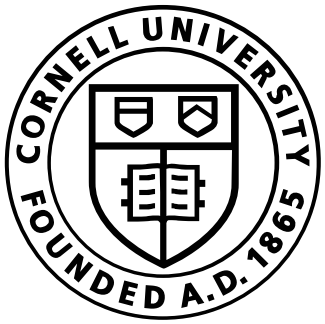
Cornell Node of the NSF-Census Research Network - Reports



Cornell University

Lars VILHUBER (Principal Investigator)

William BLOCK (co-Principal Investigator)

Description and List of Outcomes of the Cornell node of the NSF-Census Research Network

Also see our **main website** for many additional details.

Outcomes report

Our project pursued and achieved several complementary goals:

- We aimed to facilitate the **documentation and discoverability of restricted-use data**, by developing a tool, usable by researchers, to create and disseminate such documentation. We created Comprehensive Extensible Data Documentation and Access Repository (CED2AR), a web-based tool using the leading metadata standards (Data Documentation Initiative, DDI). The tool has been used to create and curate multiple releases of data documentation (see https://www2.ncrn.cornell.edu). Extensions for collaborative and crowd-sourced editing of such documentation were explored, with the former being implemented within the current data product, and the latter having been explored with various institutions.

- We addressed the issue of **linkage in high-dimensional data**, with multiple innovative papers creating novel algorithms to address these problems. These algorithms are finding applications in real-world problems that have privacy implications, f.i.

    - Chen, Shrivastava, and Steorts. "Unique entity estimation with application to the Syrian conflict". *Ann. Appl. Stat.* 12.2.

- We tackled the theoretical and practical problem of efficient trade-offs between data **quality and confidentiality** (accuracy v. privacy loss) using techniques from economics; i.e., a formal production possibilities frontier (PPF) and a formal social welfare function (SWF). We consider situations where data quality will be inefficiently under-supplied, and how statistical agencies can manage the accuracy privacy-loss tradeoff using the SWF. The research led to several influential papers,
    - Abowd, John M., and Schmutte, Ian, "Economic analysis and statistical disclosure limitation". *Brookings Papers on Economic Activity* Fall 2015 (2015).
    - Abowd, John M., and Schmutte, Ian, "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices". *American Economic Review* 109 (1 2019). doi: 10.1257/aer.20170627.
    - Abowd, John M., Ian Schmutte, William Sexton, and Lars Vilhuber, "Why the Economics Profession Must Actively Participate in the Privacy Protection Debate", *American Economic Review: Papers and Proceedings*, 2019 (see also LDI Document 51)
    - Abowd, John M., "How Will Statistical Agencies Operate When All Data Are Private?". *Journal of Privacy and Confidentiality* 7 (3). https://doi.org/10.29012/jpc.v7i3.404.

- More importantly, the research and the publications were instrumental in demonstrating to the U.S. Census Bureau that **better data protection mechanisms were needed**, contributing directly to the adoption of formal privacy measures by the U.S. Census Bureau:
    - Census Advisory Committee Meeting 2017-09-15
    - Abowd, John M., "The U.S. Census Bureau Adopts Differential Privacy", *KDD '18 Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, UK (handle.net/1813/60392 and doi.org/10.1145/3219819.3226070)

- Finally, we expanded and taught (in 2011, 2013, 2016, 2017) a multi-site distance learning class on **"Understanding Social and Economic Data" (INFO 7470)**. The course teaches students (Ph.D. students and faculty in the social sciences) basic and advanced techniques for acquiring and transforming raw information into social and economic data. The course focuses on the outputs and usage of confidential data from the U.S. Census

Bureau and other American statistical agencies. Over 500 students participated in the various iterations of the class. A textbook is being prepared, but many of the materials are available under open licenses, see f.i. https://hdl.handle.net/1813/43949.

Additional **collaborations with the U.S. Census Bureau** that emerged over the life of the project led to various other innovations. McKinney et al. (2017) produced the first total error analysis of the Quarterly Workforce Indicators that incorporates variability stemming from editing, multiple imputation, and disclosure avoidance measures. Green et al. (2017) investigated the coherence of ACS and administrative reports of workplace location. We continued supporting mechanisms to access confidential data through the publication of synthetic data combined with access and validation servers, see Synthetic Data Server @ Cornell.

Beyond individual contributions of this node, we were part of a **network of nodes**. Jointly, the nodes responded to the desire of the statistical system to have better and deeper integration with academia. Many of the nodes' members were introduced to academic-government partnerships for the first time; others deepened their collaboration. The network, and all its nodes, were honored with the American Statistical Association's "Statistical Partnerships Among Academe, Industry, and Government" (SPAIG) Award. A full report on that network's activities can be found in

- Daniel H Weinberg, John M Abowd, Robert F Belli, Noel Cressie, David C Folch, Scott H Holan, Margaret C Levenstein, Kristen M Olson, Jerome P Reiter, Matthew D Shapiro, Jolene D Smyth, Leen-Kiat Soh, Bruce D Spencer, Seth E Spielman, Lars Vilhuber, Christopher K Wikle (2018); "Effects of a government-academic partnership: Has the NSF-Census Bureau Research Network helped improve the US statistical system?, *Journal of Survey Statistics and Methodology*, smy023, doi.org/10.1093/jssam/smy023 (also handle.net/1813/52650.2).

A **detailed list of our project's outcomes**, including a bibliography, can be viewed at https://ncrncornell.github.io/reports/, and most reports produced by the project have been archived at https://hdl.handle.net/1813/30503, including a comprehensive bibliography with additional information on published articles.

In addition to the Principal Investigators listed, the project also included John ABOWD, Ping LI (previous Principal Investigators), Warren BROWN, Carl LAGOZE (Senior Researchers). Funding for the Cornell node of the NSF-Census Research Network was provided by NSF Grant #1131848 (NCRN), with additional funding by NSF Grant #1012593 (TC-

Large) and a grant by the Alfred P. Sloan Foundation, and would not have been possible without the support of the U.S. Census Bureau