International Statistical Review (2011), 79, 3, 362–384 doi:10.1111/j.1751-5823.2011.00153.x

Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database

Satkartar K. Kinney[1], Jerome P. Reiter[2], Arnold P. Reznek[3], Javier Miranda[3], Ron S. Jarmin[3] and John M. Abowd[4]

[1]*National Institute of Statistical Sciences, Research Triangle Park, NC, USA*
[2]*Duke University, Durham, NC, USA*
[3]*U.S. Census Bureau, Washington D.C., USA*
[4]*Cornell University, Ithaca, NY, USA*
E-mail: saki@niss.org

Summary

In most countries, national statistical agencies do not release establishment-level business microdata, because doing so represents too large a risk to establishments' confidentiality. One approach with the potential for overcoming these risks is to release synthetic data; that is, the released establishment data are simulated from statistical models designed to mimic the distributions of the underlying real microdata. In this article, we describe an application of this strategy to create a public use file for the Longitudinal Business Database, an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976. The U.S. Bureau of the Census and the Internal Revenue Service recently approved the release of these synthetic microdata for public use, making the synthetic Longitudinal Business Database the first-ever business microdata set publicly released in the United States. We describe how we created the synthetic data, evaluated analytical validity, and assessed disclosure risk.

Key words: Economic census; data confidentiality; synthetic data; disclosure limitation.

1 Introduction

Many national statistical agencies collect data on business establishments; however, very few disseminate establishment-level data as unrestricted public use files. Instead, they disseminate business data in highly aggregated forms, such as the County Business Patterns and the Business Dynamics Statistics released by the U. S. Census Bureau or the Business Employment Dynamics released by the Bureau of Labor Statistics.[1] The dearth of establishment-level data stems from requirements, both legal and practical, to protect the confidentiality of establishments' data. Business data are especially at risk for confidentiality breaches because (1) many establishments are highly recognizable and can be easily identified from just a few variables, e.g., the aircraft manufacturer in Seattle, Washington, or the largest restaurant in your home town, and (2) knowledge of a business's operations data, for example the total payroll for a small hardware store, could help its competitors gain an edge, so that there is incentive for malicious users of the data to attempt identifications.

Protecting confidentiality is also a concern with other types of data, such as demographic and health data; yet, there are many individual-level public-use files for those types of data. Typically, the data on these files have been altered before release using statistical disclosure limitation methods. Common methods include aggregating or coarsening data, such as releasing ages in five year intervals or geographies at low resolution; reporting exact values only above or below certain thresholds, for example reporting all incomes above 100 000 as "100 000 or more"; swapping data values for selected records, e.g., switch key variables for at-risk records with those for other records to discourage users from matching, since matches may be based on incorrect data; and, adding noise to numerical data values to reduce the possibilities of exact matching on key variables or to distort the values of sensitive variables. See Cox & Zayatz (1995), Fienberg (1994), FCSM (2005), Kinney *et al.* (2009), Skinner *et al.* (1994) and Willenborg & de Waal (2001) for overviews of these and other methods.

These disclosure limitation methods can be effective when used for small amounts of data alteration; however, they are ineffective when applied at high intensity, as is typically necessary for public-use establishment-level data. In the United States, for example, federal law prevents the release of exact values of tax data, even including the fact of filing. Hence, top-coding cannot be used on monetary data as a large fraction of exact values would be released. It also suggests that swapping would have to be done at a 100% rate, in which case the released data would be useless for any analysis involving relationships with swapped variables. Many variables of interest to researchers and policy-makers, for example number of employees and total payroll, have highly skewed distributions even within industry classifications. Hence, the amount of added noise necessary to disguise these observations could be so large as to degrade estimates of marginal distributions and attenuate estimates of relationships.

In this article, we describe the generation of an unrestricted public use establishment-level data set for the U.S. Census Bureau's Longitudinal Business Database (LBD). The LBD is essentially a census of business establishments in the U.S. with paid employees comprised of survey and administrative records. It supports an active research agenda on business entry and exit, gross employment flows, employment volatility, industrial organization, and other topics that cannot be adequately addressed without establishment-level data. Gaining research access to the confidential data in the LBD is not a trivial process, and thus a public-use version of the LBD analogous to those available for various demographic and health surveys will be quite useful to the research and policy-making communities. An initial version of the public release file was approved for release and can now be accessed via the SynLBD website.[2]

In addition to containing establishment data, the longitudinal nature of the data presents additional disclosure risk as the longitudinal structures themselves can aid in re-identification of establishments; in fact, public-use longitudinal data are nearly as rare as public-use establishment data. Additionally, as the LBD is a census, no protection is provided by sampling uncertainty. For these reasons any public release of the LBD requires a great deal of alteration in order to protect against disclosures. Further, no actual values in the LBD were permitted to be released to the public, with the possible exception of geographic and industry classification variables that are already publicly available from the County Business Patterns database. Hence, synthetic data (Little, 1993; Reiter, 2003; Rubin, 1993) presented the only viable method of generating a safe public-use LBD. Other disclosure limitation methods stand very little chance—we would venture to say no chance—of safely providing record-level data that preserve distributional features and associations across variables.

The basic procedure for generating synthetic data is to fit models for the sensitive information in the confidential data, simulate replacement values from these models, and release the simulated data for public use. This can protect confidentiality, since identification of businesses' sensitive data is difficult when the released data are not actual, collected values. Furthermore,

with appropriate data generation methods, the approach enables data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods, and software. Potential drawbacks include the complexity of the process needed to generate synthetic data and dependence of secondary analyses on the synthesis models, which could be mis-specified. For examples of reviews of synthetic data approaches, see Abowd & Vilhuber (2008), Abowd *et al.* (2009), Dreschler *et al.* (2008), Reiter (2003, 2004), and Reiter & Raghunathan (2007).

The remainder of the article is organized as follows. Section 2 describes the LBD in more detail. Section 3 describes the modeling strategies for creating a synthetic version of the LBD. Section 4 describes some checks of the usefulness of the synthetic data file via analyses of establishment characteristics, job dynamics, and correlational analyses. Section 5 describes the confidentiality properties of the synthetic LBD. Section 6 concludes with a discussion of plans for improvements to future versions of the synthetic LBD.

2 The Longitudinal Business Database

The U.S. Census Bureau Center for Economic Studies began developing the LBD in the late 1990s with the intent of extending the Longitudinal Research Database beyond the manufacturing sector (Foster *et al.*, 2006; McGuckin & Pascoe, 1988). The LBD trades depth of information on each establishment in exchange for complete industry coverage. As of this writing, it includes over 24 million establishments with employees that were active between 1976 and 2009. It is updated annually and continuously improved. Sources of information include the Business Register, Internal Revenue Service tax records, the Economic Census, the Company Organization Survey, and the Annual Survey of Manufactures.

By itself or in combination with other data sets the LBD is a powerful and unique research tool. The LBD is at the forefront of research on business dynamics, market structure, company organization, and business cycle. For example, Jarmin *et al.* (2005) use the LBD to examine producer dynamics in the U.S. retail sector and the growth of retail chains. Davis *et al.* (2006, 2008a) use the LBD to examine volatility and dispersion in business growth rates and their relation to business dynamics and employment. Davis *et al.* (2008b) use it to examine the impact of private equity on employment of target firms. Davis *et al.* (2007) use it to study entrepreneural activity and business formation.

The LBD also provides the source data for public use aggregate statistics including the Business Dynamics Statistics (BDS) and the U.S. series of the OECD's Entrepreneurship Indicators Programme. These statistics are currently used to examine the dynamics of young and small firms in the U.S. as well as for international comparisons of the creative destruction process of Haltiwanger *et al.* (2009c, a, b) and Bartelsman *et al.* (2004). For more information on the LBD see Jarmin & Miranda (2002).

Access to the LBD is regulated by Title 13 and Title 26 of the U.S. Code, meaning researchers desiring access must follow lengthy and potentially costly procedures to use the data. Currently researchers can only access it at one of several Census Bureau Research Data Centers (RDCs). Although the Census Bureau has dramatically streamlined and improved the process under which access to its confidential data is granted to external data analysts, there remain real costs to such an access model. Gaining restricted access requires a rigorous vetting process, and, for many researchers, requires travel to conduct research away from their home institutions. These costs undoubtably mean that many researchers who might benefit from using the LBD must find other ways to accomplish their research objectives or abandon them.

Despite these constraints the LBD has become the most popular data set in the RDC. Unique features set it apart from similar databases including: (1) the length of the time series—currently

Table 1
Synthetic LBD variable descriptions.

Name	Type	Description	Notation	Action
ID	Identifier	Unique Random Number for Establishment		Created
County	Categorical	Geographic Location	x_1	Not released
SIC	Categorical	Industry Code	x_2	Unmodified
Firstyear	Categorical	First Year Establishment is Observed	y_1	Synthesized
Lastyear	Categorical	Last Year Establishment is Observed	y_2	Synthesized
Year	Categorical	Year dating of annual variables		Created
Multiunit	Categorical	Multiunit Status	y_3	Synthesized
Employment	Continuous	March 12 Employment (annual)	y_4	Synthesized
Payroll	Continuous	Payroll (annual)	y_5	Synthesized

covering multiple business cycles; (2) the ability to link establishments and subsidiaries to their parent companies; (3) its wide coverage including practically all the private non-farm activity in the U.S., and (4) the ability to link to hundreds of other databases from the U.S. Census Bureau including economic census and surveys, as well as external databases such as the COMPUSTAT and VentureExpert.

A public version of the LBD will enable secondary data analysts to answer many questions without having to incur the costs of gaining access through an RDC. For questions that cannot be adequately answered by the synthetic data, the public use version would enable users to develop models, for example to create appropriate software code or investigate the need for transformations in regressions, before going to the RDC. This can make their limited time at the RDC more productive.

3 Generating the Synthetic LBD

The synthetic LBD, or SynLBD, is generated using the variables described in Table 1, which shows a more simplified structure than available in the full LBD. Additional variables, such as firm structure are available in the confidential LBD but were not used to generate the SynLBD; these will be incorporated in future versions. The LBD is a universe file (with the exception of any coverage omissions), and as such, there are no sampling weights. SynLBD is based on a cleaned version of the confidential database made available to authorized users in May 2007. Data cleaning steps include (1) removing establishments that are out of scope or that have obvious SIC coding errors, (2) editing establishments with obvious errors in payroll or employment, and (3) filling in the very small fraction of missing data with interpolated values.

For this initial version of the SynLBD, we generated a universe file comprising one record for each of 21 million establishments active in the Business Register any time between 1976 and 2001. Data for the years 2002 through 2005 were available at the time of synthesis but were excluded due to the change in industrial coding schemes from SIC to NAICS in 2001. This will be remedied in future versions of the SynLBD once the change is accounted for. The SynLBD contains each establishment's actual 3-digit SIC code, and synthesized values of first year, last year, annual payroll, annual employment, and multiunit status (recoded so as not to be annual). No geographic or firm-level information are included in this initial version, though County and State were used in the synthesis.

In Table 1, variables denoted with y_i are synthesized, i.e., their values are replaced with imputed values, and variables denoted with x_i are not synthesized. We discuss the implications for confidentiality protection of the structure of the SynLBD in Section 5. We use the terms "synthesize", "simulate", and "generate" instead of "impute" when describing the process of creating the SynLBD. While synthetic data is generated using multiple imputation, "impute" is

typically inferred as the replacement of missing values, whereas in synthetic data, imputation models are assumed to be fit on completely observed data.

The y_i are synthesized using a Dirichlet-multinomial model or a linear regression approach; see the Appendix (available on the SynLBD website) for details. We generated the synthetic data for each of nearly 500 3-digit SIC subgroups as follows:

1. Synthesize Firstyear using the Dirichlet-multinomial approach with "confidentiality prior" to draw from $f(y_1|x_1, x_2)$.
2. Synthesize Lastyear using the Dirichlet-multinomial approach with flat prior to approximate a draw from $f(y_2|y_1, x_1, x_2)$.
3. Synthesize a categorical Multiunit status using the Dirichlet-multinomial approach with flat prior to approximate a draw from $f(y_3|y_2, y_1, x_1, x_2)$.
4. Synthesize annual Employment and Payroll using normal linear regression approach with a kernel density estimator transformation applied to the response (Abowd & Woodcock, 2004), so that we approximate a draw from

$$f\left(y_4^{(t)}|y_4^{(t-1)}, y_3, y_2, y_1, x_1, x_2\right) \quad \text{and } f\left(y_5^{(t)}|y_4^{(t)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2\right),$$

where t indicates a year between 1975 and 2001. We use only one lag in the synthesis models for computational convenience, as using additional lags reduces the number of establishments available for modeling. Exploratory data analysis suggests that higher lags are not as important for prediction given the other terms in the models. Further, a large percentage of establishments have lifetimes of only one or two years.

3.1 Firstyear

Firstyear is a left-censored ordinal categorical variable representing the first year an establishment reports positive payroll. Its possible values are the years 1975 through 2001. The variable is synthesized conditional on 4-digit SIC and County using the Dirichlet-multinomial approach. We use a prior distribution with a small sample size and probabilities determined by the percentages of years in the 3-digit classification in that County. The reason for using an informative prior is to introduce uncertainty by introducing a positive probability that the synthetic Firstyear for a unique or homogenous SIC-County subgroup can take on a value not on the confidential data. That is, it is used to add noise in a controlled fashion rather than improve prediction, hence the term "confidentiality prior". Normalizing to a small sample size ensures that the prior does not impact larger and less homogenous SIC-County subgroups. Sampling from the posterior predictive distribution yields a synthetic Firstyear for each establishment that is conditional on 4-digit SIC and County.

3.2 Lastyear

Lastyear is a right-censored ordinal categorical variable representing the last year an establishment reports positive payroll. Its possible values are the years 1976 through 2005. The variable is synthesized using the Dirichlet-multinomial approach with an improper flat prior distribution, conditional on Firstyear, and 4-digit SIC. Sampling from the posterior predictive distribution yields a value for Lastyear for each establishment in the Firstyear-SIC pair. Dependencies on geographic variables are not modeled in this version of the SynLBD.

3.3 Multiunit

The variable Multiunit indicates whether or not an establishment was ever part of a multi-unit firm, i.e., whether an establishment was ever part of a parent enterprise-conducting business at multiple locations. In the confidential data, multi-unit status is a series of longitudinal binary indicators. To facilitate synthesis, we defined a categorical variable equal to 1 if the establishment was never part of a multi-unit firm; equal to 2 if it switched from single to multi-unit once; equal to 3 if it switched from multi-unit to single once; equal to 4 if it switched multiple times; and equal to 5 if the establishment was always part of a multi-unit firm. For establishments with values of 2–4, the year change occurs is of interest to researchers so this is planned for future versions of the SynLBD.

We synthesized this categorical variable using the Dirichlet-multinomial approach with a flat prior. We sampled the posterior predictive distribution in the same manner as for Firstyear and Lastyear. For each establishment, the posterior multinomial distribution was sampled to yield a value of 1–5 for that establishment, conditional on Firstyear, Life (Lastyear − Firstyear), 4-digit SIC, and State.

3.4 Payroll and Employment

Payroll and employment data are imputed for each active establishment in every year between 1976 and 2001. If the synthetic values of Firstyear and Lastyear indicate an establishment was inactive in a given year, payroll and employment values are not generated. The employment variables are imputed first, in ascending order by year, followed by the payroll variables. Separate regressions are estimated for different subgroups of data as described below. Within any subgroup and any year, our general strategy is to specify a saturated model and use the Bayesian Information Criterion (BIC) to select a parsimonious subset of variables from this model. In many 3-digit SIC groups, the payroll and employment values are highly skewed, so we synthesize using regressions with kernel density estimator transformations of the response variables (see Woodcock & Benedetto (2009) and Abowd & Woodcock (2004)) and logarithmic transformations of continuous predictors.

3.4.1 Subgrouping procedures

For each year, establishments within any 3-digit SIC in operation that year are broken down into groups for separate modeling so that automated procedures can be used. First, the establishments are dichotomized by multiunit status; establishments that change their multiunit status over time are treated as multiunit establishments. The synthesis is done separately for single unit and multiunit groups, since in many industries there are sharp differences between single unit and multiunit establishments for these variables. Second, within each status group, establishments are separated into births and continuers; see below for more details on the models for births and continuers. We use this separation because (1) the distributions for births and continuers can be quite different, and (2) the models for continuers condition on previous year values that are not available for births. Within these subgroups, the establishments are split into the highest 5% and the lowest 95% of employment level when the sample size in the observed data is sufficient for the saturated model to be full rank. This special treatment of large values generally improves the quality of the synthesis. Models are estimated separately in each of these subgroups.

Some subgroups are sparse enough that it is not possible to estimate the saturated model. Thus, we use an informative prior distribution for the regression coefficients in the Normal method. Specifically, for a given 3-digit SIC group, we find a comparable subgroup in the corresponding

2-digit SIC group. The data from this 2-digit group are used to compute the prior distribution. We use a unit information prior, which has the same amount of information about the regression coefficients as that contained in a single observation. See the Appendix for descriptions of these prior distributions.

3.4.2 Births

First-year employment and payroll are synthesized from observed data corresponding to units in their first year. The predictors in the saturated model for employment births include 4-digit SIC, years till death, and indicators for whether or not the Firstyear or Lastyear are censored, and an indicator for the penultimate year. The predictors in the saturated model for payroll are the same, with the addition of the log of current year's employment.

Large percentages of establishments start their first year after March 12 and hence have zero employment recorded for the first year. Because of this, the birth model is fit in two stages. First, a logistic regression is used to predict whether or not units in their first year had zero March 12 employment, conditional on 4-digit SIC and years to death.[3] Second, the observed data for units with nonzero employment are used to synthesize employment for units predicted to have nonzero employment using a linear regression with a kernel density estimator transformation of the response variable.

3.4.3 Continuers

The predictors in the saturated employment model for continuers include 4-digit SIC, age, years to death, log of previous year's employment, indicator variables indicating that the first year is censored or that the last year is censored, indicators for the second year, the penultimate year, and the last year, and interactions of these three indicators with the log of previous year's employment. Second-year establishments that had zero employment in the first year are separated for the imputation of employment.

The predictors in the saturated model for payroll include all 4-digit SIC, age, years to death, log of current year's employment, log of previous year's payroll, indicator variables indicating the first year is censored, the last year is censored, if the current year is the second year, the penultimate year, or the last year, and interactions of payroll and employment with the last year and penultimate year indicators.

4 Analytic Validity of the Synthetic Data

In this section we explore the analytical validity of the initial SynLBD file for reproducing key analyses obtained from the LBD. Additional summaries are provided on the SynLBD website. A full evaluation of all the possible analytical applications is beyond the scope of any one paper and might only come in time. Generally, it is expected that high-level analyses involving large groups will be well preserved, while analyses involving small groups or high-dimensional inferences will require access to the confidential data.

4.1 Establishments Characteristics

The SynLBD generally provides inferences on aggregate means and correlations that are similar to what would be obtained from the LBD. For example, Figure 1 shows gross employment levels for each year in the SynLBD are very close to those in the LBD. The average discrepancy over the 1976–2000 period is 1.3%. A similar comparison of gross payroll, not shown, indicates

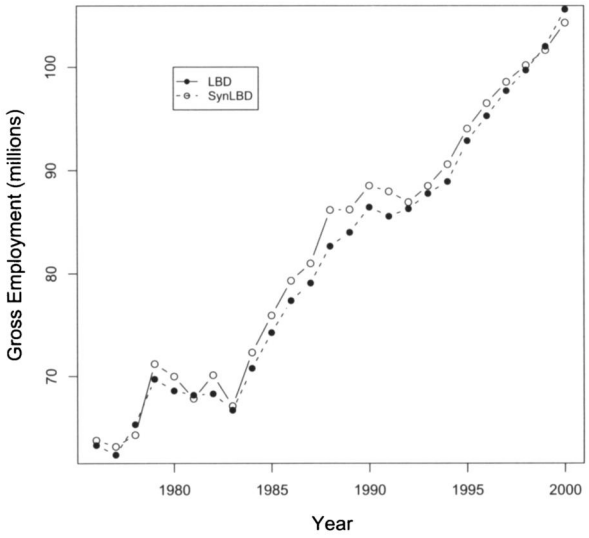

Figure 1. *Gross employment level by year, LBD versus Synthetic.*

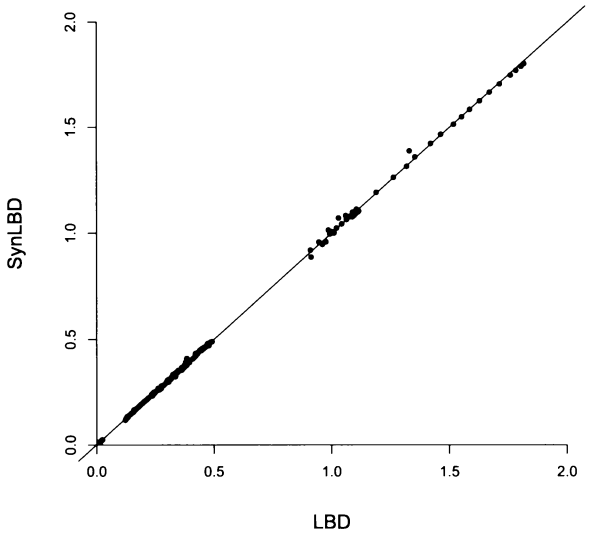

Figure 2. *Share of establishments by industry sector and year, 1976–2000.*

a slightly higher average discrepancy of 8%. Other marginal and conditional distributions that are modelled (as described in Section 3), such as first year and last year, are well preserved, as are the distributions of numbers of establishments per year and their lifetime (defined as Lastyear - Firstyear). See the SynLBD website for additional details.

The SynLBD preserves the industry and geographic distribution of establishments in the United States over time. Although neither geography nor industry are synthesized, since entries and exits are synthesized, the number of establishments in a given industry or state in a particular year may differ from the LBD. Figure 2 plots the share of establishments by sector and year in the LBD and the SynLBD. The shares cluster along the 45-degree line, illustrating that industry sectors (2-digit SIC) have similar shares of total establishments in the LBD and SynLBD. Put

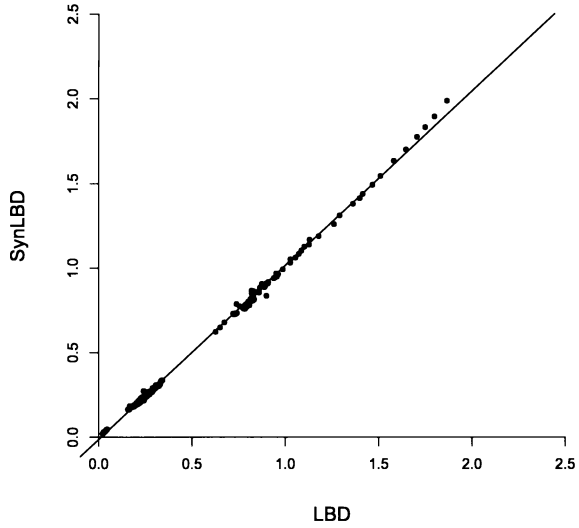

Figure 3. *Share of employment by industry sector and year, 1976–2000.*

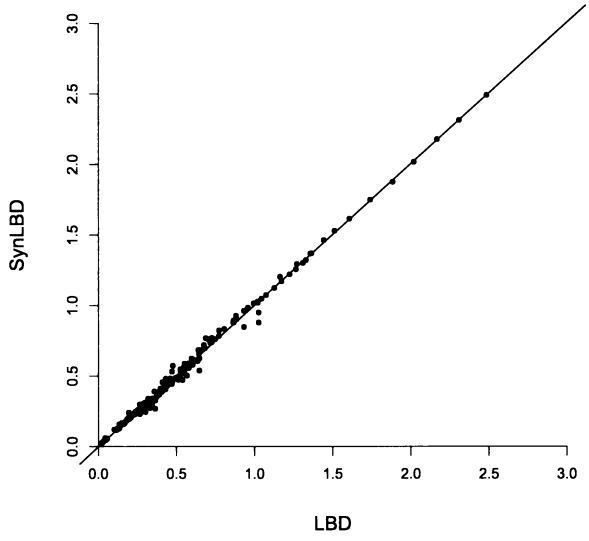

Figure 4. *Share of payroll by industry sector and year, 1976–2000.*

differently, sectors that dominate activity in the LBD also do so in the SynLBD. The SynLBD is also designed to preserve the employment and payroll weighted industry distribution of activity in the economy. Some sectors are becoming increasingly important in the economy while others are in decline; some are more labor intensive than other and yet other require higher skills and pay higher wages. The SynLBD is designed to preserve these relations in the data. As before Figures 3 and 4 show those shares fall tightly along the 45-degree line.

Figures 5 through 7 illustrate the preservation of state-level geography, though geography is not included in the current release. Figure 5 shows that the distribution of establishments across states is well preserved. Figures 6 and 7 show that the payroll and employment weighted distributions of establishments across states are reasonably well-preserved, although there are

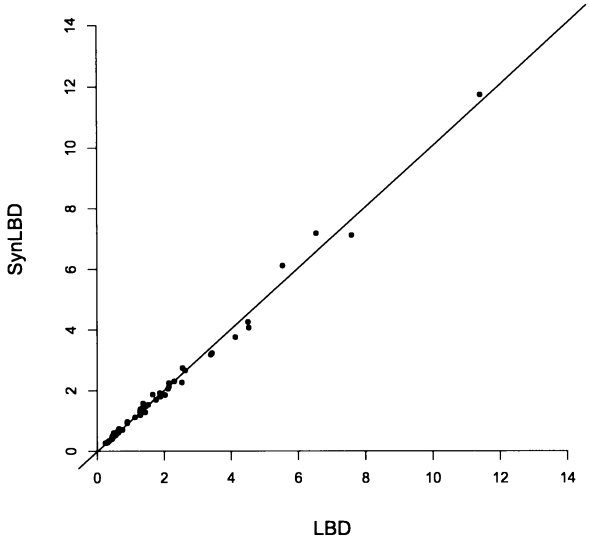

Figure 5. *Share of establishments by state and year, 1976–2000.*

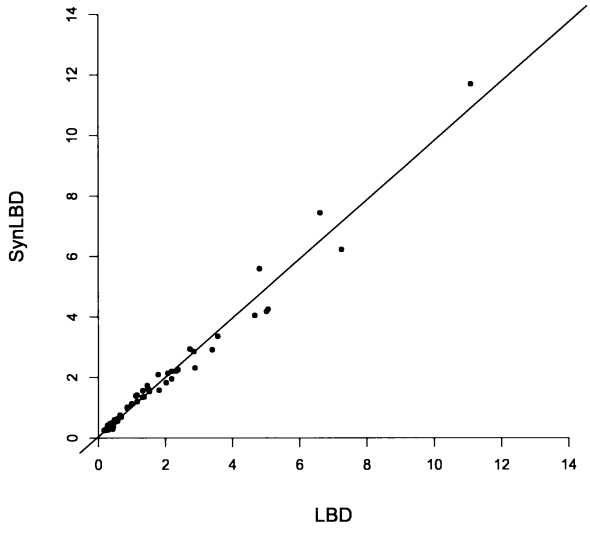

Figure 6. *Share of employment by state and year, 1976–2000.*

modest discrepancies for the larger states of New York, California, and Texas. The SynLBD employment and payroll models do not include State as a conditioning variable so it is not surprising that the distributions do not line up precisely.

4.2 Lifetime: Firstyear and Lastyear

We examine the validity of the synthesized lifetime of establishments by matching establishment entry and exit rates in the SynLBD against observed ones in the LBD for every year between 1976 and 2000. We find the entry and exit rates match up well. The average establishment entry rate in the SynLBD during this period is 11.13% versus 11.14% in the LBD. Employment size

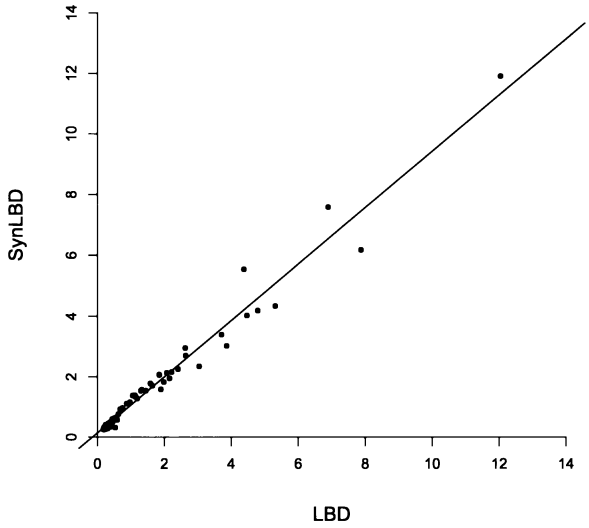

Figure 7. *Share of payroll by state and year, 1976–2000.*

for startups is the focus of intense research. In this regard we are also interested in preserving employment weighted entry and exit rates. We find the SynLBD again does well in this regard. The average employment weighted entry rate in the SynLBD during this period is 2.85% versus 3.16% in the LBD, an average discrepancy of 0.31 points.

4.3 Dynamics of Job Flows

One of the most important applications of the LBD is to generate statistics that describe the amount of job creation and destruction taking place in the economy, the number of establishments that are opening and closing and the amount of employment volatility in the economy (see Section 2). Here we reproduce measures of job flows that are common in the economics literature. These statistics are particularly good candidates for our testing purposes because (1) they make intensive use of the data by requiring computation of flow measures for each and every establishment in the file, and (2) they require reproduction of cross-sectional as well as longitudinal features of the data.

First, job creation and destruction are defined as in Davis *et al.* (1996):

$$JC_t = \sum_e \left(\frac{Z_{et}}{Z_t}\right) |\max\{0, g_{et}\}| = \sum_e \left|\max\left\{0, EMP_{et} - EMP_{e,t-1}\right\}\right| / Z_t$$

$$JD_t = \sum_e \left(\frac{Z_{et}}{Z_t}\right) |\min\{0, g_{et}\}| = \sum_e \left|\min\left\{0, EMP_{et} - EMP_{e,t-1}\right\}\right| / Z_t$$

$$NET_t = JC_t - JD_t$$

where JC_t is the sum of all employment gains from expanding establishments from year $t-1$ to year t including establishment startups, $Z_{et} = 0.5*(EMP_{et} + EMP_{e,t-1})$ is a measure of size of employer e, EMP denotes the number of employees, and $g_{et} = (EMP_{et} - EMP_{e,t-1})/Z_{et}$, is the growth rate from $t-1$ to t of employer e. Job creation is expressed as a rate by dividing through by total employment defined as the average of the total jobs in years $t-1$ and t, $Z_t = \sum Z_{et}$. Similarly, JD_t is the sum of all employment losses in year t including the sum of employment

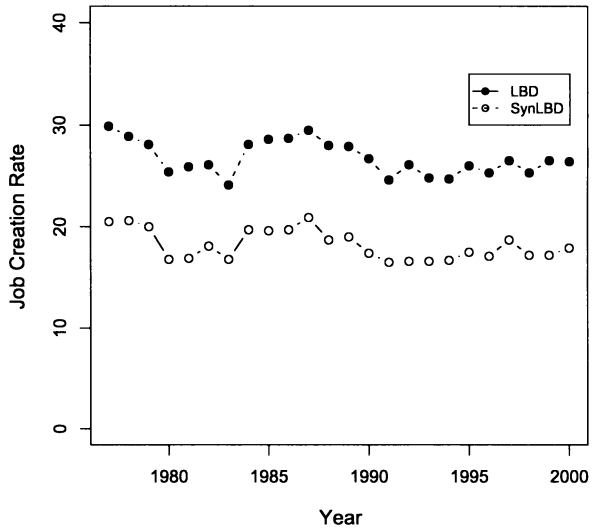

Figure 8. *Job creation rate by year, LBD versus Synthetic.*

over (1) all establishments that are last observed in year $t - 1$ and (2) employment losses for establishments that contracted between year $t - 1$ and year t. Job destruction is expressed as a rate by dividing the total employment defined as the average of the total jobs in years $t - 1$ and t. Net job creation is the job creation rate minus the job destruction rate.

Figure 8 shows the job creation rates from the SynLBD and compares them against the job creation rates from the LBD. Figure 8 shows that the SynLBD reproduces the year-to-year movements in the job creation rate rather well. However, the rate is considerably lower in the LBD, approximately 10 points lower when compared to the SynLBD. This is due to greater year-to-year changes in employment at the establishment level compared to the LBD. A large portion of establishments have constant employment from one year to the next which is not captured by the SynLBD. This is illustrated in Figure 9 , which shows a greater mass of establishments with zero growth rate in 1987 compared to the SynLBD. Year-to-year changes at the aggregate level are still preserved, as seen in Section 4.1.

A similar effect is seen in the distribution of job destruction rates, so that the net job creation rate, i.e., job creation minus job destruction, is captured quite well in the SynLBD, as shown in Figure 10. Despite the discrepancy in the series it is encouraging to see that the models track yearly movements in the job creation and destruction series for all years and through contractionary and expansionary periods. It is anticipated that this discrepancy can be reduced in future versions, and for the time being, SynLBD users are provided with information on the discrepancies in job creation and destruction rates, as well as other measures of variability of interest to economists, such as employment volatility (Davis *et al.*, 2006).

4.4 Regression Analysis

The LBD has been used to examine industry dynamics (the growth and decline of industries), the dynamics of new entrepreneurial firms (firm and establishment growth) as well as the spatial dimension of economic activity (growing and declining regions). In this section we assess how well the SynLBD captures variability in economic growth due to industry, establishment age,

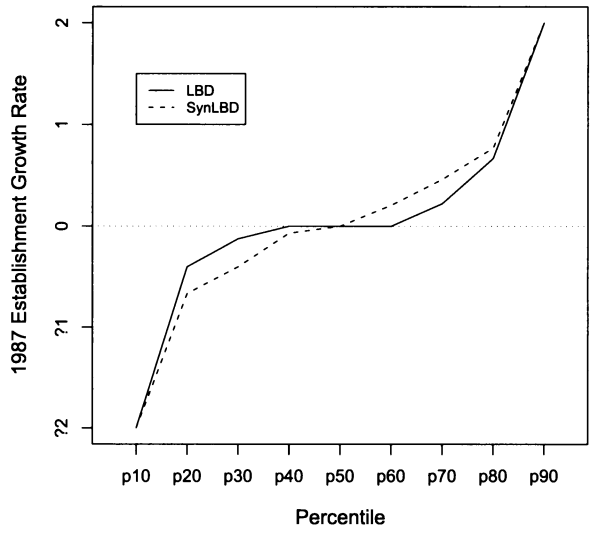

Figure 9. *Distribution of job creation rates, LBD versus Synthetic.*

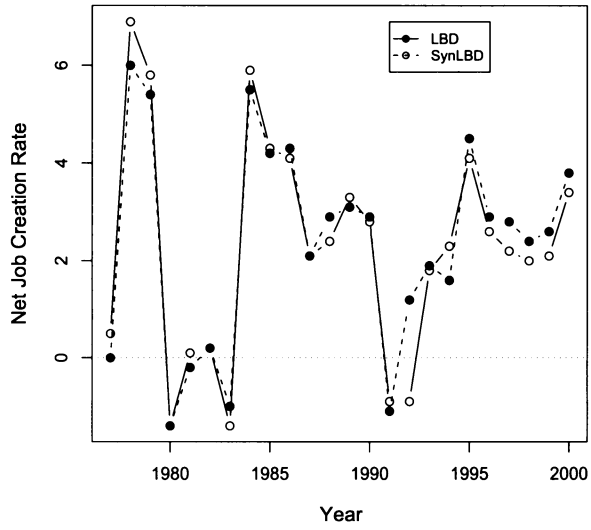

Figure 10. *Net job creation rate by year, LBD versus Synthetic.*

and geographic location, by estimating the following model:

$$LEMP_i = \alpha + \beta LEMP_{i-1} + \delta LPAY_i + \theta IND_i + \psi STATE_i + \vartheta AGE_i + \gamma MU_i + \epsilon$$

where $LEMP_i$ is the logarithm of employment of establishment $i = 1, \ldots, N$, $LEMP_{i-1}$ is the logarithm of prior year's employment, $LPAY_i$ is the logarithm of payroll of establishment $i = 1, \ldots, N$, IND_i is a vector of P 2-digit industry dummies, $STATE_i$ is a vector of dummy variables for State, AGE_i is a vector of dummy variables for age of establishment, MU_i is an indicator for multi establishment firm status, and $\epsilon \sim N(0, \sigma^2)$.

We run this model on one year of the data, 1997. This allows for a large sample on which to estimate our coefficients without loss of generality. We estimate this equation separately

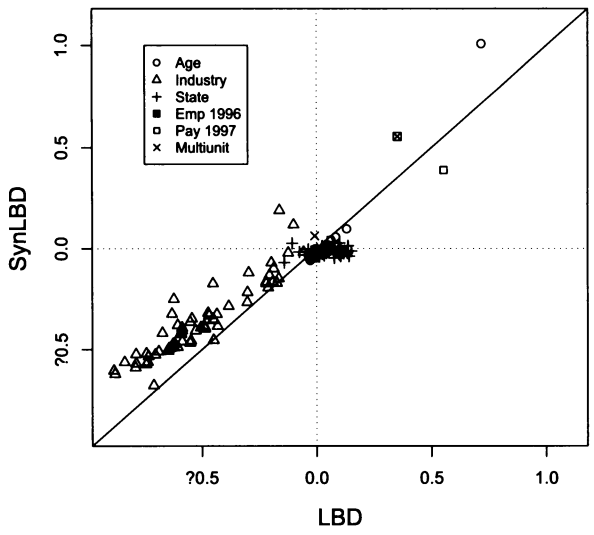

Figure 11. *Regression coefficients, LBD versus Synthetic.*

on the LBD and the SynLBD and find that the SynLBD provides similar inferences to the LBD, summarized graphically in Figure 11, although coefficients are slightly attenuated. The coefficients for State are the exception; however, this is unsurprising as State does not appear to have much impact on employment variability and was not included in the synthesis model for employment. A handful of other coefficients change sign in the SynLBD though they are close to zero. Qualitatively similar results are obtained for different years.

5 Confidentiality Properties

This section describes the features of the synthetic data that make it difficult for identity and attribute disclosures to occur. The discussion includes typical metrics for disclosure risk considered by statistical agencies and illustrative computation of the risk per the criterion of differential privacy (Dwork, 2006) from cryptography.

5.1 Structural Properties

There are many features of SynLBD that reduce disclosure risks. Broadly, the protection results from replacing actual values with simulated values drawn from probability distributions. Because the released data are essentially fully simulated, it is difficult to determine the actual values for any single source record.

In the SynLBD, arguably attribute disclosures are of greater concern than identification disclosures. The underlying universe data in the LBD are basically the same as those in the County Business Patterns (CBP) program, and the public use, tabular CBP files contain county and industry codes, which when crossed are unique for many establishments. Hence, many establishments can be identified from the CBP as being in the data that produced the LBD.

In other applications of synthetic data (e.g., the SIPP synthetic data program), disclosure risks are assessed with re-identification experiments. That is, record linkage software is used to attempt to match the synthesized data to the actual data. When there are many correct matches, the synthetic data are considered not adequately protective. Re-identification experiments are

of limited value in the SynLBD. First, with the exception of SIC, there are no real data on the SynLBD; hence, there is little to match with certainty. Second, because employment and payroll are not defined when the establishment is not in operation, and birth and death years are synthesized, it is difficult to conceive of sensible matching algorithms. One possibility is to match only when records have the same lifetimes, but how does one proceed when there are no such records? We considered several ad hoc rules in such cases, but we were not able to specify one that had sensible properties. Thus, we did not perform re-identification experiments.

As public establishment-level data are rare and synthetic data are still unfamiliar to many users, we took additional measures to avoid the perception of high disclosure risks. We released only one version of the SynLBD, as opposed to multiple versions as recommended in the literature on synthetic data. We suppressed all geographic information in the released data. Firm structure was not used in the synthesis at all and ownership links, real or synthetic, between establishments in multi-unit firms are not released. Future versions of the synthetic LBD are expected to include these features without compromising disclosure protection.

In the sections below we illustrate the unit-level differences between the LBD and SynLBD with several examples. Our comparisons were done using an unreleased version of the SynLBD containing actual ID variables enabling us to link units to the confidential LBD. Such linking would not be possible with the released SynLBD.

5.1.1 Firstyear and lastyear

Although general birth and death properties are preserved in the SynLBD, as illustrated in Section 4, at the unit level the synthesis introduces a substantial amount of variability between the real and actual values and longitudinal trends. As an illustration, Table 2 summarizes, for each 3-digit SIC, the probability that the synthetic first year equals the true first year, given the synthetic first year. That is, for a given year *yy*, the number of units with both synthetic firstyear and actual firstyear equal to *yy*, divided by the number of units with synthetic first year equal to *yy*. For virtually every industry these probabilities are quite low. The high probabilities for 1975 are due to censoring. Similar results are seen for the lastyear variable. There are a small number of exceptions in industries (at the 3-digit SIC level) that have few establishments to begin with; hence, there are few birth or death years to choose from. It is questionable whether or not these represent establishment-level disclosures. For example, if every establishment in a particular 3-digit SIC begins (or ends) in the same year, releasing that year does not distinguish individual establishments in that industry.

The synthesis of birth and death years confounds re-identification because establishments in the synthetic data exist at different times and for different lifetimes than in the real data. Thus for a given year different establishments are active in the actual data and synthetic data. By random chance, some establishments will have longitudinal characteristics that can be found in the original data; however, users cannot meaningfully attach establishment identifications to those records, since the lifetimes are random. An analogy to data swapping is appropriate here. In data swapping algorithms, commonly applied to other types of confidential data, the swap rate is normally an increasing function of some measure of the risk for a particular confidential record. Not all swap rates are unity. The protection (which is substantially weaker than the protection afforded by synthetic data) does not depend upon the complete absence of risky records in the output data set. Rather, the protection in swapping depends upon the uncertainty created by the probability of a swap. In synthetic data with a provably safe synthesizer (see Section 5.2 below), the protection comes from the fact that any confidential record can be transformed into any released record (including records that do not occur in the universe of actual records in the

Table 2
Summary statistics: Observed establishment births given synthetic births
Pr[Actual Firstyear = YYYY|Synthetic Firstyear = YYYY].

First (Birth) Year		% of Births Over Industries		
Synthetic	Actual	Minimum	Mean	Maximum
1975	1975	1.52	25.41	88.89
1976	1976	0.12	5.12	75.00
1977	1977	0.43	5.09	71.43
1978	1978	0.46	3.65	16.22
1979	1979	0.27	3.90	50.00
1980	1980	0.36	3.46	25.00
1981	1981	0.26	3.91	50.00
1982	1982	0.36	3.69	50.00
1983	1983	0.39	4.10	50.00
1984	1984	0.69	3.79	19.30
1985	1985	0.15	3.75	23.73
1986	1986	0.41	3.92	33.33
1987	1987	0.35	4.19	25.00
1988	1988	0.48	4.25	52.48
1989	1989	0.63	4.28	25.15
1990	1990	0.47	3.91	25.00
1991	1991	0.56	4.18	50.00
1992	1992	0.45	3.94	17.39
1993	1993	0.67	3.86	25.00
1994	1994	0.53	4.33	50.00
1995	1995	0.35	4.16	16.67
1996	1996	0.20	4.12	16.67
1997	1997	0.10	4.04	18.60
1998	1998	0.46	3.85	20.00
1999	1999	0.28	4.64	43.02
2000	2000	0.31	4.46	33.33
2001	2001	0.35	4.22	25.27

confidential data) with strictly positive probability. This property ensures that there are no exact disclosures and bounds the inferential disclosures.

5.1.2 Payroll and employment

The synthesis models for payroll and employment are, at their core, regressions with transformed variables that preserve low-dimensional relationships and sacrifice high dimensional ones, which can compromise confidentiality because they limit the ability of the synthesizer to transition points in the support of the confidential data to different points in the support of the synthetic data. This smoothing results in variability in the predictive models used for synthesis (e.g., large variances around the regression lines), which enhances the ability of the synthesizer to transition away from a given point in the support of the confidential data. Thus, any establishment's synthesized employment or payroll values can differ substantially from its original values.

We illustrate how original and synthetic values can differ with a simplified example. Suppose that we seek to synthesize the 1995 employment for all establishments in a particular 3-digit SIC category with synthetic first year of 1995. We fit a smooth approximation to the distribution of 1995 employment for all such establishments in the first year, adding other establishments from the 2-digit SIC group if necessary to increase sample size. We then randomly sample values from this approximation, which creates the synthetic data. If our approximation is reasonable over most of the support of the employment distribution, we should reproduce the main features of the data. But, an individual establishment's values can be drawn from anywhere in the distribution.

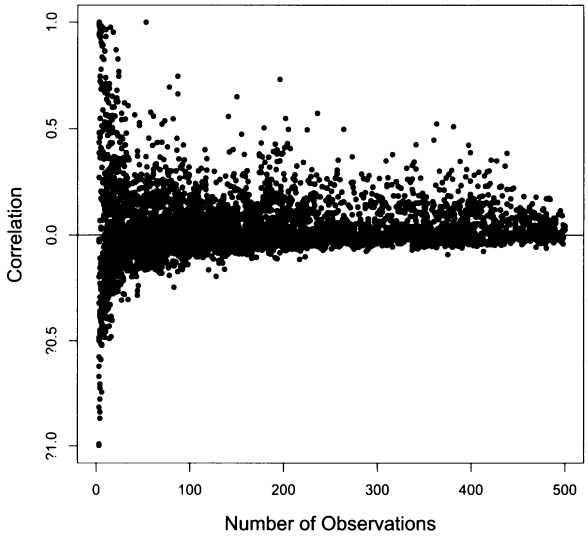

Figure 12. *Pearson correlations of payroll with synthetic payroll versus n:* $2 < n < 500$.

For example, there is nothing in the models that forces a large establishment in the real data to be a large establishment in the synthetic data (where "large" is measured within the SIC3). Hence, the values can and do move around substantially. Similar variability applies when generating employee or payroll in continuing years.

To summarize the substantial variability between the records in the confidential data and their images in the SynLBD, we computed the Pearson correlations, by year within industry, of actual versus synthetic employment, and actual versus synthetic payroll. Only establishments active in both the observed and synthetic data at the same time are used. The results for payroll are shown in Figure 12 . The horizontal axis shows the number of establishments in the (year, SIC) group for which the Pearson correlation was calculated, from 1 to 500. Correlations when $n > 500$ are near zero and are not shown. The vertical axis shows the value of the correlation between the confidential variable and the synthetic variable for all establishments in the (year, SIC) group that had data, from -1 to 1. The correlations are almost uniformly small; in fact, most are near zero. The few exceptions occur in industries that contain years in which very small numbers of establishments occur in both the observed and synthetic data. For example, the correlations equal either to 1 or -1 arise when there are only two observations. The results for employment are similar.

There may be establishments in the synthetic data that have payroll and employment values in any particular year that look like those for genuine establishments. In high density regions of the distribution, this represents minimal disclosure risk; for example, there are many restaurants in the U.S. with five employees, so releasing a synthetic employee size of five for some restaurant—without any information about geography and firm links—tells us nothing new. In the tails of the distribution, and particularly the upper tails, similarity of real and synthetic values could represent a risk in the following sense. If an intruder knew that a certain establishment—not firm, since firm structure is not preserved—was by far the largest employer in a particular SIC3 across the entire country, the intruder might speculate that the largest synthetic values in that SIC3 cell are reasonable estimates of the confidential value of employment.

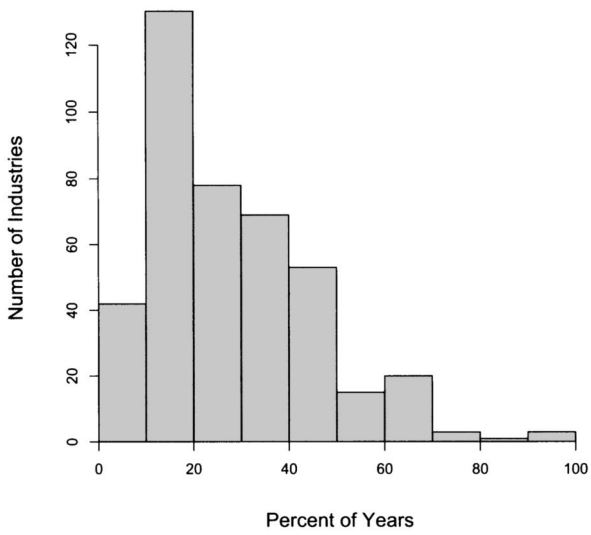

Figure 13. *Histogram: Percent distance between actual and synthetic employment.*

This risk is negligible for several reasons. First, because we simulate data from low dimensional models, for most SIC3 groups there should be large variance in the generated values of the tails of the distributions. For example, the largest value in a SIC3 by year group could be simulated to be 5 000, other times 2 000, and other times 1 000. The first order statistic's value depends on the random draws. Second, because these are longitudinal data, the simulation variance propagates across years and variables; that is, it cumulates as one continues to calculate the joint posterior predictive distribution from the sequence of its conditional distributions given the variables that have already been synthesized. This, along with variation in the lifetime distributions, makes it difficult to find synthetic establishments with very similar longitudinal payroll and employment histories as real establishments.

We investigated the risk associated with outliers by comparing the maximum synthetic payroll/employment to the maximum real payroll/employment within each industry in each year. Figure 13 displays a histogram of each industry's percentage of years that the synthetic and real maximum employment are within 5% of each other. The unit of analysis summarized in the histogram is a (year, SIC) group. The horizontal axis shows the percentage of years within the SIC where the absolute value of the difference between the maximum synthetic value and the maximum actual value is less than 5% of the actual maximum. The vertical axis shows the count of the number of SICs that are found in each range shown on the horizontal axis. For example, 42 SICs have synthetic employment data such that in less than 10% of the years does the synthetic maximum employment value come within 5% of the actual maximum value. Similarly, 130 SICs have synthetic employment data such that in 10% to 20% of the years the synthetic maximum employment is within 5% of the actual value. The graph for payroll is similar.

For both employment and payroll, the percentage of years in which the synthetic maximum value comes within 5% of the actual maximum value is less than 50% for the vast majority of industries (height of the bar). This indicates that the synthetic data generation process injects enough variation to cause the synthetic and real maxima to differ by more than 5% for the vast majority of industries and years. Furthermore, there is nothing in the synthetic LBD that indicates when the synthetic maxima are close to the real maxima and when they are not. Hence,

in any given year and any given 3-digit SIC, an intruder cannot be sure if using the synthetic maximum gives a close estimate of the actual maximum. This provides protection akin to the protection afforded in swapping, but stronger: the intruder who believes that he or she has a good estimate cannot be sure of being correct. And, because the data were synthesized using methods that do not simply reproduce the points in the original data, the intruder never gets real values (unlike swapping, where the swapped data are still actual values from the confidential data albeit assigned to different entities). The protection is also stronger than typical swapping applications, because the attacker will be completely off a large percentage of the time. In fact, the differential privacy limit (see Section 5.2 below) ensures that even the most successful attacker outcome—assuming the attacker knew which synthetic industry-years came closest to their confidential values—will generate an incorrect inference about the confidential value with a strictly positive probability.

5.2 Differential Privacy Computations

The SynLBD were generated using what the cryptography-based computer science privacy literature calls a "randomized sanitizer". In this case the randomized sanitizer is the data synthesizer described in Section 3. A few properties of this synthesizer can be used in conjunction with the theorems in the privacy-preserving data-mining literature to formally characterize the confidentiality protections in the SynLBD. First, given SIC, for all the synthetic variables the domain of the joint posterior predictive distribution is the same as the domain of the underlying confidential data. Second, given SIC, for every point in the domain of the joint distribution of the confidential data (*i.e.*, every point in the sample space, treating the confidential LBD as the universe), every point in the domain of the conditional distribution of the synthetic data, given the confidential data, has positive probability. This implies that there is no record in the confidential data that is reproduced in the synthetic data with probability one.

Under these two conditions, Dwork's (2006) differential privacy theorem can be applied to assert that there are no exact disclosures (attribute, identity, or inferential) relative to a world in which the attacker knows everything possible except the exact value of some variable for some entity. The same theorem implies that there is a finite bound on the maximum amount that any attacker with any information set (including all of the confidential data, save one observation) can learn about any entity in the universe (bound on the inferential disclosure). These results do not depend upon the values in the underlying confidential data. They are properties of the synthesizer—hence, properties of the confidentiality protection—regardless of the actual values that occur in the confidential data. The differential privacy result holds for every entity in the universe and every variable in the confidential data except those that are unsynthesized (SIC in this case).

In this section, we estimate the differential privacy limit for one industry in the synthetic LBD. In principle, it is possible to repeat this exercise for all industries and for all combinations of variables. In practice, computational routines do not exist for datasets with the complexity of the SynLBD.

For discrete data, the differential privacy bound is computed as the maximum of the logarithm of the ratio of elements of a transition matrix giving the probability of going from a row value, which comprises values of the confidential data, to a column value, which comprise values of the synthetic data. The numerator and denominator of the ratio come from the same column and the candidate rows are all those rows that can be reached by a single change in the confidential data (one value of one variable for one record). In the SynLBD, the complete transition matrix has billions of rows and columns; thus, it is too large to compute

completely. An additional complication is that computation of the conditional probabilities that form the elements of the transition matrix requires evaluation of the joint probability of all outcomes in the support of the posterior for each outcome in the actual confidential data. Even though each of these probabilities is computable from the statistical analysis that produced the synthesizer, it was not feasible at this time to do the exact calculation of the transition matrix probabilities.

Instead of estimating the complete transition matrix, we focused on the transition matrix from the actual employment and payroll values for the same year to the synthetic values for those variables in that year. We formed discrete categories for each variable based on intervals defined by the quantiles: [0,90) [90,99) [99,max). Thus, the lowest category consists of all establishments whose actual (*resp.*, synthetic) value for employment (*resp.*, payroll) was strictly below the 90th percentile of the employment (*resp.*, payroll) distribution in the confidential data for that year and SIC. This results in a discrete random variable with 10 unique outcomes for the confidential and synthetic data. We arrange these values with employment [0,90) [90,99) [99,max) in the first three positions and payroll [0,90) [90,99) [99,max) in the last three positions. The support of the confidential data distribution is, therefore, [000000], [100100], [100010], [100001], [010100], [010010], [010001], [001100], [001010], [001001]. The support of the synthetic data variable is identical. The value [100010], for example, means that employment was in the quantile range [0,90), while payroll was in the quantile range [90,99). The value [001001], for example, means that employment and payroll were both in the quantile range [99,max). The value [000000] means that the establishment is not present in that year.

The transition matrix measures the conditional probability of going from a particular value, say [100100], in the confidential data, to each of the ten possible outcomes in the synthetic data. The diagonal elements of the transition matrix measure the probability that the quantile category in the confidential and synthetic data are in the same range. The off-diagonal elements measure the probability that a particular confidential quantile-range outcome maps to each of the different possibilities. Hence, Pr[Synthetic=[001001]|Real=[100100]] is the probability that an observation with employment and payroll both in the [0,90) quantile range becomes an observation with employment and payroll both in the [99,max) quantile range (loosely, both employment and payroll go from common values in the confidential data to extreme values in the synthetic data) and Pr[Synthetic=[001001]|Real=[001001]] is the probability of going from the extreme value quantile range for both variables in the confidential data to the extreme quantile range in the synthetic data for both variables.

Using the 100 replicates from the synthesizer for industry 573, we estimated the 100 elements in this transition matrix (pooling the estimated probabilities across years). The differential privacy bound for this transition matrix was then computed. That bound is 1.9098, which corresponds to an odds ratio of 6.751. The estimated differential privacy bound is low (2 is often used as a reference, corresponding to an odds ratio of 7.39). This property of the synthesizer can be interpreted as follows. An attacker armed with the knowledge of the identities of all but one establishment in industry 573 and with the values of employment and payroll for all of those establishments (coded in the quantile ranges used above) could improve his estimate of the employment and payroll ranges for the unknown establishment by a maximum of 6.751:1 using the synthetic data. That is, if the intruder thought that *a priori* each of the three quantile ranges was equally probable for the establishment with unknown data, after seeing the publication value in the synthetic data, one of those ranges would be 6.751 times more likely than the other two (which one depends upon which synthesized value gets published). That is the best inference that the intruder can make even though the intruder knows all of the synthetic data and all but one row of the confidential data. The protection prevents an attribute disclosure even in this extreme case.

Several caveats apply to this estimate. First, it is only an estimate for two variables (employment and payroll), not the complete set of synthetic variables. Second, it uses the synthesizer to estimate the probabilities as a substitute for using the statistical estimates of the posterior distribution and the exact formulas because the former calculation is easier to program—but not as accurate unless we produce thousands of replicates. Third, industry 573 is a relatively large industry. Attempts to repeat the analysis on a more concentrated industry (372) resulted in sampling zeroes in the pooled (over years) transition matrix, implying that 100 replicates is too few to get a good estimate of the probabilities and that the differential privacy for that industry is larger (the synthetic data are more informative).

6 Concluding Remarks

Work on an expanded and improved SynLBD is underway. Changes and improvements planned for the second phase include: switching from SIC to NAICS; using an updated LBD with additional years of data; and the addition of multiple implicates, geography, firm size and age for establishments in multi-establishment firms, and year of status change for establishments that change their multiunit status during their lifetime. Additionally, we plan to explore the use of more flexible approaches, such as CART models, for modeling payroll and employment so that the bias observed in the job creation and destruction rates is reduced without compromising analytic validity in other areas. As is the nature of a longitudinal database, the LBD continues to grow. Ultimately the Census Bureau would like to have a mechanism by which the SynLBD can be updated regularly as the LBD is updated. The nature of the synthesis, at least in its current form, is that adding years of data will require repeating the entire synthesis process since the Firstyear and Lastyear variables are synthesized. The risk implications of such a scenario is another area of research.

Improvements to the LBD will be guided by a key feature of the infrastructure that accompanies the current release of the SynLBD: analysts who use the SynLBD can request the Census Bureau to provide results of the same analysis run on the confidential LBD data. This mechanism will encourage researchers that might otherwise be hesitant to use synthetic data for research and provide the synthesis team with feedback on important uses of the SynLBD, suggesting areas of improvement. Access to the SynLBD initially will be via remote desktop with all analyses taking place on a Census Bureau server.

Acknowledgements

The research in this paper was conducted by U.S. Census Bureau employees and Special Sworn Status researchers at the Triangle Census Research Data Center. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. This work was supported by NSF grant ITR-0427889. We are grateful to the National Science Foundation for their financial support. We also wish to thank John Haltiwanger, Nick Greenia, Karen Masken, Kevin McKinney, Lars Vilhuber, and Laura Zayatz for their helpful comments.

Notes

[1] See http://www.census.gov/econ/cbp/index.html, http://www.ces.census.gov/index.php/bds, and http://www.bls.gov/bdm/.

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

[3]Employment in the LBD is a point in time measure. It captures employment present during the payroll period including the week of March 12. Establishments that begin operations after March 12 will show zero employment and positive payroll for that year.

References

Abowd, J. & Vilhuber, L. (2008). How protective are synthetic data? In *Privacy in Statistical Databases, Lecture Notes in Computer Science 5262*, Eds. J. Domingo-Ferrer & Y. Saygin, pp. 239–246. Berlin: Springer-Verlag.

Abowd, J.M., Gehrke, J. & Vilhuber, L. (2009). Parameter exploration for synthetic data with privacy guarantees for *OnTheMap*. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*.

Abowd, J.M. & Woodcock, S.D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases*, Eds. J. Domingo-Ferrer & V. Torra, pp. 290–297. New York: Springer-Verlag.

Bartelsman, E., Haltiwanger, J. & Scarpetta, S. (2004). Microeconomic evidence of creative destruction in industrial and developing countries. Discussion Paper 04-114/3, Tinbergen Institute.

Cox, L.H. & Zayatz, L.V. (1995). An agenda for research in statistical disclosure limitation. *J. Official Statist.*, **11**, 205–220.

Davis, S., Haltiwanger, J., Jarmin, R.S. & Miranda, J. (2007). Volatility and dispersion in business growth rates: Publicly traded versus privately held firms. In *NBER Macroeconomics Annual*, Eds. D. Acemoglu, K. Rogoff & M. Woodford, pp.107–180. Cambridge, MA: NBER Inc.

Davis, S., Haltiwanger, J. & Schuh, S. (1996). *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Davis, S.J., Faberman, R.J., Haltiwanger, J., Jarmin, R.S. & Miranda, J. (2008a). Business volatility, job destruction, and unemployment. Technical Report CES-WP-08-26, U.S. Census Bureau Center for Economic Studies.

Davis, S.J., Haltiwanger, J., Jarmin, R.S., Krizan, C.J., Miranda, J., Nucci, A. & Sandusky, K. (2007). Measuring the dynamics of young and small businesses: Integrating the employer and nonemployer universes. Tech. Rep. 13226, NBER.

Davis, S.J., Haltiwanger, J., Jarmin, R.S., Lerner, J. & Miranda, J. (2008b). Private equity and employment. Technical Report CES-WP-05-30, U.S. Census Bureau Center for Economic Studies.

Dreschler, J., Bender, S. & Rassler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. *Trans. Data Privacy*, **1**, 105–130.

Dwork, C. (2006). Differential privacy. In *ICALP, Lecture Notes in Computer Science 4052*, Eds. M. Bughesi, B. Preneel, V. Sasson & I. Wegner, pp. 1–12. Heidelberg: Springer.

Federal Committee on Statistical Methodology (2005). Report on statistical disclosure limitation methodology. Office of Management and Budget, Statistical Policy Working Paper 22.

Fienberg, S.E. (1994). Conflicts between the need for access to statistical information and demands for confidentiality. *J. Official Statist.*, **9**, 115–132.

Foster, L., Haltiwanger, J. & Kim, N. (2006). Gross job flows for the U.S. manufacturing sector: Measurement from the longitudinal research database. Technical Report CES-WP-06-30, U. S. Census Bureau Center for Economic Studies.

Haltiwanger, J., Jarmin, R.S. & Miranda, J. (2009a). Entrepreneurship across states. Technical Report BDS Briefing 3, Kauffman Foundation.

Haltiwanger, J., Jarmin, R.S. & Miranda, J. (2009b). High growth and failure of young firms. Technical Report BDS Briefing 4, Kauffman Foundation.

Haltiwanger, J., Jarmin, R.S. & Miranda, J. (2009c). Jobs created from business startups in the United States. Technical Report BDS Briefing 2, Kauffman Foundation.

Jarmin, R., Klimek, S. & Miranda, J. (2005). The role of retail chains: National, regional, and industry results. Technical Report CES-WP-05-30, U.S. Census Bureau Center for Economic Studies.

Jarmin, R. & Miranda, J. (2002). The longitudinal business database. CES Working Paper CES-WP-02-17, U.S. Census Bureau, Center for Economic Studies.

Kinney, S.K., Karr, A.F. & Gonzalez, J.F., Jr. (2009). Data confidentiality: The next five years summary and guide to papers. *J. Privacy Confident.*, **1**(2), 125–134.

Little, R.J.A. (1993). Statistical analysis of masked data. *J. Official Statist.*, **9**, 407–426.

McGuckin, R. & Pascoe, G. (1988). The longitudinal research database (LRD): Status and research possibilities. Technical Report CES-WP-88-2, U.S. Census Bureau Center for Economic Studies.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Method.*, **29**, 181–189.

Reiter, J.P. (2004). New approaches to data dissemintation: A glimpse into the future (?). *Chance*, **17**, 12–16.
Reiter, J.P. & Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Assoc.*, **102**, 1462–1471.
Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *J. Official Statist.*, **9**, 462–468.
Skinner, C., Marsh, C., Openshaw, S. & Wymer, C. (1994). Disclosure control for census microdata. *J. Official Statist.*, **10**, 31–51.
Willenborg, L. & de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.
Woodcock, S.D. & Benedetto, G. (2009). Distribution-preserving statistical disclosure limitation. *Comput. Stat. Data Anal.*, **53**, 4228–4242.

Résumé

Dans la plupart des pays, les instituts nationaux de statistique ne publient pas les micro-données relatives aux entreprises. Les publier présente en effet un risque trop élevé de rupture de confidentialité. Ce risque peut être évité par un recours à des données synthétiques—des données simulées à partir de modèles statistiques reproduisant la loi des véritables micro-données. Dans cet article, nous décrivons une application de cette stratégie à la création d'une telle base de données à partir des résultats du recensement économique annuel des entreprises américaines. Cette base de donnée comprend plus de 20 millions d'entreprises sur une période remontant à 1976. L'U.S. Bureau of Census et l'Internal Revenue Service ont récemment approuvé la publication sous forme synthétique de ces micro-données, faisant ainsi de la Longitudinal Business Database le premier ensemble de micro-données de ce type accessible au public aux Etats-Unis. Nous expliquons la façon dont cette base de données synthétiques a été créée, comment sa validité a été testée, et comment son risque de rupture de confidentialité a été évalué.

[Received February 2011, accepted July 2011]