 12

Contribution to the Project: Software development

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: 1

Contribution to the Project: Expertise on ACS, INFO7470.

Funding Support: NSF (this grant)

International Collaboration: No

International Travel: No

Sylvérie Herbert**Email:** sh2258@cornell.edu**Most Senior Project Role:** Graduate Student (research assistant)**Nearest Person Month Worked:** 6**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:** 0**Contribution to the Project:** Assistance on confidentiality research**Funding Support:** 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:** 0**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

Ian Schmutte**Email:** schmutte@uga.edu**Most Senior Project Role:** Consultant**Nearest Person Month Worked:** 1**Contribution to the Project:** Co-authored various papers**Funding Support:** This grant**International Collaboration:** No**International Travel:** No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location
Bureau of Labor Statistics	Other Nonprofits	Washington, DC

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:

Bureau of Labor Statistics

Organization Type: Other Nonprofits

Organization Location: Washington, DC

Partner's Contribution to the Project:

Collaborative Research

More Detail on Partner and Contribution: We have been working with BLS staff to allow them to use the CED2AR software and create codebooks on confidential data available through the FSRDC system.

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 crowd sourcing 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 privacy protected public use data products, will enable more researchers to discover and use data, leading to new discoveries in the social sciences. INFO7470 trained 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. The discussion on privacy and accuracy has impacted the thinking in the Federal Statistical System about these new methods, including the implementation of differential privacy by the U.S. Census Bureau, and the consideration of formal methods by other parts of the FSS.

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?

The discussion about privacy and accuracy has broadly broken out of academia, and is being discussed and reported on at the New York Times, NPR, and other media. Triggered by political decisions in part, the academic discussion of the necessity to weigh privacy and accuracy of statistics that are used by government and society has contributed to that discussion, and articles funded by this grant have been cited in the mainstream media.

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.