1 Bibliography December 29, 2018

References

John M. Abowd: How Will Statistical Agencies Operate When All Data Are Private?

John M. Abowd. *How Will Statistical Agencies Operate When All Data Are Private?* Document 30. Labor Dynamics Institute, Cornell University, 2016.

Abstract: The dual problems of respecting citizen privacy and protecting the confidentiality of their data 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.

John M. Abowd: How Will Statistical Agencies Operate When All Data Are Private?

John M. Abowd. "How Will Statistical Agencies Operate When All Data Are Private?" In: *Journal of Privacy and Confidentiality* 7.3 (2017). DOI: 10.29012/jpc.v7i3.404.

Abstract: The dual problems of respecting citizen privacy and protecting the confidentiality of their data 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.

John M. Abowd: Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do

John M. Abowd. *Why Statistical Agencies Need to Take Privacy-loss Budgets Seriously, and What It Means When They Do*. Document 32. Labor Dynamics Institute, Cornell University, 2016.

Abstract: To appear on fcsm.sites.usa.gov, as presented to the 2016 FCSM Statistical Policy Seminar.

John M. Abowd et al.: Sorting Between and Within Industries: A Testable Model of Assortative Matching

John M. Abowd, Francis Kramarz, Sebastien Perez-Duarte, and Ian M. Schmutte. *Sorting Between and Within Industries: A Testable Model of Assortative Matching*. Document 40. Labor Dynamics Institute, 2017. URL: http://digitalcommons.ilr.cornell.edu/ldi/28/.

Abstract: We test Shimer's (2005) theory of the sorting of workers between and within industrial sectors based on directed search with coordination frictions, deliberately maintaining its static general equilibrium framework. We fit the model to sector-specific wage, vacancy and output data, including publicly-available statistics that characterize the distribution of worker and employer wage heterogeneity across sectors. Our empirical method is general and can be applied to a broad class of assignment models. The results indicate that industries are the loci of sorting–more productive workers are employed in more productive industries. The evidence confirms that strong assortative matching can be present even when worker and employer components of wage heterogeneity are weakly correlated.

John M. Abowd et al.: Sorting Between and Within Industries: A Testable Model of Assortative Matching

John M. Abowd, Francis Kramarz, Sebastien Perez-Duarte, and Ian M. Schmutte. "Sorting Between and Within Industries: A Testable Model of Assortative Matching". In: *Annals of Economics and Statistics* (129 2018), pp. 1–32. DOI: 10.15609/annaeconstat2009.129.0001.

Abstract: We test Shimer's (2005) theory of the sorting of workers between and within industrial sectors based on directed search with coordination frictions, deliberately maintaining its static general equilibrium framework. We fit the model to sector-specific wage, vacancy and output data, including publicly-available statistics that characterize the distribution of worker and employer wage heterogeneity across sectors. Our empirical method is general and can be applied to a broad class of assignment models. The results indicate that industries are the loci of sorting-more productive workers are employed in more productive industries. The evidence confirms that strong assortative matching can be present even when worker and employer components of wage heterogeneity are weakly correlated.

John M. Abowd et al.: Noise infusion as a confidentiality protection measure for graph-based statistics

John M. Abowd and Kevin L. McKinney. "Noise infusion as a confidentiality protection measure for graph-based statistics". In: *Statistical Journal of the IAOS* 32.1 (2016), pp. 127–135. DOI: 10.3233/SJI-160958.

Abstract: We use the bipartite graph representation of longitudinally linked employer-employee data, and the associated projections onto the employer and employee nodes, respectively, to characterize the set of potential statistical summaries that the trusted custodian might produce. We consider noise infusion as the primary confidentiality protection method. We show that a relatively straightforward extension of the dynamic noise-infusion method used in the U.S. Census Bureau's Quarterly Workforce Indicators can be adapted to provide the same confidentiality guarantees for the graph-based statistics: all inputs have been modified by a minimum percentage deviation (i.e., no actual respondent data are used) and, as the number of entities contributing to a particular statistic increases, the accuracy of that statistic approaches the unprotected value. Our method also ensures that the protected statistics will be identical in all releases based on the same inputs.

File: :https\://ecommons.cornell.edu/bitstream/handle/1813/42338/AbowdMcKinney-with%20galley%20corrections.pdf?sequence=2&isAllowed=y:URL;:AbowdMcKinney-SJIAOS2016.pdf:PDF.

John M. Abowd et al.: Modeling Endogenous Mobility in Wage Determination

John M. Abowd, Kevin L. McKinney, and Ian M. Schmutte. *Modeling Endogenous Mobility in Wage Determination*. Document 28. Labor Dynamics Institute, 2016. URL: http://digitalcommons.ilr.cornell.edu/ldi/28/.

Abstract: We evaluate the bias from endogenous job mobility in fixed-effects estimates of worker- and firm-specific earnings heterogeneity using longitudinally linked employer-employee data from the LEHD infrastructure file system of the U.S. Census Bureau. First, we propose two new residual diagnostic tests of the assumption that mobility is exogenous to unmodeled determinants of earnings. Both tests reject exogenous mobility. We relax the exogenous mobility assumptions by modeling the evolution of the matched data as an evolving bipartite graph using a Bayesian latent class framework. Our results suggest that endogenous mobility biases estimated firm effects toward zero. To assess validity, we match our estimates of the wage components to out-of-sample estimates of revenue per worker. The corrected estimates attribute much more of the variation in revenue per worker to variation in match quality and worker quality than the uncorrected estimates.

John M. Abowd et al.: Earnings Inequality and Mobility Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data

John M. Abowd, Kevin L. Mckinney, and Nellie Zhao. "Earnings Inequality and Mobility Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data". In: *Journal of Labor Economics* 36.S1 (2018), pp. 183–300. DOI: 10.1086/694104.

Abstract: Using earnings data from the U.S. Census Bureau, this paper analyzes the role of the employer in explaining the rise in earnings inequality in the United States. We first

establish a consistent frame of analysis appropriate for administrative data used to study earnings inequality. We show that the trends in earnings inequality in the administrative data from the Longitudinal Employer-Household Dynamics Program are inconsistent with other data sources when we do not correct for the presence of misused SSNs. After this correction to the worker frame, we analyze how the earnings distribution has changed in the last decade. We present a decomposition of the year-to-year changes in the earnings distribution from 2004-2013. Even when simplifying these flows to movements between the bottom 20%, the middle 60% and the top 20% of the earnings distribution, about 20.5 million workers undergo a transition each year. Another 19.9 million move between employment and nonemployment. To understand the role of the firm in these transitions, we estimate a model for log earnings with additive fixed worker and firm effects using all jobs held by eligible workers from 2004-2013. We construct a composite log earnings firm component across all jobs for a worker in a given year and a non-firm component. We also construct a skill-type index. We show that, while the difference between working at a low- or middle-paying firm are relatively small, the gains from working at a top-paying firm are large. Specifically, the benefits of working for a high-paying firm are not only realized today, through higher earnings paid to the worker, but also persist through an increase in the probability of upward mobility. High-paying firms facilitate moving workers to the top of the earnings distribution and keeping them there.

John M. Abowd et al.: Economic analysis and statistical disclosure limitation

John M. Abowd and Ian Schmutte. "Economic analysis and statistical disclosure limitation". In: *Brookings Papers on Economic Activity* Fall 2015 (2015). URL: http://www.brookings.edu/about/projects/bpea/papers/2015/economic-analysis-statistical-disclosure-limitation.

Abstract: This paper explores the consequences for economic research of methods used by statistical agencies to protect confidentiality of their respondents. We first review the concepts of statistical disclosure limitation for an audience of economists who may be unfamiliar with these methods. Our main objective is to shed light on the effects of statistical disclosure limitation for empirical economic research. In general, the standard approach of ignoring statistical disclosure limitation leads to incorrect inference. We formalize statistical disclosure methods in a model of the data publication process. In the model, the statistical agency collects data from a population, but published a version of the data that have been intentionally distorted. The model allows us to characterize what it means for statistical disclosure limitation to be ignorable, and to characterize what happens when it is not. We then consider the effects of statistical disclosure limitation for regression analysis, instrumental variable analysis, and regression discontinuity design. Because statistical agencies do not always report the methods they use to protect confidentiality, we use our model to characterize settings in which statistical disclosure limitation methods are discoverable; that is, they can be learned from the released data. We conclude with advice for researchers, journal editors, and statistical agencies.

John M. Abowd et al.: Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods

John M. Abowd and Ian Schmutte. *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*. Document 22. Labor Dynamics Institute, 2015. URL: http://digitalcommons.ilr.cornell.edu/ldi/22/.

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 (?, ?)-differential privacy with (?, ?)-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.

John M. Abowd et al.: An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices

John M. Abowd and Ian M. Schmutte. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices". In: *American Economic Review* (forthcoming).

John M. Abowd et al.: An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices

John M. Abowd and Ian M. Schmutte. *An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices*. Working Papers 18-35. Center for Economic Studies, U.S. Census Bureau, 2018. URL: https://ideas.repec.org/p/cen/wpaper/18-35.html.

Abstract: Statistical agencies face a dual mandate to publish accurate statistics while protecting respondent privacy. Increasing privacy protection requires decreased accuracy. Recognizing this as a resource allocation problem, we propose an economic solution: operate where the marginal cost of increasing privacy equals the marginal benefit. Our model of production,

from computer science, assumes data are published using an efficient differentially private algorithm. Optimal choice weighs the demand for accurate statistics against the demand for privacy. Examples from U.S. statistical programs show how our framework can guide decision-making. Further progress requires a better understanding of willingness-to-pay for privacy and statistical accuracy.

John M. Abowd et al.: An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices

John M. Abowd and Ian M. Schmutte. *An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices*. preprint. arXiv, 2018. URL: https://arxiv.org/abs/1808.06303.

John M. Abowd et al.: Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods

John M. Abowd and Ian M. Schmutte. *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*. Document 37. Labor Dynamics Institute, 2017. URL: http://digitalcommons.ilr.cornell.edu/ldi/37/.

Abstract: We consider the problem of determining the optimal accuracy of public statistics when increased accuracy requires a loss of privacy. To formalize this allocation problem, we use tools from statistics and computer science to model the publication technology used by a public statistical agency. We derive the demand for accurate statistics from first principles to generate interdependent preferences that account for the public-good nature of both data accuracy and privacy loss. We first show data accuracy is inefficiently under-supplied by a private provider. Solving the appropriate social planner's problem produces an implementable publication strategy. We implement the socially optimal publication plan for statistics on income and health status using data from the American Community Survey, National Health Interview Survey, Federal Statistical System Public Opinion Survey and Cornell National Social Survey. Our analysis indicates that welfare losses from providing too much privacy protection and, therefore, too little accuracy can be substantial.

John M. Abowd et al.: Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods

John M. Abowd and Ian M. Schmutte. *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*. Working Papers 17-37. Center for Economic Studies, U.S. Census Bureau, 2017. URL: https://ideas.repec.org/p/cen/wpaper/17-37.html.

Abstract: We consider the problem of determining the optimal accuracy of public statistics when increased accuracy requires a loss of privacy. To formalize this allocation problem, we use tools from statistics and computer science to model the publication technology used by a public statistical agency. We derive the demand for accurate statistics from first principles

to generate interdependent preferences that account for the public-good nature of both data accuracy and privacy loss. We first show data accuracy is inefficiently undersupplied by a private provider. Solving the appropriate social planner's problem produces an implementable publication strategy. We implement the socially optimal publication plan for statistics on income and health status using data from the American Community Survey, National Health Interview Survey, Federal Statistical System Public Opinion Survey and Cornell National Social Survey. Our analysis indicates that welfare losses from providing too much privacy protection and, therefore, too little accuracy can be substantial.

John M. Abowd et al.: Disclosure Limitation and Confidentiality Protection in Linked Data

John M. Abowd, Ian M. Schmutte, and Lars Vilhuber. *Disclosure Limitation and Confidentiality Protection in Linked Data*. Working Papers 18-07. Center for Economic Studies, U.S. Census Bureau, 2018. URL: https://ideas.repec.org/p/cen/wpaper/18-07.html.

Abstract: Confidentiality protection for linked administrative data is a combination of access modalities and statistical disclosure limitation. We review traditional statistical disclosure limitation methods and newer methods based on synthetic data, input noise infusion and formal privacy. We discuss how these methods are integrated with access modalities by providing three detailed examples. The first example is the linkages in the Health and Retirement Study to Social Security Administration data. The second example is the linkage of the Survey of Income and Program Participation to administrative data from the Internal Revenue Service and the Social Security Administration. The third example is the Longitudinal Employer-Household Dynamics data, which links state unemployment insurance records for workers and firms to a wide variety of censuses and surveys at the U.S. Census Bureau. For examples, we discuss access modalities, disclosure limitation methods, the effectiveness of those methods, and the resulting analytical validity. The final sections discuss recent advances in access modalities for linked administrative data.

John M. Abowd et al.: A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs

John M. Abowd, Lars Vilhuber, and William Block. "A Proposed Solution to the Archiving and Curation of Confidential Scientific Inputs". In: *Privacy in Statistical Databases*. Ed. by Josep Domingo-Ferrer and Ilenia Tinnirello. Vol. 7556. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 216–225. ISBN: 978-3-642-33626-3. DOI: 10.1007/978-3-642-33627-0_17.

Abstract: We develop the core of a method for solving the data archive and curation problem that confronts the custodians of restricted-access research data and the scientific users of such data. Our solution recognizes the dual protections afforded by physical security and access limitation protocols. It is based on extensible tools and can be easily incorporated into existing instructional materials.

Chen et al.: Unique entity estimation with application to the Syrian conflict

Beidi Chen, Anshumali Shrivastava, and Rebecca C. Steorts. "Unique entity estimation with application to the Syrian conflict". In: *Ann. Appl. Stat.* 12.2 (2018), pp. 1039–1067. DOI: 10.1214/18-AOAS1163.

Abstract: Entity resolution identifies and removes duplicate entities in large, noisy databases and has grown in both usage and new developments as a result of increased data availability. Nevertheless, entity resolution has tradeoffs regarding assumptions of the data generation process, error rates, and computational scalability that make it a difficult task for real applications. In this paper, we focus on a related problem of unique entity estimation, which is the task of estimating the unique number of entities and associated standard errors in a data set with duplicate entities. Unique entity estimation shares many fundamental challenges of entity resolution, namely, that the computational cost of all-to-all entity comparisons is intractable for large databases. To circumvent this computational barrier, we propose an efficient (near-linear time) estimation algorithm based on locality sensitive hashing. Our estimator, under realistic assumptions, is unbiased and has provably low variance compared to existing random sampling based approaches. In addition, we empirically show its superiority over the state-of-the-art estimators on three real applications. The motivation for our work is to derive an accurate estimate of the documented, identifiable deaths in the ongoing Syrian conflict. Our methodology, when applied to the Syrian data set, provides an estimate of 191,874±1,772 documented, identifiable deaths, which is very close to the Human Rights Data Analysis Group (HRDAG) estimate of 191,369. Our work provides an example of challenges and efforts involved in solving a real, noisy challenging problem where modeling assumptions may not hold. This project was started when Shrivastava and Steorts were funded by NCRN grants to Cornell and CMU, respectively.

Drechsler et al.: Synthetic Longitudinal Business Databases for International Comparisons

Jörg Drechsler and Lars Vilhuber. "Synthetic Longitudinal Business Databases for International Comparisons". In: *Privacy in Statistical Databases*. Ed. by Josep Domingo-Ferrer. Vol. 8744. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 243–252. ISBN: 978-3-319-11256-5. DOI: 10.1007/978-3-319-11257-2_19.

Abstract: International comparison studies on economic activity are often hampered by the fact that access to business microdata is very limited on an international level. A recently launched project tries to overcome these limitations by improving access to Business Censuses from multiple countries based on synthetic data. Starting from the synthetic version of the longitudinally edited version of the U.S. Business Register (the Longitudinal Business Database, LBD), the idea is to create similar data products in other countries by applying the synthesis methodology developed for the LBD to generate synthetic replicates that could be distributed without confidentiality concerns. In this paper we present some first results

of this project based on German business data collected at the Institute for Employment Research.

Green et al.: Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files

Andrew S. Green, Mark J. Kutzbach, and Lars Vilhuber. *Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files*. Working Papers 17-34. Center for Economic Studies, U.S. Census Bureau, 2017. URL: https://ideas.repec.org/p/cen/wpaper/17-34.html.

Abstract: Commuting flows and workplace employment data have a wide constituency of users including urban and regional planners, social science and transportation researchers, and businesses. The U.S. Census Bureau releases two, national data products that give the magnitude and characteristics of home to work flows. The American Community Survey (ACS) tabulates households' responses on employment, workplace, and commuting behavior. The Longitudinal Employer-Household Dynamics (LEHD) program tabulates administrative records on jobs in the LEHD Origin-Destination Employment Statistics (LODES). Design differences across the datasets lead to divergence in a comparable statistic: county-to-county aggregate commute flows. To understand differences in the public use data, this study compares ACS and LEHD source files, using identifying information and probabilistic matching to join person and job records. In our assessment, we compare commuting statistics for job frames linked on person, employment status, employer, and workplace and we identify person and job characteristics as well as design features of the data frames that explain aggregate differences. We find a lower rate of within-county commuting and farther commutes in LODES. We attribute these greater distances to differences in workplace reporting and to uncertainty of establishment assignments in LEHD for workers at multi-unit employers. Minor contributing factors include differences in residence location and ACS workplace edits. The results of this analysis and the data infrastructure developed will support further work to understand and enhance commuting statistics in both datasets.

Haney et al.: Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics

Samuel Haney, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, and Mark Kutzbach. "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics". In: *Proceedings of the 2017 ACM International Conference on Management of Data* (2017). DOI: 10.1145/3035918.3035940.

Abstract: National statistical agencies around the world publish tabular summaries based on combined employer-employee (ER-EE) data. The privacy of both individuals and business establishments that feature in these data are protected by law in most countries. These data are currently released using a variety of statistical disclosure limitation (SDL) techniques that do not reveal the exact characteristics of particular employers and employees, but lack

provable privacy guarantees limiting inferential disclosures. In this work, we present novel algorithms for releasing tabular summaries of linked ER-EE data with formal, provable guarantees of privacy. We show that state-of-the-art differentially private algorithms add too much noise for the output to be useful. Instead, we identify the privacy requirements mandated by current interpretations of the relevant laws, and formalize them using the Pufferfish framework. We then develop new privacy definitions that are customized to ER-EE data and satisfy the statutory privacy requirements. We implement the experiments in this paper on production data gathered by the U.S. Census Bureau. An empirical evaluation of utility for these data shows that for reasonable values of the privacy-loss parameter $>= 1$, the additive error introduced by our provably private algorithms is comparable, and in some cases better, than the error introduced by existing SDL techniques that have no provable privacy guarantees. For some complex queries currently published, however, our algorithms do not have utility comparable to the existing traditional SDL algorithms. Those queries are fodder for future research.

Haney et al.: Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics

Samuel Haney, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics". In: *Proceedings of the 2017 International Conference on Management of Data*. SIGMOD '17. ACM, 2017. DOI: 10.1145/3035918.3035940.

Abstract: National statistical agencies around the world publish tabular summaries based on combined employer-employee (ER-EE) data. The privacy of both individuals and business establishments that feature in these data are protected by law in most countries. These data are currently released using a variety of statistical disclosure limitation (SDL) techniques that do not reveal the exact characteristics of particular employers and employees, but lack provable privacy guarantees limiting inferential disclosures. In this work, we present novel algorithms for releasing tabular summaries of linked ER-EE data with formal, provable guarantees of privacy. We show that state-of-the-art differentially private algorithms add too much noise for the output to be useful. Instead, we identify the privacy requirements mandated by current interpretations of the relevant laws, and formalize them using the Pufferfish framework. We then develop new privacy definitions that are customized to ER-EE data and satisfy the statutory privacy requirements. We implement the experiments in this paper on production data gathered by the U.S. Census Bureau. An empirical evaluation of utility for these data shows that for reasonable values of the privacy-loss parameter $\epsilon \geq 1$, the additive error introduced by our provably private algorithms is comparable, and in some cases better, than the error introduced by existing SDL techniques that have no provable privacy guarantees. For some complex queries currently published, however, our algorithms do not have utility comparable to the existing traditional SDL algorithms. Those queries are fodder for future research.

Haney et al.: Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics

Samuel Haney, Ashwin Machanavajjhala, John M Abowd, Matthew Graham, and Mark Kutzbach. *Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics*. Preprint 1813:49652. Cornell University, 2017. URL: http://hdl.handle.net/1813/49652.

Abstract: Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics Haney, Samuel; Machanavajjhala, Ashwin; Abowd, John M; Graham, Matthew; Kutzbach, Mark National statistical agencies around the world publish tabular summaries based on combined employeremployee (ER-EE) data. The privacy of both individuals and business establishments that feature in these data are protected by law in most countries. These data are currently released using a variety of statistical disclosure limitation (SDL) techniques that do not reveal the exact characteristics of particular employers and employees, but lack provable privacy guarantees limiting inferential disclosures. In this work, we present novel algorithms for releasing tabular summaries of linked ER-EE data with formal, provable guarantees of privacy. We show that state-of-the-art differentially private algorithms add too much noise for the output to be useful. Instead, we identify the privacy requirements mandated by current interpretations of the relevant laws, and formalize them using the Pufferfish framework. We then develop new privacy definitions that are customized to ER-EE data and satisfy the statutory privacy requirements. We implement the experiments in this paper on production data gathered by the U.S. Census Bureau. An empirical evaluation of utility for these data shows that for reasonable values of the privacy-loss parameter >=1, the additive error introduced by our provably private algorithms is comparable, and in some cases better, than the error introduced by existing SDL techniques that have no provable privacy guarantees. For some complex queries currently published, however, our algorithms do not have utility comparable to the existing traditional "This Article is brought to you for free and open access by the Centers, Institutes, Programs at DigitalCommons@ILR. It has been accepted for inclusion in Labor Dynamics Institute by an authorized administrator of DigitalCommons@ILR. For more information, please contact hlmdigital@cornell.edu."

Lagoze et al.: Data Management of Confidential Data

Carl Lagoze, William C. Block, Jeremy Williams, John M. Abowd, and Lars Vilhuber. "Data Management of Confidential Data". In: *International Journal of Digital Curation* 8.1 (2013). Presented at 8th International Digital Curation Conference 2013, Amsterdam. See also http://hdl.handle.net/1813/30924, pp. 265–278. DOI: 10.2218/ijdc.v8i1.259.

Abstract: Social science researchers increasingly make use of data that is confidential because it contains linkages to the identities of people, corporations, etc. The value of this data lies in the ability to join the identifiable entities with external data such as genome data, geospatial information, and the like. However, the confidentiality of this data is a barrier to its utility and curation, making it difficult to fulfill US federal data management mandates and interfering

with basic scholarly practices such as validation and reuse of existing results. We describe the complexity of the relationships among data that span a public and private divide. We then describe our work on the CED2AR prototype, a first step in providing researchers with a tool that spans this divide and makes it possible for them to search, access, and cite that data.

Lagoze et al.: Encoding Provenance of Social Science Data: Integrating PROV with DDI

Carl Lagoze, William C. Block, Jeremy Williams, and Lars Vilhuber. "Encoding Provenance of Social Science Data: Integrating PROV with DDI". In: *5th Annual European DDI User Conference*. 2013. DOI: http://dx.doi.org/10.3886/eDDILagoze.

Abstract: Provenance is a key component of evaluating the integrity and reusability of data for scholarship. While recording and providing access provenance has always been important, it is even more critical in the web environment in which data from distributed sources and of varying integrity can be combined and derived. The PROV model, developed under the auspices of the W3C, is a foundation for semantically-rich, interoperable, and web-compatible provenance metadata. We report on the results of our experimentation with integrating the PROV model into the DDI metadata for a complex, but characteristic, example social science data. We also present some preliminary thinking on how to visualize those graphs in the user interface.

File: :LagozeEtAl2013:PDF.

Lagoze et al.: CED²AR: The Comprehensive Extensible Data Documentation and Access Repository

Carl Lagoze, Lars Vilhuber, Jeremy Williams, Benjamin Perry, and William C. Block. "CED²AR: The Comprehensive Extensible Data Documentation and Access Repository". In: *ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014)*. Presented at the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014). ACM/IEEE. London, United Kingdom: Institute of Electrical & Electronics Engineers (IEEE), 2014. DOI: 10.1109/JCDL.2014.6970178.

Abstract: Social science researchers increasingly make use of data that is confidential because it contains linkages to the identities of people, corporations, etc. The value of this data lies in the ability to join the identifiable entities with external data such as genome data, geospatial information, and the like. However, the confidentiality of this data is a barrier to its utility and curation, making it difficult to fulfill US federal data management mandates and interfering with basic scholarly practices such as validation and reuse of existing results. We describe the complexity of the relationships among data that span a public and private divide. We then describe our work on the CED2AR prototype, a first step in providing researchers with a tool that spans this divide and makes it possible for them to search, access, and cite that data.

Lagoze et al.: Encoding Provenance Metadata for Social Science Datasets

Carl Lagoze, Jeremy Willliams, and Lars Vilhuber. "Encoding Provenance Metadata for Social Science Datasets". In: *Metadata and Semantics Research*. Ed. by Emmanouel Garoufallou and Jane Greenberg. Vol. 390. Communications in Computer and Information Science. Springer International Publishing, 2013, pp. 123–134. ISBN: 978-3-319-03436-2. DOI: 10.1007/978-3-319-03437-9_13.

Abstract: Recording provenance is a key requirement for data-centric scholarship, allowing researchers to evaluate the integrity of source data sets and reproduce, and thereby, validate results. Provenance has become even more critical in the web environment in which data from distributed sources and of varying integrity can be combined and derived. Recent work by the W3C on the PROV model provides the foundation for semantically-rich, interoperable, and web-compatible provenance metadata. We apply that model to complex, but characteristic, provenance examples of social science data, describe scenarios that make scholarly use of those provenance descriptions, and propose a manner for encoding this provenance metadata within the widely-used DDI metadata standard.

Li et al.: One Permutation Hashing

Ping Li, Art Owen, and Cun-Hui Zhang. "One Permutation Hashing". In: *Advances in Neural Information Processing Systems 25*. Ed. by P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger. 2012, pp. 3122–3130. URL: http://papers.nips.cc/paper/4778-one-permutation-hashing.

Abstract: While minwise hashing is promising for large-scale learning in massive binary data, the preprocessing cost is prohibitive as it requires applying (e.g.,) k=500 permutations on the data. The testing time is also expensive if a new data point (e.g., a new document or a new image) has not been processed. In this paper, we develop a simple **one permutation hashing** scheme to address this important issue. While it is true that the preprocessing step can be parallelized, it comes at the cost of additional hardware and implementation. Also, reducing k permutations to just one would be much more **energy-efficient**, which might be an important perspective as minwise hashing is commonly deployed in the search industry. While the theoretical probability analysis is interesting, our experiments on similarity estimation and SVM & logistic regression also confirm the theoretical results.

File: 4778-one-permutation-hashing.pdf:http\://papers.nips.cc/paper/4778-one-permutation-hashing.pdf:PDF.

Li et al.: b-Bit Minwise Hashing in Practice

Ping Li, Anshumali Shrivastava, and Arnd Christian König. "b-Bit Minwise Hashing in Practice". In: *Internetware 2013*. 2013. URL: http://www.nudt.edu.cn/internetware2013/.

Abstract: Minwise hashing is a standard technique in the context of search for approximating set similarities. The recent work [26, 32] demonstrated a potential use of b-bit minwise

hashing [23, 24] for efficient search and learning on massive, high-dimensional, binary data (which are typical for many applications in Web search and text mining). In this paper, we focus on a number of critical issues which must be addressed before one can apply b-bit minwise hashing to the volumes of data often used industrial applications. Minwise hashing requires an expensive preprocessing step that computes k (e.g., 500) minimal values after applying the corresponding permutations for each data vector. We developed a parallelization scheme using GPUs and observed that the preprocessing time can be reduced by a factor of 20 80 and becomes substantially smaller than the data loading time. Reducing the preprocessing time is highly beneficial in practice, e.g., for duplicate Web page detection (where minwise hashing is a major step in the crawling pipeline) or for increasing the testing speed of online classifiers. Another critical issue is that for very large data sets it becomes impossible to store a (fully) random permutation matrix, due to its space requirements. Our paper is the first study to demonstrate that b-bit minwise hashing implemented using simple hash functions, e.g., the 2-universal (2U) and 4-universal (4U) hash families, can produce very similar learning results as using fully random permutations. Experiments on datasets of up to 200GB are presented.

File: a13-li.pdf:http\://ecommons.library.cornell.edu/bitstream/1813/37986/2/a13-li.pdf:PDF.

Li et al.: GPU-based minwise hashing: GPU-based minwise hashing

Ping Li, Anshumali Shrivastava, and Arnd Christian König. "GPU-based minwise hashing: GPU-based minwise hashing". In: *Proceedings of the 21st World Wide Web Conference (WWW 2012) (Companion Volume)*. 2012, pp. 565–566. DOI: 10.1145/2187980.2188129.

Abstract: Minwise hashing is a standard technique for efficient set similarity estimation in the context of search. The recent work of b-bit minwise hashing provided a substantial improvement by storing only the lowest b bits of each hashed value. Both minwise hashing and b-bit minwise hashing require an expensive preprocessing step for applying k (e.g., k=500) permutations on the entire data in order to compute k minimal values as the hashed data. In this paper, we developed a parallelization scheme using GPUs, which reduced the processing time by a factor of 20-80. Reducing the preprocessing time is highly beneficial in practice, for example, for duplicate web page detection (where minwise hashing is a major step in the crawling pipeline) or for increasing the testing speed of online classifiers (when the test data are not preprocessed).

Li et al.: Entropy Estimations Using Correlated Symmetric Stable Random Projections

Ping Li and Cun-Hui Zhang. "Entropy Estimations Using Correlated Symmetric Stable Random Projections". In: *Advances in Neural Information Processing Systems 25*. Ed. by P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger. 2012, pp. 3185–

3193. URL: http://papers.nips.cc/paper/4667-entropy-estimations-using-correlated-symmetric-stable-random-projections.

Abstract: Methods for efficiently estimating the Shannon entropy of data streams have important applications in learning, data mining, and network anomaly detections (e.g., the DDoS attacks). For nonnegative data streams, the method of Compressed Counting (CC) based on maximally-skewed stable random projections can provide accurate estimates of the Shannon entropy using small storage. However, CC is no longer applicable when entries of data streams can be below zero, which is a common scenario when comparing two streams. In this paper, we propose an algorithm for entropy estimation in general data streams which allow negative entries. In our method, the Shannon entropy is approximated by the finite difference of two correlated frequency moments estimated from correlated samples of symmetric stable random variables. Our experiments confirm that this method is able to substantially better approximate the Shannon entropy compared to the prior state-of-the-art.

File: 4667-entropy-estimations-using-correlated-symmetric-stable-random-projections.pdf:http\://papers.nips.cc/paper/4667-entropy-estimations-using-correlated-symmetric-stable-random-projections.pdf:PDF.

Li et al.: Exact Sparse Recovery with L0 Projections

Ping Li and Cun-Hui Zhang. "Exact Sparse Recovery with L0 Projections". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA: ACM, 2013, pp. 302–310. ISBN: 978-1-4503-2174-7. DOI: 10.1145/2487575.2487694.

Abstract: Many applications (e.g., anomaly detection) concern sparse signals. This paper focuses on the problem of recovering a K-sparse signal x R/1×N, i.e., K « N and N/i=1 1xi 0 = K. In the mainstream framework of compressed sensing (CS), × is recovered from M linear measurements y = xS R/1×M, where S RN×M is often a Gaussian (or Gaussian-like) design matrix. In our proposed method, the design matrix S is generated from an -stable distribution with 0. Our decoding algorithm mainly requires one linear scan of the coordinates, followed by a few iterations on a small number of coordinates which are "undetermined" in the previous iteration. Our practical algorithm consists of two estimators. In the first iteration, the (absolute) minimum estimator is able to filter out a majority of the zero coordinates. The gap estimator, which is applied in each iteration, can accurately recover the magnitudes of the nonzero coordinates. Comparisons with linear programming (LP) and orthogonal matching pursuit (OMP) demonstrate that our algorithm can be significantly faster in decoding speed and more accurate in recovery quality, for the task of exact spare recovery. Our procedure is robust against measurement noise. Even when there are no sufficient measurements, our algorithm can still reliably recover a significant portion of the nonzero coordinates.

McKinney et al.: Total Error and Variability Measures with Integrated Disclosure Limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in On The Map

Kevin L. McKinney, Andrew S. Green, Lars Vilhuber, and John M. Abowd. *Total Error and Variability Measures with Integrated Disclosure Limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in On The Map*. Working Papers 17-71. Center for Economic Studies, U.S. Census Bureau, 2017. URL: https://ideas.repec.org/p/cen/wpaper/17-71.html.

Abstract: We report results from the rst comprehensive total quality evaluation of five major indicators in the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program Quarterly Workforce Indicators (QWI): total employment, beginning-of-quarter employment, full-quarter employment, total payroll, 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 (LODES) workplace reports as displayed in OnTheMap (OTM). The evaluation is conducted by generating multiple threads of the edit and imputation models used in the LEHD Infrastructure File System. These threads conform to the Rubin (1987) multiple imputation model, with each thread or implicate being the output of formal probability models that address coverage, edit, and imputation errors. Design-based sampling variability and nite 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 five publication variables under study are estimated very accurately for tabulations involving at least 10 jobs. Tabulations involving three to nine jobs have quality in the range generally deemed acceptable. Tabulations involving zero, one or two jobs, which are generally suppressed in the QWI and synthesized in LODES, 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 based on a similar number of workers.

Miranda et al.: Using Partially Synthetic Data to Replace Suppression in the Business Dynamics Statistics: Early Results

Javier Miranda and Lars Vilhuber. "Using Partially Synthetic Data to Replace Suppression in the Business Dynamics Statistics: Early Results". In: *Privacy in Statistical Databases*. Ed. by Josep Domingo-Ferrer. Vol. 8744. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 232–242. ISBN: 978-3-319-11256-5. DOI: 10.1007/978-3-319-11257-2_18.

Abstract: The Business Dynamics Statistics is a product of the U.S. Census Bureau that provides measures of business openings and closings, and job creation and destruction, by a variety of cross-classifications (firm and establishment age and size, industrial sector, and

geography). Sensitive data are currently protected through suppression. However, as additional tabulations are being developed, at ever more detailed geographic levels, the number of suppressions increases dramatically. This paper explores the option of providing public-use data that are analytically valid and without suppressions, by leveraging synthetic data to replace observations in sensitive cells.

Miranda et al.: Using partially synthetic microdata to protect sensitive cells in business statistics

Javier Miranda and Lars Vilhuber. "Using partially synthetic microdata to protect sensitive cells in business statistics". In: *Statistical Journal of the IAOS* 32.1 (2016), pp. 69–80. DOI: 10.3233/SJI-160963.

Abstract: We describe and analyze a method that blends records from both observed and synthetic microdata into public-use tabulations on establishment statistics. The resulting tables use synthetic data only in potentially sensitive cells. We describe different algorithms, and present preliminary results when applied to the Census Bureau's Business Dynamics Statistics and Synthetic Longitudinal Business Database, highlighting accuracy and protection afforded by the method when compared to existing public-use tabulations (with suppressions).

File: :MirandaVilhuber-SJIAOS2016.pdf:PDF.

Reeder et al.: Codebook for the SIPP Synthetic Beta 7.0 (DDI-C and PDF)

Lori B. Reeder, Jordan C. Stanley, and Lars Vilhuber. *Codebook for the SIPP Synthetic Beta 7.0 (DDI-C and PDF)*. Codebook. Labor Dynamics Institute. Cornell University, 2018. DOI: 10.5281/zenodo.1477097.

Abstract: The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receipt, and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA and the IRS. The purpose of the SSB is to provide access to linked data that are usually not publicly available due to confidentiality concerns. To overcome these concerns, Census has synthesized, or modeled, all the variables in a way that changes the record of each individual in a manner designed to preserve the underlying covariate relationships between the variables. The only variables that were not altered by the synthesis process and still contain their original values are gender and a link to the first reported marital partner in the survey. Eight SIPP panels (1990, 1991, 1992, 1993, 1996, 2001, 2004, 2008) form the basis for the SSB, with a large subset of variables available across all the panels selected for inclusion and harmonization across the years. Administrative data were added and some editing was done to correct for logical inconsistencies in the IRS/SSA earnings and benefits data.

Reeder et al.: Codebook for the SIPP Synthetic Beta 7.0 (PDF version)

Lori B. Reeder, Jordan C. Stanley, and Lars Vilhuber. *Codebook for the SIPP Synthetic Beta 7.0 (PDF version)*. PDF and DDI code V20181102b-pdf. Cornell Institute for Social and Economic Research and Labor Dynamics Institute. Cornell University, 2018. DOI: 10.5281/zenodo.1477099.

Abstract: The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receipt, and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA and the IRS. The purpose of the SSB is to provide access to linked data that are usually not publicly available due to confidentiality concerns. To overcome these concerns, Census has synthesized, or modeled, all the variables in a way that changes the record of each individual in a manner designed to preserve the underlying covariate relationships between the variables. The only variables that were not altered by the synthesis process and still contain their original values are gender and a link to the first reported marital partner in the survey. Eight SIPP panels (1990, 1991, 1992, 1993, 1996, 2001, 2004, 2008) form the basis for the SSB, with a large subset of variables available across all the panels selected for inclusion and harmonization across the years. Administrative data were added and some editing was done to correct for logical inconsistencies in the IRS/SSA earnings and benefits data.

Reeder et al.: Codebook for the SIPP Synthetic Beta v7 [Online]

Lori B. Reeder, Jordan C. Stanley, and Lars Vilhuber. *Codebook for the SIPP Synthetic Beta v7 [Online]*. 2018. URL: http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb/v/v7.

Reeder et al.: Codebook for the SIPP Synthetic Beta v6.0.2 [Online]

Lori B. Reeder, Martha Stinson, Kelly E. Trageser, and Lars Vilhuber. *Codebook for the SIPP Synthetic Beta v6.0.2 [Online]*. 2015. URL: http://www2.ncrn.cornell.edu/ced2ar-web/codebooks/ssb/v/v602.

Schneider et al.: A new method for protecting interrelated time series with Bayesian prior distributions and synthetic data

Matthew J. Schneider and John M. Abowd. "A new method for protecting interrelated time series with Bayesian prior distributions and synthetic data". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2015), n/a–n/a. DOI: 10.1111/rssa.12100.

Abstract: Organizations disseminate statistical summaries of administrative data via the Web for unrestricted public use. They balance the trade-off between protection of confidentiality and quality of inference. Recent developments in disclosure avoidance techniques include the incorporation of synthetic data, which capture the essential features of underlying data by releasing altered data generated from a posterior predictive distribution. The US Census Bureau collects millions of interrelated time series microdata that are hierarchical and contain many 0s and suppressions. Rule-based disclosure avoidance techniques often require the suppression of count data for small magnitudes and the modification of data based on a small number of entities. Motivated by this problem, we use zero-inflated extensions of Bayesian generalized linear mixed models with privacy-preserving prior distributions to develop methods for protecting and releasing synthetic data from time series about thousands of small groups of entities without suppression based on the magnitudes or number of entities. We find that, as the prior distributions of the variance components in the Bayesian generalized linear mixed model become more precise towards zero, protection of confidentiality increases and the quality of inference deteriorates. We evaluate our methodology by using a strict privacy measure, empirical differential privacy and a newly defined risk measure, the probability of range identification, which directly measures attribute disclosure risk. We illustrate our results with the US Census Bureau's quarterly workforce indicators.

Shrivastava et al.: Beyond Pairwise: Provably Fast Algorithms for Approximate k-Way Similarity Search

Anshumali Shrivastava and Ping Li. "Beyond Pairwise: Provably Fast Algorithms for Approximate k-Way Similarity Search". In: *Advances in Neural Information Processing Systems 26*. Ed. by C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger. Curran Associates, Inc., 2013, pp. 791–799. URL: http://papers.nips.cc/paper/5216-beyond-pairwise-provably-fast-algorithms-for-approximate-k-way-similarity-search/.

Abstract: We go beyond the notion of pairwise similarity and look into search problems with k-way similarity functions. In this paper, we focus on problems related to 3-way Jaccard similarity: R3way = |S1 S2 S3| |S1 S2 S3| , S1, S2, S3 C, where C is a size n collection of sets (or binary vectors). We show that approximate R3way similarity search problems admit fast algorithms with provable guarantees, analogous to the pairwise case. Our analysis and speedup guarantees naturally extend to k-way resemblance. In the process, we extend traditional framework of locality sensitive hashing (LSH) to handle higher-order similarities, which could be of independent theoretical interest. The applicability of R3way search is shown on the "Google Sets" application. In addition, we demonstrate the advantage of R3way resemblance over the pairwise case in improving retrieval quality.

File: 5216-beyond-pairwise-provably-fast-algorithms-for-approximate-k-way-similarity-search.pdf:http\://papers.nips.cc/paper/5216-beyond-pairwise-provably-fast-algorithms-for-approximate-k-way-similarity-search.pdf:PDF.

Shrivastava et al.: Fast Near Neighbor Search in High-Dimensional Binary Data

Anshumali Shrivastava and Ping Li. "Fast Near Neighbor Search in High-Dimensional Binary Data". In: *The European Conference on Machine Learning (ECML 2012)*. 2012. URL: http://www.ecmlpkdd2012.net/.

Abstract: Abstract. Numerous applications in search, databases, machine learning, and computer vision, can benefit from efficient algorithms for near neighbor search. This paper proposes a simple framework for fast near neighbor search in high-dimensional binary data, which are common in practice (e.g., text). We develop a very simple and effective strategy for sub-linear time near neighbor search, by creating hash tables directly using the bits generated by b-bit minwise hashing. The advantages of our method are demonstrated through thorough comparisons with two strong baselines: spectral hashing and sign (1-bit) random projections.

File: 1125548.pdf:http\://www.cs.bris.ac.uk/~flach/ECMLPKDD2012papers/1125548.pdf:PDF.

Shrivastava et al.: Graph Kernels via Functional Embedding

Anshumali Shrivastava and Ping Li. "Graph Kernels via Functional Embedding". In: *CoRR* abs/1404.5214 (2014). URL: http://arxiv.org/abs/1404.5214.

Abstract: We propose a representation of graph as a functional object derived from the power iteration of the underlying adjacency matrix. The proposed functional representation is a graph invariant, i.e., the functional remains unchanged under any reordering of the vertices. This property eliminates the difficulty of handling exponentially many isomorphic forms. Bhattacharyya kernel constructed between these functionals significantly outperforms the state-of-the-art graph kernels on 3 out of the 4 standard benchmark graph classification datasets, demonstrating the superiority of our approach. The proposed methodology is simple and runs in time linear in the number of edges, which makes our kernel more efficient and scalable compared to many widely adopted graph kernels with running time cubic in the number of vertices.

Shrivastava et al.: In Defense of MinHash Over SimHash

Anshumali Shrivastava and Ping Li. "In Defense of MinHash Over SimHash". In: *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 33. Reykjavik, Iceland, 2014. URL: http://jmlr.org/proceedings/papers/v33/shrivastava14.html.

Abstract: MinHash and SimHash are the two widely adopted Locality Sensitive Hashing (LSH) algorithms for large-scale data processing applications. Deciding which LSH to use for a particular problem at hand is an important question, which has no clear answer in the existing literature. In this study, we provide a theoretical answer (validated by experiments) that MinHash virtually always outperforms SimHash when the data are binary, as common

in practice such as search. The collision probability of MinHash is a function of resemblance similarity (R), while the collision probability of SimHash is a function of cosine similarity (S). To provide a common basis for comparison, we evaluate retrieval results in terms of S for both MinHash and SimHash. This evaluation is valid as we can prove that MinHash is a valid LSH with respect to S, by using a general inequality S2 R S2−S. Our worst case analysis can show that MinHash significantly outperforms SimHash in high similarity region. Interestingly, our intensive experiments reveal that MinHash is also substantially better than SimHash even in datasets where most of the data points are not too similar to each other. This is partly because, in practical data, often R Sz−S holds where z is only slightly larger than 2 (e.g., z 2.1). Our restricted worst case analysis by assuming Sz−S R S2−S shows that MinHash indeed significantly outperforms SimHash even in low similarity region. We believe the results in this paper will provide valuable guidelines for search in practice, especially when the data are sparse.

File: shrivastava14.pdf:http\://jmlr.org/proceedings/papers/v33/shrivastava14.pdf:PDF.

Srivastava et al.: Testing for Membership to the IFRA and the NBU Classes of Distributions

Radhendushka Srivastava, Ping Li, and Debasis Sengupta. "Testing for Membership to the IFRA and the NBU Classes of Distributions". In: *Journal of Machine Learning Research - Proceedings Track for the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)* 22 (2012), pp. 1099–1107. URL: http://www.jmlr.org/proceedings/papers/v22/srivastava12.html.

Abstract: This paper provides test procedures to determine whether the probability distribution underlying a set of non-negative valued samples belongs to the Increasing Failure Rate Average (IFRA) class or the New Better than Used (NBU) class. Membership of a distribution to one of these classes is known to have implications which are important in reliability, queuing theory, game theory and other disciplines. Our proposed test is based on the Kolmogorov-Smirnov distance between an empirical cumulative hazard function and its best approximation from the class of distributions constituting the null hypothesis. It turns out that the least favorable distribution, which produces the largest probability of Type I error of each of the tests, is the exponential distribution. This fact is used to produce an appropriate cut-off or p-value. Monte Carlo simulations are conducted to check small sample size (i.e., significance) and power of the test. Usefulness of the test is illustrated through the analysis of a set of monthly family expenditure data collected by the National Sample Survey Organization of the Government of India.

File: srivastava12.pdf:http\://www.jmlr.org/proceedings/papers/v22/srivastava12/srivastava12.pdf:PDF.

Sun et al.: Fast Multi-task Learning for Query Spelling Correction

Xu Sun, Anshumali Shrivastava, and Ping Li. "Fast Multi-task Learning for Query Spelling Correction". In: *The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*. 2012, pp. 285–294. DOI: 10.1145/2396761.2396800.

Abstract: In this paper, we explore the use of a novel online multi-task learning framework for the task of search query spelling correction. In our procedure, correction candidates are initially generated by a ranker-based system and then re-ranked by our multi-task learning algorithm. With the proposed multi-task learning method, we are able to effectively transfer information from different and highly biased training datasets, for improving spelling correction on all datasets. Our experiments are conducted on three query spelling correction datasets including the well-known TREC benchmark dataset. The experimental results demonstrate that our proposed method considerably outperforms the existing baseline systems in terms of accuracy. Importantly, the proposed method is about one order of magnitude faster than baseline systems in terms of training speed. Compared to the commonly used online learning methods which typically require more than (e.g.,) 60 training passes, our proposed method is able to closely reach the empirical optimum in about 5 passes.

Sun et al.: Query spelling correction using multi-task learning

Xu Sun, Anshumali Shrivastava, and Ping Li. "Query spelling correction using multi-task learning". In: *Proceedings of the 21st World Wide Web Conference (WWW 2012)(Companion Volume)*. 2012, pp. 613–614. DOI: 10.1145/2187980.2188153.

Abstract: This paper explores the use of online multi-task learning for search query spelling correction, by effectively transferring information from different and biased training datasets for improving spelling correction across datasets. Experiments were conducted on three query spelling correction datasets, including the well-known TREC benchmark data. Our experimental results demonstrate that the proposed method considerably outperforms existing baseline systems in terms of accuracy. Importantly, the proposed method is about one-order of magnitude faster than baseline systems in terms of training speed. In contrast to existing methods which typically require more than (e.g.,) 50 training passes, our algorithm can very closely approach the empirical optimum in around five passes.

Vilhuber: ncrncornell/ced2ar-nber-ces-codebook: Codebook for NBER-CES Manufacturing Industry Database

Lars Vilhuber. *ncrncornell/ced2ar-nber-ces-codebook: Codebook for NBER-CES Manufacturing Industry Database*. PDF and DDI code. Labor Dynamics Institute. Cornell University, 2015. DOI: 10.5281/zenodo.2527908.

Abstract: 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/.

Vilhuber: ncrncornell/ced2ar-nqwi-codebook: Codebook for the National QWI [Codebook file]

Lars Vilhuber. *ncrncornell/ced2ar-nqwi-codebook: Codebook for the National QWI [Codebook file]*. PDF and DDI code. Labor Dynamics Institute. Cornell University, 2015. DOI: 10.5281/zenodo.2527906.

Abstract: Codebook for the early research version of National QWI. Live version of the DDI codebook at https://www2.ncrn.cornell.edu/ced2ar-web/codebooks/nqwi/.

Vilhuber: ncrncornell/ced2ar-synlbd-codebook: DDI Codebook for the Synthetic LBD

Lars Vilhuber. *ncrncornell/ced2ar-synlbd-codebook: DDI Codebook for the Synthetic LBD*. PDF and DDI code. Labor Dynamics Institute. Cornell University, 2016. DOI: 10.5281/zenodo.2527910.

Abstract: 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/.

Vilhuber et al.: Synthetic establishment microdata around the world

Lars Vilhuber, John M. Abowd, and Jerome P. Reiter. "Synthetic establishment microdata around the world". In: *Statistical Journal of the IAOS* 32.1 (2016), pp. 65–68. DOI: 10.3233/SJI-160964.

Abstract: In contrast to the many public-use microdata samples available for individual and household data from many statistical agencies around the world, there are virtually no establishment or firm microdata available. In large part, this difficulty in providing access to business microdata is due to the skewed and sparse distributions that characterize business data. Synthetic data are simulated data generated from statistical models. We organized sessions at the 2015 World Statistical Congress and the 2015 Joint Statistical Meetings, highlighting work on synthetic *establishment* microdata. This overview situates those papers, published in this issue, within the broader literature.

File: :VilhuberAbowdReiter-SJIAOS2016.pdf:PDF.

Vilhuber et al.: Proceedings from the Synthetic LBD International Seminar

Lars Vilhuber, Saki Kinney, and Ian Schmutte. *Proceedings from the Synthetic LBD International Seminar*. Document 44. Labor Dynamics Institute, Cornell University, 2017. URL: http://digitalcommons.ilr.cornell.edu/ldi/44/.

Abstract: On May 9, 2017, we hosted a seminar to discuss the conditions necessary to implement the SynLBD approach with interested parties, with the goal of providing a straightforward toolkit to implement the same procedure on other data. The proceedings summarize the discussions during the workshop. Funding for the workshop was provided by the National Science Foundation (Grants 1012593; 1131848) and the Alfred P. Sloan Foundation (G-2015-13903). Organizational support was provided by the Labor Dynamics Institute at Cornell University.

Vilhuber et al.: Making Confidential Data Part of Reproducible Research

Lars Vilhuber and Carl Lagoze. *Making Confidential Data Part of Reproducible Research*. Document 41. Labor Dynamics Institute, Cornell University, 2017. URL: http://digitalcommons.ilr.cornell.edu/ldi/41/.

Vilhuber et al.: Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy

Lars Vilhuber and Ian Schmutte. *Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy*. Preprint 1813:46197. Cornell University, 2017. URL: http://hdl.handle.net/1813/46197.

Abstract: Proceedings from the 2016 NSF–Sloan Workshop on Practical Privacy Vilhuber, Lars; Schmutte, Ian; Abowd, John M. On October 14, 2016, we hosted a workshop that brought together economists, survey statisticians, and computer scientists with expertise in the field of privacy preserving methods: Census Bureau staff working on implementing cutting-edge methods in the Bureau's flagship public-use products mingled with academic researchers from a variety of universities. The four products discussed as part of the workshop were 1. the American Community Survey (ACS); 2. Longitudinal Employer-Household Data (LEHD), in particular the LEHD Origin-Destination Employment Statistics (LODES); the 3. 2020 Decennial Census; and the 4. 2017 Economic Census. The goal of the workshop was to 1. Discuss the specific challenges that have arisen in ongoing efforts to apply formal privacy models to Census data products by drawing together expertise of academic and governmental researchers 2. Produce short written memos that summarize concrete suggestions for practical applications to specific Census Bureau priority areas.

Vilhuber et al.: Proceedings from the 2017 Cornell-Census-NSF-Sloan Workshop on Practical Privacy

Lars Vilhuber and Ian Schmutte. *Proceedings from the 2017 Cornell-Census-NSF-Sloan*

Workshop on Practical Privacy. Document 43. Labor Dynamics Institute, Cornell University, 2017. URL: http://digitalcommons.ilr.cornell.edu/ldi/43/.

Abstract: These proceedings report on a workshop hosted at the U.S. Census Bureau on May 8, 2017. Our purpose was to gather experts from various backgrounds together to continue discussing the development of formal privacy systems for Census Bureau data products. This workshop was a successor to a previous workshop held in October 2016 (Vilhuber and Schmutte 2017). At our prior workshop, we hosted computer scientists, survey statisticians, and economists, all of whom were experts in data privacy. At that time we discussed the practical implementation of cutting-edge methods for publishing data with formal, provable privacy guarantees, with a focus on applications to Census Bureau data products. The teams developing those applications were just starting out when our first workshop took place, and we spent our time brainstorming solutions to the various problems researchers were encountering, or anticipated encountering. For these cutting-edge formal privacy models, there had been very little effort in the academic literature to apply those methods in real-world settings with large, messy data. We therefore brought together an expanded group of specialists from academia and government who could shed light on technical challenges, subject matter challenges and address how data users might react to changes in data availability and publishing standards. In May 2017, we organized a follow-up workshop, which these proceedings report on. We reviewed progress made in four different areas. The four topics discussed as part of the workshop were 1. the 2020 Decennial Census; 2. the American Community Survey (ACS); 3. the 2017 Economic Census; 4. measuring the demand for privacy and for data quality. As in our earlier workshop, our goals were to 1. Discuss the specific challenges that have arisen in ongoing efforts to apply formal privacy models to Census data products by drawing together expertise of academic and governmental researchers; 2. Produce short written memos that summarize concrete suggestions for practical applications to specific Census Bureau priority areas.

Vilhuber et al.: Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy

Lars Vilhuber and Ian M. Schmutte. *Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy*. Document 33. Labor Dynamics Institute, Cornell University, 2017. URL: http://digitalcommons.ilr.cornell.edu/ldi/33/.

Abstract: On October 14, 2016, we hosted a workshop that brought together economists, survey statisticians, and computer scientists with expertise in the field of privacy preserving methods: Census Bureau staff working on implementing cutting-edge methods in the Bureau?s flagship public-use products mingled with academic researchers from a variety of universities. The four products discussed as part of the workshop were 1. the American Community Survey (ACS); 2. Longitudinal Employer-Household Data (LEHD), in particular the LEHD Origin-Destination Employment Statistics (LODES); the 3. 2020 Decennial Census; and the 4. 2017 Economic Census. The goal of the workshop was to 1. Discuss the specific challenges that have arisen in ongoing efforts to apply formal privacy models to Census data

products by drawing together expertise of academic and governmental researchers 2. Produce short written memos that summarize concrete suggestions for practical applications to specific Census Bureau priority areas. Funding for the workshop was provided by the National Science Foundation (CNS-1012593) and the Alfred P. Sloan Foundation. Organizational support was provided by the Research and Methodology Directorate at the U.S. Census Bureau and the Labor Dynamics Institute at Cornell University.

Weinberg et al.: Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Improve the U.S. Statistical System?

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 Smyth, Leen-Kiat Soh, Bruce D. Spencer, Seth E. Spielman, Lars Vilhuber, and Christopher K. Wikle. *Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Improve the U.S. Statistical System?* Working Papers 17-59r. Center for Economic Studies, U.S. Census Bureau, 2017. URL: https://ideas.repec.org/p/cen/wpaper/17-59r.html.

Abstract: The National Science Foundation-Census Bureau Research Network (NCRN) was established in 2011 to create interdisciplinary research nodes on methodological questions of interest and significance to the broader research community and to the Federal Statistical System (FSS), particularly the Census Bureau. The activities to date have covered both fundamental and applied statistical research and have focused at least in part on the training of current and future generations of researchers in skills of relevance to surveys and alternative measurement of economic units, households, and persons. This paper discusses some of the key research findings of the eight nodes, organized into six topics: (1) Improving census and survey data collection methods; (2) Using alternative sources of data; (3) Protecting privacy and confidentiality by improving disclosure avoidance; (4) Using spatial and spatio-temporal statistical modeling to improve estimates; (5) Assessing data cost and quality tradeoffs; and (6) Combining information from multiple sources. It also reports on collaborations across nodes and with federal agencies, new software developed, and educational activities and outcomes. The paper concludes with an evaluation of the ability of the FSS to apply the NCRN's research outcomes and suggests some next steps, as well as the implications of this research-network model for future federal government renewal initiatives.

Weinberg et al.: Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Improve the U.S. Statistical System?

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 Smyth, Leen-Kiat Soh, Bruce D. Spencer, Seth E. Spielman, Lars Vilhuber, and Christopher K. Wikle. "Effects of a Government-Academic Partnership: Has the NSF-Census Bureau Research Network Helped Improve the U.S. Statistical System?" In: *Journal of Survey Statistics and Methodology* (2018), smy023. DOI: 10.1093/jssam/smy023.

Abstract: The National Science Foundation-Census Bureau Research Network (NCRN) was established in 2011 to create interdisciplinary research nodes on methodological questions of interest and significance to the broader research community and to the Federal Statistical System (FSS), particularly the Census Bureau. The activities to date have covered both fundamental and applied statistical research and have focused at least in part on the training of current and future generations of researchers in skills of relevance to surveys and alternative measurement of economic units, households, and persons. This paper discusses some of the key research findings of the eight nodes, organized into six topics: (1) Improving census and survey data collection methods; (2) Using alternative sources of data; (3) Protecting privacy and confidentiality by improving disclosure avoidance; (4) Using spatial and spatio-temporal statistical modeling to improve estimates; (5) Assessing data cost and quality tradeoffs; and (6) Combining information from multiple sources. It also reports on collaborations across nodes and with federal agencies, new software developed, and educational activities and outcomes. The paper concludes with an evaluation of the ability of the FSS to apply the NCRN's research outcomes and suggests some next steps, as well as the implications of this research-network model for future federal government renewal initiatives.