Preview of Award 1131848 - Annual Project Report

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 Ping Li, Co-Principal Investigator Lars Vilhuber, Co-Principal Investigator
Submitting Official (if other than PD\PI):	N/A
Submission Date:	N/A
Recipient Organization:	Cornell University
Project/Grant Period:	10/01/2011 - 09/30/2016
Reporting Period:	10/01/2012 - 09/30/2013
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).

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: The INFO7470 class was given in Spring of 2013.

	Development of the CED2AR tool is well underway. Work on specific codebooks is nearly complete, and one of the datasets (Synthetic LBD) was showcased in a session at ISI 2013. Development of algorithms to implement High Performance Logistic Regression Methods for Data Edits and Imputation is well underway.
Specific Objectives:	The INFO7470 class is aimed at American Ph.D. students 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. No equivalent class exists, and take-up is wide, and generally found to be useful. The CED2AR tool is the platform on which to achieve the overarching goal of liberating metadata on confidential data. Once stable, it will be used to host the data within the confidential confines (the Census Bureau) as well as the filtered metadata outside (where it will be deployed first) the HP Logistic Regression methods are to be applied to confidential Census data in order to address the problem of multiple-response variables with incomplete coded links or unknown responses.
Significant Results:	The INFO7470 class reached 120 signed-up students in 11 universities and campuses, including the Census Bureau. A development version of CED2AR was shown to partners, and a production site released after the reporting period. Significant effort was put into cleaning up metadata (or creating metadata) for poorly or undocumented datasets. HP Logistic Regression and related algorithms were developed, and applications shown outside of the Census context.
Key outcomes or Other achievements:	We've trained a signficant number of students and staff at the Census Bureau. We have created previously inexistant metadata, and will be making it available in an easily accessible form.

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

120 students have been trained as part of the INFO7470 class.

The project worked with Cornell's CS 5150 ("Software engineering") and students picked a component of CED2AR as a class project. One of our staff members participated as part of a degree program, and one of the students was subsequently hired as a staff programmer.

A graduate student in computer science is engaged with CED2AR, working with information scientists and economists, making this a cross-disciplinary project. A economist graduate student was brought on in the fall of 2013.

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

Our outreach has been through conferences, working papers and publications, and presentations to statistical agencies, as well as continued engagement of the Census Bureau in particular through the PIs of the project. A session on one of the synthetic datasets that will be highlighted in CED2AR was prepared and presented at ISI 2013.

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

INFO7470 will be developed into a MOOC-style online class, and made available to the scientific community in 2015

again.

CED2AR will be released. Additional metadata will be generated, and experience gained from the workflow will be translated into a tool for easier metadata creation by end-users instead of data librarians, and integrated into a Census-RDC workflow that will faciliate the replicability of research results within secure data centers.

Products

Journals

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.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = No ; DOI: 10.2218/ijdc.v8i1.259

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)*. 565.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/2187980.2188129

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)*. 613.

Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1145/2187980.2188153

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.

Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1007/978-3-642-33627-0_17.

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.

Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://books.nips.cc/papers/files/nips25/NIPS2012_1436.pdf.

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.

Status = PUBLISHED; Acknowledgement of Federal Support = Yes ; Peer Reviewed = Yes ; OTHER: http://books.nips.cc/papers/files/nips25/NIPS2012_1456.pdf.

Thesis/Dissertations

Conference Papers and Presentations

Ping Li and Anshumali Shrivastava and K\"onig, Arnd Christian (2013). *b-Bit Minwise Hashing in Practice*. Internetware'13. Changsha, China.

Status = AWAITING_PUBLICATION; Acknowledgement of Federal Support = Yes

Ping Li and Cun-Hui Zhang (2012). *Exact Sparse Recovery with L0 Projections*. 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Chicago, USA.

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). Bristol, UK.

Status = PUBLISHED; Acknowledgement of Federal Support = Yes

Other Publications

Technologies or Techniques
Nothing to report.

Patents
Nothing to report.

Inventions
Nothing to report.

Licenses
Nothing to report.

Websites

Title: NCRN-Cornell node website

URL: http://www.ncrn.cornell.edu

Description: Provides information on the activities of the project.

Title: INFO7470

URL: http://www.vrdc.cornell.edu/info7470/

Description: The course is designed to teach students basic and advanced techniques for acquiring and transforming raw information into social and economic data. The 2013 version is particularly aimed at American Ph.D. students 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. Major emphasis is placed on U.S. Census Bureau data that are accessible from the Bureau's Research Data Center network. Graduate students and faculty who are planning to use RDC-based data, or are seriously considering it, should pay particular attention to the labs related to the proposal process. The RDC-accessible data products covered in the course include the internal files used to manage the Census Bureau's household and establishment frames; the Longitudinal Employer-Household Dynamics (LEHD) micro data; the Longitudinal Business Database (LBD) and its predecessor the Longitudinal Research Database (LRD); internal versions of the Survey of Income and Program Participation (SIPP), Current Population Survey (CPS), American Community Survey (ACS), American Housing Survey (AHS), and the 1990, 2000, and 2010 Decennial Censuses of Population and Housing; the Employer and Non-employer Business Registers (BR and SSEL); the Censuses and Annual Surveys of Manufactures, Mining, Services, Retail Trade, Wholesale Trade, Construction, Transportation, Communications, and Utilities;

Business Expenditures Survey; Characteristics of Business Owners; and others. Students will also be introduced to the NSF-sponsored Virtual Research Data Center and Social Science Gateway to XSEDE.

Other Products

Product Type: Audio or Video Products

Description: Online recordings of the INFO7470 class, freely available at

http://www.vrdc.cornell.edu/info7470/video.php

for webstreaming.

Other: Educational aids or Curricula

Product Type: The curriculum for the INFO7470 class on "Understanding Social and Economic Data" is generally viewed as useful, and is available online

Description:

http://www.vrdc.cornell.edu/info7470/course_outline.php

Other:

Participants

Research Experience for Undergraduates (REU) funding

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Warren Brown	Staff Scientist (doctoral level)	2
Carl Lagoze	Faculty	1
Flavio Stanchi	Graduate Student (research assistant)	1
William C Block	Co PD/PI	1
Anshumali Shrivastava	Graduate Student (research assistant)	3
Jeremy Williams	Other Professional	2
Benjamin Perry	Other Professional	4
Margo Anderson	Faculty	0
Aleksandra B. Slavkovic	Faculty	0
Nicholas Nagle	Faculty	0
John M Abowd	PD/PI	2
Ping Li	Co PD/PI	2

Lars Vilhuber	Co PD/PI	2

What other organizations have been involved as partners?

Name	Location
ICPSR	Ann Arbor, MI
US Census Bureau	Washington, DC

Have other collaborators or contacts been involved? Y

Impacts

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

We expect the impact to occur in 2013-2014. The groundwork has been laid.

What is the impact on other disciplines?

The project is cross-disciplinary (information science, computer science, statistics, economics) and influences all four domains.

What is the impact on the development of human resources?

The metadata-oriented outcomes of this project will facilitate knowledge discovery and creation by graduate students and other scientists 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?

(practices) The improvements to metadata standards are being proposed to international standards bodies, and thus likely to be adopted by many institutions worldwide.

What is the impact on information resources that form infrastructure?

The metadata-oriented outcomes of this project will facilitate knowledge discovery and creation of data, thus improving the overall set of information resources. Such resources do not need to, and won't, be housed within the project.

What is the impact on technology transfer?

(public use) The improvements to metadata standards are being proposed to international standards bodies, and thus likely to be adopted by many institutions worldwide.

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

Changes

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.