Usage and outcomes of the Synthetic Data Server

Lars Vilhuber[1] John M. Abowd[23]

October 30, 2015

THIS VERSION PRELIMINARY AND INCOMPLETE: PLEASE DO NOT CITE

[1]Labor Dynamics Institute, Cornell University
[2]Labor Dynamics Institute, Cornell University
[3]Both authors gratefully acknowledge funding from NSF under Grants 1042181 and 0941226, as well as funding from the Alfred P. Sloan Foundation.

Abstract

The Synthetic Data Server (SDS) at Cornell University was set up to provide early access to new synthetic data products by the U.S. Census Bureau. These datasets are made available to interested researchers in a controlled environment, prior to a more generalized release. Over the past 5 years, 4 synthetic datasets were made available on the server, and over 100 users have accessed the server over that time period. This paper reports on interim outcomes of the activity: results of validation requests from a user perspective, functioning of the feedback loop due to validation and user input, and the role of the SDS as a access gateway to and educational tool for other mechanisms of accessing detailed person, household, establishment, and firm statistics.

1 History of the Synthetic Data Server

The Synthetic Data Server (SDS)[1] was set up to provide early access to new synthetic data products by the U.S. Census Bureau. These datasets are made available to interested researchers in a controlled environment, prior to a more generalized release. Following the award of NSF grant SES-1042181 (Vilhuber and Abowd, 2010) in September 2010, the SDS replaced a more limited SIPP Synthetic Beta (SSB) server in December 2010, expanding functionality, computing power, and data access. A hardware upgrade in June 2014, funded through NSF Grant BCS-0941226 (Abowd, 2010) ensured longer-term viability and increased computational capability, as the user base expanded and users pushed the types of analyses being performed with the data.

As of March 2015, the SDS provides access to three datasets:

- SIPP Synthetic Beta (SSB) v5.1 (released in 2013)
- SIPP Synthetic Beta (SSB) v6.0 (released in 2015)
- Synthetic LBD (SynLBD) v2. (released in 2011)

2 Access to the server

Access requests are reviewed for feasibility, but are not otherwise restricted. Once the data provider has reviewed the application for feasibility, the server provider (Cornell University) sets up accounts on the system, and provides users with instructions on how to gain access to the remote graphical desktop interface (using NX technology[2]). In order to prevent users from removing datasets from the server, requests for removal are moderated, but not censored. The server provides access to a variety of statistical software (SAS, Stata, R, Matlab, and others as requested), and is only restricted by the software available on the Census Bureau's validation servers (see below).

3 Datasets

3.1 SIPP Synthetic Beta (SSB)

The SDS was launched with SSB v5.0 (U.S. Census Bureau, 2011a) on December 1, 2010. SSB v5.1 (U.S. Census Bureau, 2013) was released on July 26, 2013, and SSB v6.0 (U.S. Census Bureau, 2015) was released in March 2015. The purpose[3] of the SIPP Synthetic Beta (SSB) is to provide access to linked data that are usually not publicly available due to confidentiality concerns. In order to overcome that barrier, Census Bureau staff economists

[1] http://www.vrdc.cornell.edu/news/synthetic-data-server/

[2] `http://www.nomachine.com`

[3] This section is derived from the Census Bureau's SSB website `http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html`.

and statisticians in collaboration with researchers at Cornell University, the Social Security Administration (SSA) and the Internal Revenue Service (IRS), first created an internal, confidential file that integrates person-level micro-data from a household survey (the Survey of Income and Program Participation (SIPP)) with administrative tax (IRS Form W-2 records) and benefit data (SSA records of retirement and disability benefit receipt). Some editing was done to correct for logical inconsistencies in the IRS/SSA earnings and benefits data. Data that are missing due to missing survey interviews or missing administrative data are multiply-imputed. The resulting data sets are called the **Completed Gold Standard Files** and contain all original, non-missing, confidential values and imputed values in place of originally missing data.

Then all variables are synthesized, or modeled, 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. The goal of the SSB is to produce results that are qualitatively the same as results from the Completed Gold Standard Files. The synthesis process itself involves estimating the joint distribution of all the variables in the data and taking random draws from this modeled distribution. These draws are then used to replace actual data values. This process is repeated multiple times to create a set of 16 files, also called implicates. For more details, see Benedetto et al. (2013) and the SSB website at the Census Bureau[4].

The creation of the SSB was funded by the U.S. Census Bureau and the Social Security Administration (SSA) with additional funding from NSF Grants SES-0427889 (Abowd et al., 2004b) and SES-0339191 (Abowd et al., 2004a), and is currently being maintained by the U.S. Census Bureau.

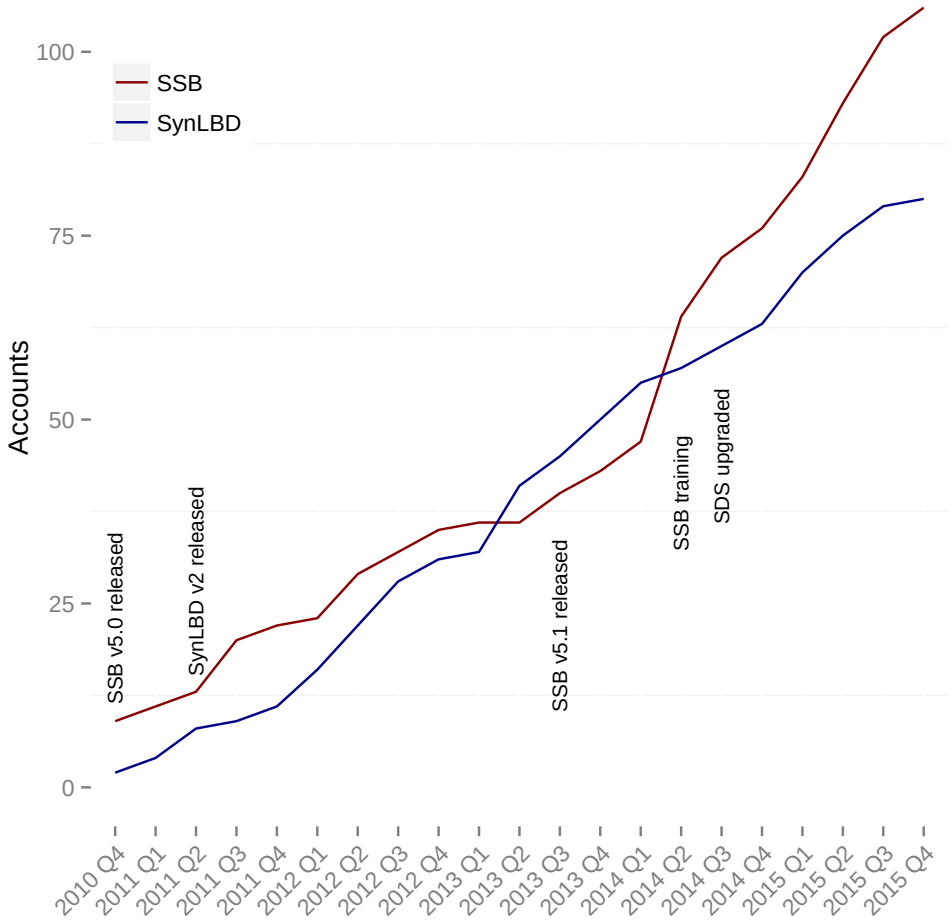
3.2 Synthetic Longitudinal Business Database (SynLBD)

The Synthetic Longitudinal Business Database (LBD) is produced by the U.S. Census Bureau in collaboration with Duke University, Cornell University, the National Institute of Statistical Sciences (NISS), the IRS and the National Science Foundation (NSF). The purpose of the SynLBD is to provide users with access to a longitudinal business data product that can be used outside of a secure Census Bureau facility, without disclosing confidential information. The SynLBD is created by synthesizing information from the (confidential) LBD (Miranda and Jarmin, 2002) on establishments' employment and payroll, establishments' birth and death years, and multi-unit status, conditional on industrial classification. Geography and firm-level information are not yet available on the SynLBD. The Census Bureau's Disclosure Review Board and their counterparts at IRS have reviewed the content of the file, and allowed the release of these data for public use. See Kinney et al. (2011b), Kinney et al. (2011a), and the SynLBD website at the U.S. Census Bureau[5] for detailed information on methods.

[4] `http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html`

[5] `http://www.census.gov/ces/dataproducts/synlbd/index.html`

Figure 1: Account creation and salient events



The creation of the LBD Synthetic Beta file v2 was funded by NSF Grant SES-0427889 (Abowd et al., 2004b), and continuing development is funded by the U.S. Census Bureau.

4 Usage and impact

Since the start of the Synthetic Data Server project, account growth has been steady (see Figure 1). As of 2015 Q4, 186 accounts had been created on the server. In general, the rate of account requests increased subsequent to conference presentations or after specific events, such as the 2014 workshop (held by the University of Michigan NCRN node in collaboration with the U.S. Census Bureau and Cornell University) dedicated to teaching graduate students on how to use and leverage the SSB.

4.1 Feedback loop

Both of the current data providers have incorporated feedback from users into their data products. The SSB with which the SDS was launched was itself the second public iteration, after the original release of v4.0. The growth of the supporting IT infrastructure, first from its predecessor to the original instantiation of the SDS, and subsequently in the 2014 IT upgrade mentioned earlier, reflected the growing interest that followed adaptations. In the case of the SSB, such feedback first lead to the incorporation of additional variables in v5.1 (U.S. Census Bureau, 2013), and subsequently to further enhancements in v6.0 (U.S. Census Bureau, 2015). In the case of the Synthetic LBD (SynLBD), only one iteration has so far been released on the SDS, but the key shortcomings in the structure of the SynLBD v2.0 (U.S. Census Bureau, 2011b) – the absence of North American Industry Classification System (NAICS) codes, of any sort of geographic detail, and of indicators of firm structure – have been reflected in the rejected access applications. As a reminder, access is granted when the technical requests are feasibly satisfied by the data on the SDS. In the case of the SynLBD, we can quantify additional data requested that lead to applications being rejected. Out of 100 applications for access to the SynLBD received through 2015-08-10, 21 (21%) were denied because they were not feasible using the synthetic data (this does not take into account applications that were partially feasible, which were generally approved). Of those denied,

- 6 (28.57%) had requested firm-level variables,
- 11 (52.38%) had requested data to perform conditional geographic analysis, and
- 1 (4.76%) had requested data for specific NAICS industries.

Such numbers, of course, do not take into account potential requestors that did not apply because a reading of the documentation revealed that the analysis was not feasible.

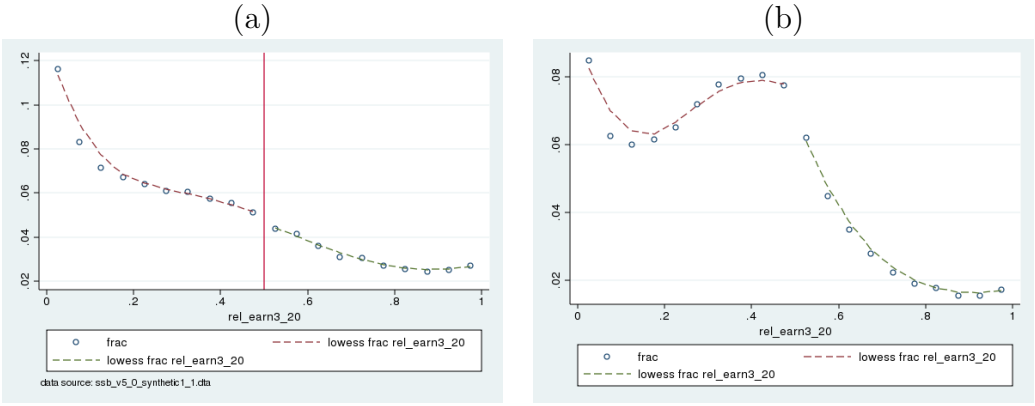
4.2 Validation requests

The data are by their nature preliminary, and users are discouraged from using results based solely on the synthetic data. Validation requests are encouraged and free, subject to following certain rules, outlined on the data providers' websites[6]. Generally, validation requires that users provide all programs and auxiliary input files, documentation of the results similar to a disclosure review request at Federal Statistical Research Data Center (FSRDC), and that all programs run error-free (replicability requirement). Results obtained from confidential data are subject to all the disclosure avoidance protocols in effect at the time of their release. That being said, the requirements are no more onerous than generic replication requirements, and turnaround may be as short as one week.

[6]SIPP Synthetic Beta (SSB) Website at the U.S. Census Bureau: `http://www.census.gov/programs-surveys/sipp/methodology/sipp-synthetic-beta-data-product.html`, Synthetic LBD (SynLBD) at the U.S. Census Bureau: `http://www.census.gov/ces/dataproducts/synlbd/`

An analysis of SynLBD validation requests was performed as of 2015-08-10. Out of 79 projects, 5 had requested validation. A recurring issue has been that users are unfamiliar with the constraints imposed by the validation procedure, in particular that all such validation requests are treated as an authorized release of results from an analysis of confidential data, and are thus subject to review by the Census Bureau's Disclosure Review Board (DRB). In particular, the *quantity* of results requested surpasses not the ability of the confidential servers to compute, but rather is judged to be too high by the rules of the DRB (60% of validation requests).

Figure 2: From Bertrand et al (2015), their Figure I



Note: See text for details on computation and provenance.

An example for a successful validation request combined with a cautionary note for users of the SSB is illustrated by Figure 2. In preparing Bertrand et al. (2015), the authors of that paper performed an analysis of the distribution of relative household income using a variety of datasets, including the SSB. They obtained fairly robust results across a variety of datasets and time (see their Figure III, reproduced here in Figure 3): there is a distinct break in the distribution of couples when the wife's income surpassed 50%. However, the analysis with the SSB produced a very different result, as illustrated in Figure 2, Panel (a): there was no such break. The authors requested validation, using the protocols described above, and which the Census Bureau was able to accomplish in a very short time. The Census Bureau ran the same models on the confidential data, subjected the proposed publication statistics to conventional statistical disclosure limitation (in this case just rounding and release in the form of a graph), and released Figure 2, Panel (b), corresponding to Figure I in Bertrand et al. (2015). The results obtained when their analysis was replicated against the confidential files yielded a different result, consistent with other datasets. The "success" alluded to earlier is on both sides of the interaction. The researchers were able to very quickly ascertain that their model, when tested against the confidential data, yielded a result in line with other results obtained from other data, and proceeded to publish their paper. The Census Bureau, in exchange, obtained valuable feedback on the quality of the synthesis models, which they were able to take into account for the next iteration of the data production cycle, and which is the statutory justification for the researchers' use of the validation process. The cautionary

note is that while useful for exploring the data and for testing models, not every model will yield valid results on the synthetic data.[7]

Figure 3: From Bertrand et al (2015)

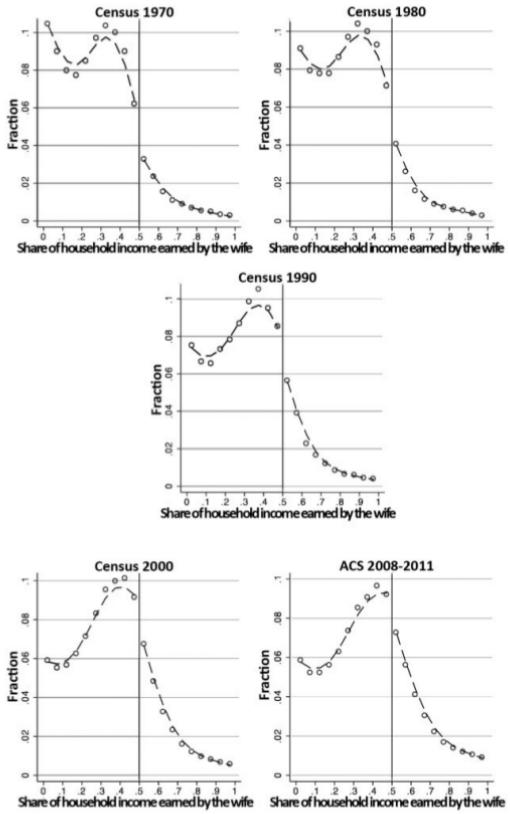


FIGURE III
Distribution of Relative Income over Time (Census Bureau Data)

More generally, the question as to the statistical precision of the results obtained from the synthetic data can be assessed. For this purpose, we replicated the key models of various SynLBD validation requests, and computed the overlap of parameter estimates as suggested by Karr et al. (2006). We compute the *interval overlap measure* $J_{k,m}$ for parameter k in model m. Consider the overlap of confidence intervals (L, U) for $\beta_{k,m}$ (estimated from the confidential data) and (L^*, U^*) for $\beta^*_{k,m}$ (from the synthetic data). Let $L^{over} = \max(L, L^*)$ and $U^{over} = \min(U, U^*)$. Then the average overlap in confidence intervals is

$$J^*_{k,m} = \frac{1}{2}\left[\frac{U^{over} - L^{over}}{U - L} + \frac{U^{over} - L^{over}}{U^* - L^*}\right]$$

We then average $J^*_{k,m}$ over all estimated models and parameters, by validation request. Table 1 presents results from 4 validation requests as of 2015-10-29. Validation results vary

[7]We thank Marianne Bertrand for allowing us to use this example, and for kindly having provided the graphs for from the analyses using the Gold Standard File (GSF) and the SSB.

widely. The correct counterfactual is running these validation requests against synthetic data that does not claim analytical validity, such as synthetic data generated from unidimensional distributions of variables. Results are pending.

Table 1: Confidence interval overlap $J^*_{k,m}$

User	Request	Mean	75th	90th	Max
A	1	0.160	0.246	0.725	0.889
A	2	0.101	0	0.523	0.924
B	1	0.869	1.000	1.000	1.000
C	1	0.219	0.509	0.725	0.995

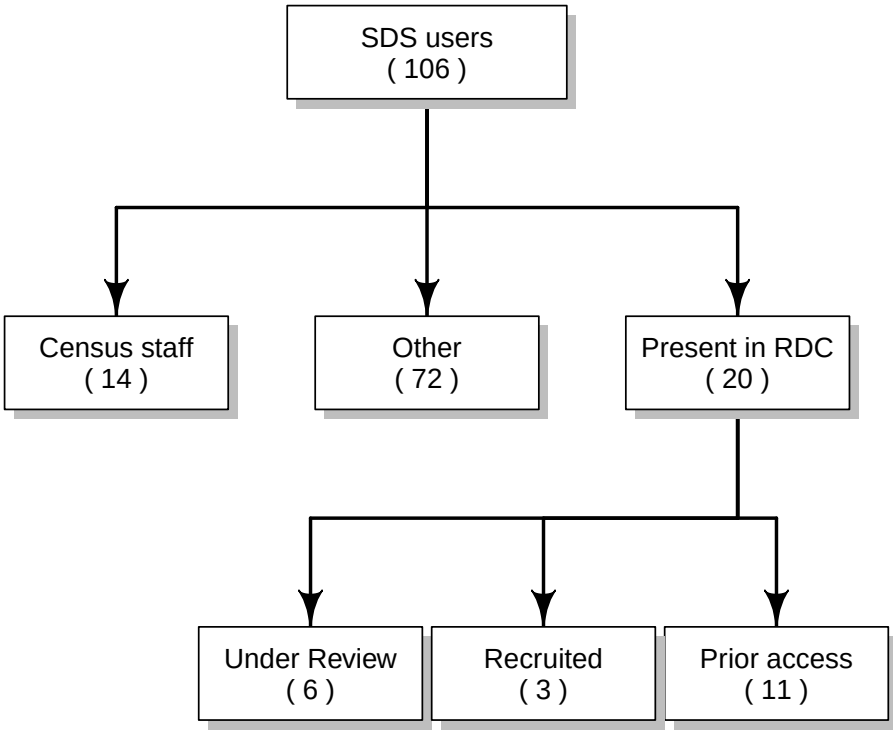
4.3 Prior and subsequent experience in the Census RDC

On 2014-10-14, we researched what kind of exposure users of the SDS had had to confidential data at the Census Bureau, by investigating whether they had prior or subsequent Census RDC projects. Firgure 4 summarizes the results.

One of the (initially) unintended consequences was the use of the SDS as a gateway to more in-depth use of confidential data, in particular data available through the FSRDC. Conversely, once made aware of the availability of data similar in content to the data in the FSRDC, researchers may wish to use the synthetic data, if it is sufficient for their research purposes. To assess the extent of such connections with the FSRDC, we extracted a list of current and past users of the SDS as of July 2014, and asked the Census Bureau to provide information on whether the users were known to the FSRDC account management system. The response obtained in October 2014, is summarized in Figure 4.[8] Of the 106 users we identified in this way, 14 were Census internal users, i.e., users who were actively engaged in ongoing Census projects, presumably related to the validation exercises themselves. 20 other users were also present in some form in the Census RDC system. 11 users had been authorized for at least one RDC project prior to their access to the SDS. More of interest, however, are the 3 users who obtained Census RDC access after their initial access to the SDS, and the additional 6 who were still waiting for the approval of their RDC project, on average 516 days after having started their SDS project. We don't have firm evidence of the relationship between the RDC projects and the SDS project, but from personal conversations of the authors with presenters at RDC conferences, at least some of the RDC projects were direct continuation of SDS projects, in domains that were not covered by the synthetic data, but the analysis for which was prepared for on the SDS. The average delay between project start on SDS and project start on RDC projects for those projects that were authorized was 400 days.

[8] Private correspondence with Barbara Downs, Lead Research Data Center Administrator, October 14, 2014

Figure 4: Connection between Census RDC usage and Synthetic Data Server



5 Other activities

5.1 Online codebooks

Using software developed under NSF Grant SES-1131848 (Abowd et al., 2011), online codebooks (Reeder et al., 2014, 2015; Vilhuber, 2013) were developed, enhancing the available documentation, and enabling users to better explore the feasibility of their projects with the synthetic and the confidential data. The online documentation can be found at `http://www2.ncrn.cornell.edu/ced2ar-web`. The availability of complete and transparent documentation, outlining the provenance, is an important factor in establishing confidence in the methods used to generate the synthetic data, as well as tracing the provenance appropriately in results. Ideally, this applies to documentation of (metadata for) both the confidential and the synthetic datasets.

5.2 Future activities and expected results

Based on the results from the past few years on the SDS, we are enhancing and expanding the data available through the SDS. In particular, we are exploring the following enhancements. First, we intend to make available on the SDS new datasets that are (a) not redistributable (b) but are under our full control for the validation process. This will allow us to implement a tighter (faster) coupling between the synthetic data generating process and the model validation. Sample datasets include a custom longitudinally-linked extract of the U.S. Office of Personnel Management (OPM) data and Brazilian microdata.

Second, we will develop new synthetic data generating processes, based on provably private algorithms (Dwork and Roth, 2014). The basic idea is to generate synthetic data with both better analytical validity (adapted to the models actually estimated) and better (provable) privacy (Abowd and Schmutte, 2015).

The results from these new research directions will provide realistic guidance and applicable toolkits to data providers of a variety of domains, as well as in the short term providing researchers the ability to access new datasets using the proven mechanism of the SDS.

6 Conclusion

The Synthetic Data Server (SDS) has been used by a large number of users. Outcomes of this experiment in the use of analytically valid datasets with validation of results are varied. A few users have published papers that directly leverage the validation setup. Others have leveraged the ability to do meaningful data exploration on the synthetic data, while waiting for more far-reaching projects to be approved for access to the underlying confidential data. Finally, the server has been used as a valuable teaching and data exploration tool for young researchers, lowering but not eliminating the cost of access to confidential data.

References

Abowd, J. M. (2010). CDI-Type II: Collaborative research: Integrating statistical and computational approaches to privacy. Grant 0941226, National Science Foundation. $409,296.00.

Abowd, J. M., Haltiwanger, J., and Jarmin, R. (2004a). Eitm: Developing the tools to understand human performance: An empirical infrastructure to foster research collaboration. Grant 0339191, National Science Foundation. $ 337,455.00.

Abowd, J. M. and Schmutte, I. (2015). Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, Fall 2015.

Abowd, J. M., Shapiro, M., Raghunathan, T., Jarmin, R., and Roehrig, S. (2004b). ITR-(ECS+ASE)-(dmc+int): Info tech challenges for secure access to confidential social science data. Grant 0427889, National Science Foundation.

Abowd, J. M., Vilhuber, L., Li, P., and Block, W. (2011). NCRN-MN: Cornell Census-NSF Research Node: Integrated research support, training and data documentation. Grant 1131848, National Science Foundation.

Benedetto, G., Stinson, M., and Abowd, J. M. (2013). The creation and use of the sipp synthetic beta. Technical report, US Census Bureau.

Bertrand, M., Kamenica, E., and Pan, J. (2015). Gender identity and relative income within households. *The Quarterly Journal of Economics*, 130(2).

Dwork, C. and Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*. now publishers, Inc. Also published as "Foundations and Trends in Theoretical Computer Science" Vol. 9, Nos. 3–4 (2014) 211-407.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):1–9.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011a). Appendix to 'Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database'. online document, Center for Economic Studies, U.S. Census Bureau.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011b). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, 79(3):362–384.

Miranda, J. and Jarmin, R. (2002). The Longitudinal Business Database. Discussion Paper CES-WP-02-17, U.S. Census Bureau, Center for Economic Studies.

Reeder, L. B., Stinson, M., Trageser, K. E., and Vilhuber, L. (2014). Codebook for the SIPP Synthetic Beta v5.1 [codebook file]. DDI-C document, Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, USA.

Reeder, L. B., Stinson, M., Trageser, K. E., and Vilhuber, L. (2015). Codebook for the SIPP Synthetic Beta v6.0 [codebook file]. DDI-C document, Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, USA.

U.S. Census Bureau (2011a). SIPP Synthetic Beta version 5.0. [computer file], U.S. Census Bureau [producer] and Cornell University, Synthetic Data Server [distributor], Washington,DC and Ithaca, NY, USA.

U.S. Census Bureau (2011b). Synthetic LBD Beta version 2.0. [computer file], U.S. Census Bureau and Cornell University, Synthetic Data Server [distributor], Washington,DC and Ithaca, NY, USA.

U.S. Census Bureau (2013). SIPP Synthetic Beta version 5.1. [computer file], U.S. Census Bureau [producer] and Cornell University, Synthetic Data Server [distributor], Washington,DC and Ithaca, NY, USA.

U.S. Census Bureau (2015). SIPP Synthetic Beta version 6.0. [computer file], U.S. Census Bureau [producer] and Cornell University, Synthetic Data Server [distributor], Washington,DC and Ithaca, NY, USA.

Vilhuber, L. (2013). Codebook for the synthetic lbd version 2.0 [codebook file]. Ddi-c document, Comprehensive Extensible Data Documentation and Access Repository (CED2AR), Cornell Institute for Social and Economic Research and Labor Dynamics Institute [distributor]. Cornell University, Ithaca, NY, USA.

Vilhuber, L. and Abowd, J. M. (2010). Synthetic data user testing and dissemination. Grant 1042181, National Science Foundation. $252,465.00.

$Id: report_on_SDS_2015_SOLE.tex 2647 2015-10-30 15:16:01Z lv39 $