My Desktop
Prepare & Submit Proposals
Proposal Status
Proposal Functions
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 - 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:	John M Abowd, Principal Investigator William C Block, Co-Principal Investigator Lars Vilhuber, Co-Principal Investigator
Recipient Organization:	Cornell University
Project/Grant Period:	10/01/2011 - 09/30/2016
Reporting Period:	10/01/2014 - 09/30/2015
Submitting Official (if other than PD\PI):	Lars Vilhuber Co-Principal Investigator
Submission Date:	10/21/2015
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Lars Vilhuber

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" (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: CED2AR is the official repository for documentation on two of the Census Bureau's datasets (see publications). We also have an early functioning prototype available within the U.S. Census Bureau. Ongoing work with CISER, the Roper Center, and the Census Bureau to implement CED2AR functionality are progressing.

Specific Objectives:

Significant Results: Abowd and Schmutte (2015) show that 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. 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.

Key outcomes or Other achievements:

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

We have hired a new post-doc under this grant (Long Zhang), who will assist with the confidentiality research that we have ramped up. A post-doc mentoring plan is attached. 6 undergraduate students were engaged either over the course of the academic year or in our summer lab, from a variety of fields (CS, economists, engineering). A graduate student in statistics graduated, and took up a new position at Rice University. We continue to have another graduate student work on our projects.

*** 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.

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

The last year will see consolidation of the software development on CED²AR, making it robust to subsequent financing or handing it over to the community. Several key elements are still being developed, in particular a very promising approach to crowd-sourcing metadata. Work on the perception of confidentiality, informing the work by Abowd and Schmutte, is progressing in collaboration with researchers at the Census Bureau, using data collected on weekly surveys.

Supporting Files

Filename	Description	Uploaded By	Uploaded On
Postdoc_Mentoring_Plan.pdf	Post-doc mentoring plan.	Lars Vilhuber	08/28/2015

Products

Books

Book Chapters

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 = No; Peer Reviewed = Yes

John M. Abowd and Ian Schmutte (2015). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*. Fall 20 . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; ISSN: 00072303

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

Anshumali Shrivastava and Ping Li (2014). *A New Space for Comparing Graphs*. ASONAM. . Status = PUBLISHED; Acknowledgement of Federal Support = Yes

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 = OTHER; 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

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). *CED²AR: 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 = OTHER; 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 = OTHER; 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 = OTHER; 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 = OTHER; Acknowledgement of Federal Support = Yes

Other Products

Other Publications

Patents

Technologies or Techniques

(Minor) enhancement to Wordpress software (used for academic websites), to parse RSS feeds by dSpace implementations (such as ecommons.cornell.edu) and generate Bibtex bibliographic reference files. Source code available at https://github.com/ncrncornell/wp-plugins, sample implementation at https://www.ncrn.cornell.edu/publications/

https://github.com/ncrncornell/wp-plugins

R wrapper around the original source code of `mtcd`, a standalone C++ implementation of the statistical model proposed in "Synthesizing Truncated Count Data for Confidentiality" and originally created by the Duke NCRN node. Website is https://github.com/ncrncornell/Rmtcd

https://github.com/ncrncornell/Rmtcd

Source code for CED²AR released at https://github.com/ncrncornell/ced2ar, live implementation at https://www2.ncrn.cornell.edu/ced2ar-web/search, demo site at https://demo.ncrn.cornell.edu/ced2ar-web/search

https://github.com/ncrncornell/ced2ar

Thesis/Dissertations

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

Websites

Github site for NCRN Cornell
https://github.com/ncrncornell/

Github site for software source code released by NCRN node

Github site for R wrapper
https://github.com/ncrncornell/Rmtcd

The repository contains a R wrapper around the original source code of `mtcd`, a standalone C++ implementation of the statistica model proposed in "Synthesizing Truncated Count Data for Confidentiality" and originally created by the Duke NCRN node.

`mtcd`

is a standalone C++ implementation of the statistical model proposed in "Synthesizing Truncated Count Data for Confidentiality"

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

All the information about the node is catalogues 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
Abowd, John	PD/PI	1
Block, William	Co PD/PI	1
Vilhuber, Lars	Co PD/PI	3
Lagoze, Carl	Faculty	1
Kambhampaty, Venkata	Other Professional	12
Perry, Benjamin	Other Professional	11
Williams, Jeremy	Other Professional	0
Brown, Warren	Staff Scientist (doctoral level)	2
Edwards, Anne	Staff Scientist (doctoral level)	0
Shrivastava, Anshumali	Graduate Student (research assistant)	3
Stanchi, Flavio	Graduate Student (research assistant)	6

Full details of individuals who have worked on the project:

John M Abowd
Email: john.abowd@cornell.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Leader of work on Confidentiality and Privacy as well as High-Dimensional Computational Statistics

Funding Support: Funding also through Census IPA, Sloan Foundation.

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, supervision of software developers, conference presentations.

Funding Support: This grant and Cornell University

International Collaboration: No
International Travel: Yes, United Kingdom - 0 years, 0 months, 3 days

Lars Vilhuber
Email: lars.vilhuber@cornell.edu

Most Senior Project Role: Co 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

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

Venkata Kambhampaty
Email: vkambhampaty@cornell.edu
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

Benjamin Perry
Email: bap63@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 11

Contribution to the Project: Programming

Funding Support: This grant

International Collaboration: No
International Travel: Yes, United Kingdom - 0 years, 0 months, 3 days

---

Jeremy Williams
Email: jw568@cornell.edu
Most Senior Project Role: Other Professional
Nearest Person Month Worked: 0

Contribution to the Project: Programming and writing up the results of project research.

Funding Support: This grant and Cornell University

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

Anshumali Shrivastava
Email: ansh@cs.cornell.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 3

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: 6

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
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.

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 t 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.

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 discover and use data, leading to new discoveries in the social sciences.

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 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.