JOHN M. ABOWD
*Cornell University*

IAN M. SCHMUTTE
*University of Georgia*

# *Economic Analysis and Statistical Disclosure Limitation*

**ABSTRACT**   This paper explores the consequences for economic research of methods used by data publishers to protect the privacy of their respondents. We review the concept of statistical disclosure limitation for an audience of economists who may be unfamiliar with these methods. We characterize what it means for statistical disclosure limitation to be *ignorable.* When it is not ignorable, we consider the effects of statistical disclosure limitation for a variety of research designs common in applied economic research. Because statistical agencies do not always report the methods they use to protect confidentiality, we also characterize settings in which statistical disclosure limitation methods are *discoverable;* that is, they can be learned from the released data. We conclude with advice for researchers, journal editors, and statistical agencies.

This paper is about the potential effects of statistical disclosure limitation (SDL) on empirical economic modeling. We study the methods that public and private providers use before they publish data. Advances in SDL have unambiguously made more data available than ever before, while protecting the privacy and confidentiality of identifiable information on individuals and businesses. But modern SDL intrinsically distorts the underlying data in ways that are generally not clear to the researcher and that may compromise economic analyses, depending on the specific hypotheses under study. In this paper, we describe how SDL works. We provide tools to evaluate the effects of SDL on economic modeling, as well as some concrete guidance to researchers, journal editors, and data providers on assessing and managing SDL in empirical research.

Some of the complications arising from SDL methods are highlighted by J. Trent Alexander, Michael Davern, and Betsey Stevenson (2010). These

authors show that the percentage of men and women by age in public-use microdata samples (PUMS) from Census 2000 and selected American Community Surveys (ACS) differs dramatically from published tabulations based on the complete census and the full ACS for individuals age 65 and older. This result was caused by an acknowledged misapplication of confidentiality protection procedures at the Census Bureau. As such, it does not reflect a failure of this specific approach to SDL. Indeed, it highlights the value to the Census Bureau of making public-use data available—researchers draw attention to problems in the data and data processing. Correcting these problems improves future data publications.

This episode reflects a deeper tension in the relationship between the federal statistical system and empirical researchers. The Census Bureau does not release detailed information on the specific SDL methods and parameters used in the decennial census and ACS public-use data releases, which include data swapping, coarsening, noise infusion, and synthetic data. Although the agency originally announced that it would not release new public-use microdata samples that corrected the errors discovered by Alexander, Davern, and Stevenson (2010), shortly after that announcement it did release corrections for all the affected Census 2000 and ACS PUMS files.[1] There is increased concern about the application of these SDL procedures without some prior input from data analysts outside the Census Bureau who specialize in the use of these PUMS files. More broadly, this episode reveals the extent to which modern SDL procedures are a black box whose effect on empirical analysis is not well understood.

In this paper, we pry open the black box. First, we characterize the interaction between modern SDL methods and commonly used econometric models in more detail than has been done elsewhere. We formalize the data publication process by modeling the application of SDL to the underlying confidential data. The data provider collects data from a frame defining an underlying, finite population, edits these data to improve their quality, applies SDL, then releases tabular and (sometimes) microdata public-use files. Scientific analysis is conducted on the public-use files.

Our model characterizes the consequences for estimation and inference if the researcher ignores the SDL, treating the published data as though they were an exact copy of the clean confidential data. Whether SDL is ignorable or not depends on the properties of the SDL model and on the

---

1. See the online appendix, section B.1. Supplemental materials and online appendices to all papers in this volume may be found at the *Brookings Papers* web page, www.brookings.edu/bpea, under "Past Editions."

analysis of interest. We illustrate ignorable and nonignorable SDL for a variety of analyses that are common in applied economics.

A key problem with the approach of most statistical agencies to modern SDL systems is that they do not publish critical parameters. Without knowing these parameters, it is not possible to determine whether the magnitude of nonignorable SDL is substantial. As the analysis by Alexander, Davern, and Stevenson (2010) suggests, it is sometimes possible to "discover" the SDL methods or features based on related estimates from the same source. This ability to infer the SDL model from the data is useful in settings where limited information is available. We illustrate this method with a detailed application in section IV.B.

For many analyses, SDL methods that have been properly applied will not substantially affect the results of empirical research. The reasons are straightforward. First, the number of data elements subject to modification is probably limited, at least relative to more serious data quality problems such as reporting error, item missingness, and data edits. Second, the effects of SDL on empirical work will be most severe when the analysis targets subpopulations where information is most likely to be sensitive. Third, SDL is a greater concern, as a practical matter, for inference on model parameters. Even when SDL allows unbiased or consistent estimators, the variance of those estimators will be understated in analyses that do not explicitly correct for the additional uncertainty.

Arthur Kennickell and Julia Lane (2006) explicitly warned economists about the problems of ignoring statistical disclosure limitation methods. Like us, they suggested specific tools for assessing the effects of SDL on the quality of empirical research. Their application was to the Survey of Consumer Finances, which was the first American public-use product to use multiple imputation for editing, missing-data imputation, and SDL (Kennickell 1997). Their analysis was based on the efforts of statisticians to explicitly model the trade-off between confidentiality risk and data usefulness (Duncan and Fienberg 1999; Karr and others 2006).

The problem for empirical economics is that statistical agencies must develop a general-purpose strategy for publishing data for public consumption. Any such publication strategy inherently advantages certain analyses over others. Economists need to be aware of how the data publication technology, including its SDL aspects, might affect their particular analyses. Furthermore, economists should engage with data providers to help ensure that new forms of SDL reflect the priorities of economic research questions and methods. Looking to the future, statisticians and computer scientists have developed two related ways to address these issues more

systematically: synthetic data combined with validation servers and privacy-protected query systems. We conclude with a discussion of how empirical economists can best prepare for this future.

## I.  Conceptual Framework and Motivating Examples

In this section we lay out the conceptual framework that underlies our analysis, including our definitions of *ignorable* versus *nonignorable* SDL. We also offer two motivating examples of SDL use that will be familiar to social scientists and economists: randomized response for eliciting sensitive information from survey respondents and the effect of topcoding in analyzing income quantiles.

### I.A.  Key Concepts

Our goal is to help researchers understand when the application of SDL methods affects the analysis. To organize this discussion, we introduce key concepts that we develop in a formal model in the online appendix. We assume the analyst is interested in estimating features of the model that generated the confidential data. However, the analyst only observes the data after the provider has applied SDL. The SDL is, therefore, a distinct part of the process that generates the published data.

We say the SDL is *ignorable* if the analyst can recover the estimates of interest and make correct inferences using the published data without explicitly accounting for SDL—that is, by using exactly the same model as would be appropriate for the confidential data. In applied economic research it is common to implicitly assume that the SDL is ignorable, and our definition is an explicit extension of the related concept of *ignorable missing data.*

If the data analyst cannot recover the estimate of interest without the parameters of the SDL model, the SDL can then be said to be *nonignorable.* In this case, the analyst needs to perform an SDL-aware analysis. However, the analyst can only do so if either (i) the data provider publishes sufficient details of the SDL model's application to the confidential data, or (ii) the analyst can recover the parameters of the SDL model based on prior information and the published data. In the first case, we call the nonignorable SDL *known.* In the second case, we call the nonignorable SDL *discoverable.*

### I.B.  Motivating Examples

Consider two examples of SDL familiar to most social scientists. The first is randomized response, which allows a respondent to answer

a sensitive question truthfully without revealing the answer to the interviewer. This yields more accurate responses, since respondents are more likely to answer truthfully, but at the cost of adding noise to the data. The second example is income topcoding, which is a form of SDL that protects the privacy of high-income households. This example highlights the fact that the ignorability of SDL is a function not just of the SDL method but also of the estimand of interest.

RANDOMIZED RESPONSE   Stanley Warner (1965) proposed a survey technique in which the respondent is presented with one of two questions that can both be answered either "yes" or "no." The interviewer does not know the question. The respondent opens an envelope drawn from a basket of identical envelopes, reads the question silently, responds "yes" or "no," and then destroys the question. With a certain probability the question is sensitive (for example, "Have you ever committed a violent crime?"), and with a complementary probability the question is innocuous (for example, "Is your birthday between July 1st and December 31st?"). Again, the interviewer records only the "yes" or "no" answer and never sees the true question.

If one runs this single-question survey on a sample of 100 people chosen randomly, the estimated proportion of "yes" answers has an expected value equal to the probability that the respondent was asked the sensitive question times the population probability (in our example) of having committed a violent crime plus the complement of the probability that the respondent was asked the sensitive question times one-half. If the sample mean proportion of "yes" answers is 26 percent, then to recover the implied estimate for the population probability of having committed a violent crime one needs to know the probability that the sensitive question was asked. The standard error of the estimated proportion of "yes" answers is 4.4 percent, but the standard error for the estimated population proportion of having committed a violent crime is 4.4 percent divided by the probability that the respondent was asked the sensitive question.

Why is this a form of statistical disclosure limitation? Because no one other than the respondent knows which question was asked, this procedure places bounds on the amount of information that anyone, including the interviewer, can learn about the respondent's answer to the sensitive question. (See section II.B for a complete discussion.) This form of SDL is obviously not ignorable. The data analyst does not care about the 26 percent but wants to estimate the proportion of people who have committed a violent crime. The data publisher adds the following documentation about the SDL

parameters: *Only half the respondents were asked the sensitive question; the other half were asked a question for which half the people in the population would answer "yes."* Now the analyst can estimate that the proportion who committed a violent crime is 2 percent, and its standard error is 8.8 percent. Notice that the SDL affected both the mean and the standard error of the estimate.

CONSEQUENCES OF TOPCODING FOR QUANTILE ESTIMATION Richard Burkhauser and others (2012) provide a simple, vivid example of the consequences of SDL for economic analysis. Because of SDL, changes in the upper tail of the income distribution are largely hidden from view in research based on public-use microdata, most often the Current Population Survey (CPS). Because income is a sensitive data item, and large incomes can be particularly revealing in combination with other information, the Census Bureau and the Bureau of Labor Statistics both censor incomes above a certain threshold in their public-use files. The topcoding of income protects privacy, but it also limits what can be done with the data.

Burkhauser and others (2012) report that the income topcode results in 4.6 percent of observations being censored. Thus, the topcoded data are perfectly fine for measuring the evolution of the 90-10 quantile ratio but completely useless for measuring the evolution of incomes among the top 1 percent of households, as was revealed when Thomas Piketty and Emmanuel Saez (2003) analyzed uncensored income data based on Internal Revenue Service (IRS) tax filings. Piketty and Saez (2003) showed that trends in income inequality look quite different in the administrative record data than in the CPS. Using restricted-access CPS data, Burkhauser and others (2012) showed that the difference between the administrative and survey data was largely due to censoring in the survey data.

If we could observe all the confidential data, *Y,* they would have probability distribution function $p_Y(Y)$ and cumulative distribution function $F_Y(Y)$. For studying income inequality, interest centers on the quantiles of $F_Y$, defined by the inverse cumulative distribution function $Q_Y$. When drawing inferences about the quantiles of the income distribution, topcoding is irrelevant for all quantiles that fall below the top-coding threshold, *T.* We say top-coding is *ignorable* if, for a given quantile point of interest $p \in [0, 1]$, $Q_Z(p) = Q_Y(p)$, where $Q_Z(p)$ is the quantile function of the published data, *Z.*

This very familiar example highlights several features of ignorable and nonignorable SDL. First, whether SDL can be ignored depends on both the properties of the SDL mechanism and the specific estimand of interest.

Second, assessing the effect of SDL requires knowledge of the mechanism. If the value of the topcode threshold $T$ were not published, it would not be possible for the researcher to assess whether a specific quantile of interest could be learned from the published data. The researcher might learn the topcode by inspecting the published data. In this case, we say the topcode is a *discoverable* form of SDL.

The work of Jeff Larrimore and others (2008) also illustrates how, when armed with information about SDL methods and access to the confidential data, researchers can improve their analysis with minimal change to the risk of harmful or unlawful data disclosure. Larrimore and others (2008) published new data for 24 separate income series for 1976–2006 that contain the mean values of incomes above the topcode values within cells, disaggregated by race, gender, and employment status. They show that these cell means can be used with the public-use CPS microdata to analyze the income distribution in ways that would otherwise require direct access to the confidential microdata.

In the randomized response example, the SDL model is *known* as long as the probability that the sensitive question was asked is disclosed. Without disclosure of this probability, the researcher is unable to perform an *SDL-aware analysis* because it is *not discoverable.* By contrast, an undisclosed topcode level may still be discoverable by a researcher through inspection of the data.

## II. The Basics of Statistical Disclosure Limitation

The key principle of confidentiality is that individual information should only be used for the statistical purposes for which it was collected. Moreover, that information should not be used in a way that might harm the individual (Duncan, Jabine, and de Wolf 1993, p. 3). This principle embodies two distinct ideas. First, individuals have a property right of privacy covering their personal information. Second, once such personal data have been shared with a trusted curator, individuals should be protected against uses that could lead to harm. These ideas are reflected in the development and implementation of SDL among data providers. For the United States, the Federal Committee on Statistical Methodology (Harris-Kojetin and others 2005) has produced a very thorough summary of the objectives and practices of SDL.

The constant evolution of information technology makes it challenging to translate the principle of confidentiality into policy and practice. The statutes that govern how statistical agencies approach SDL explicitly

prohibit any breach of confidentiality.[2] However, statisticians and computer scientists have formally proven that it is impossible to publish data without compromising confidentiality, at least probabilistically. We touch in our conclusion on how public policy should adapt in light of new ideas about SDL and privacy protection. The current period of tension also characterizes the broader co-evolution of science and public policy around SDL, which we briefly review.

### II.A. What Does SDL Protect?

SDL may appear to protect against unrealistic, fictitious, or overblown threats. Reports of data security breaches, in which hackers abscond with terabytes of sensitive individual information, are increasingly common, but it has been roughly six decades since the last reported breach of data privacy within the federal statistical system (Anderson and Seltzer 2007, for household data; Anderson and Seltzer 2009, for business data). One is hard-pressed to find a report of the American Community Survey, for example, being "hacked." Yet it is important to acknowledge that the principle of confidentiality for statistical agencies arose from very real and deliberate attempts by other government agencies to use the data collected for statistical purposes in ways that were directly harmful to specific individuals and businesses.

Laws to protect data confidentiality arose from the need to separate the statistical and enforcement activities of the federal government (Anderson and Seltzer 2007; 2009). These laws were subsequently weakened and violated in a small but influential number of cases. For example, the U.S. government obtained access to confidential decennial census information to help locate German and Japanese Americans during World Wars I and II, and from the economic census to assist with war planning. The privacy laws were subsequently strengthened, in part because businesses were quite reluctant to provide information to the Census Bureau for fear that it could either be used for tax or antitrust proceedings or be used by their

---

2. U.S. Code Title 13, Section 9, governing the Census Bureau, prohibits "any publication whereby the data furnished by any particular establishment or individual under this title can be identified" (see https://www.law.cornell.edu/uscode/text/13/9, accessed August 6, 2015). U.S. Code Title 5, Section 552a (part of the Confidential Information Protection and Statistical Efficiency Act of 2002), which governs all federal statistical agencies, requires them to "establish appropriate administrative, technical, and physical safeguards to insure the security and confidentiality of records and to protect against any anticipated threats or hazards to their security or integrity which could result in substantial harm, embarrassment, inconvenience, or unfairness to any individual on whom information is maintained" (see https://www.law.cornell.edu/uscode/text/5/552a, accessed August 6, 2015).

competitors to reveal trade secrets. The statistical agencies therefore also have a pragmatic interest in laws that protect individual and business information against intrusions by other parts of the federal and state governments, since these laws directly affect willingness to participate in censuses and surveys.

The modern proliferation of data and advances in computing technology have led to new concerns about data privacy. We now understand that it is possible to identify an individual from a very small number of demographic attributes. In a much-cited study, Latanya Sweeney (2000) shows how then publicly available hospital records might be linked to survey data to compromise confidentiality. Arvind Narayanan and Vitaly Shmatikov (2008) show that supposedly anonymous user data published by Netflix can be re-identified. Although no harm was documented in these cases, they highlight the potential for harm in the world of big data.

Paul Ohm (2010) argues that for every individual there may be a "database of ruin" that can be constructed by linking together existing non-ruinous data. That is, there may be one database with some embarrassing or damaging information, and another database with personally identifiable information to which it may be linked, perhaps through a sequence of intermediate databases. In some cases, there are clear financial incentives to seek out such a database of ruin. A potential employer or insurer may have an interest in learning health information that a prospective employee would rather not disclose. If such information could be easily and cheaply gleaned by combining publicly available data, economic intuition suggests that firms might do so, despite the absence of documented instances of such behavior. An alternative perspective is offered by Jane Yakowitz (2011), who argues for legal reforms that reduce the emphasis on hypothetical threats to privacy and expand the emphasis on the benefits from providing accurate, timely socioeconomic data.

### II.B.  Concepts and Methods of SDL

Modern SDL methods are designed to allow high-quality statistical information to be published while protecting confidentiality. Since many applied researchers may have an incomplete awareness of and knowledge about the ways in which SDL distorts published data, we provide an overview of the most common SDL methods applied to economic and demographic data. For a more technical and detailed treatment, we refer the reader to two recent works on SDL and formal privacy models: *Statistical Confidentiality: Principles and Practice* by George Duncan, Mark Elliot, and Juan-José

Salazar-González (2011), and "The Algorithmic Foundations of Differential Privacy" by Cynthia Dwork and Aaron Roth (2014).

A TAXONOMY OF THREATS TO CONFIDENTIALITY  Confidentiality may be violated in many related ways. An *identity disclosure* occurs if the identity of a specific individual is completely revealed in the data. This can occur because a unique identifier is released or because the information released about a respondent is enough to uniquely identify him or her in the data. An *attribute disclosure* occurs when it is possible to deduce from the published data a specific confidential attribute of a given respondent.

Modern SDL and formal privacy systems treat disclosure risk probabilistically. From this perspective, the problem is not merely that published data might perfectly identify a respondent or his or her attributes. Rather, it is that the published data might allow a user to infer a respondent's identity or attributes with high probability. This concept, known as *inferential disclosure,* was introduced by Tore Dalenius (1977) and formalized by Duncan and Diane Lambert (1986) in statistics, and by Shafi Goldwasser and Silvio Micali (1982) in computer science.

Suppose the published data are denoted $Z$. A confidential variable $y_i$ is associated with a specific respondent $i$. The prior beliefs of a user about the value of $y_i$ are represented by a probability distribution, $p(y_i)$, that reflects information from all other sources. Then $p(y_i|Z)$ represents the updated—posterior—beliefs of the user about the value of $y_i$ after the data $Z$ are published. An inferential disclosure has occurred if the posterior beliefs are too large relative to prior beliefs.

Our example of randomized response from section I.B provides intuition about inferential disclosure. The probability that the respondent will answer "yes" given that the truth is "yes" is 75 percent. The probability that the respondent will answer "yes" given that the truth is "no" is 25 percent. These two probabilities are *entirely* determined by the probability that the respondent was asked the sensitive question and the probability that the answer to the innocuous question is "yes." They *do not depend* on the unknown population probability of having committed a violent crime. The ratio of these two probabilities is the Bayes factor—the ratio of the posterior odds that the truth is "yes" versus "no" given the survey answer "yes" to the prior odds of "yes" versus "no." The interviewer learns from a "yes" answer that the respondent is three times as likely as a random person to have committed a violent crime, and that is *all* the interviewer learns. Had the violent crime question been asked directly, the interviewer could have updated his posterior beliefs by a much larger factor—potentially infinite if the respondent answers truthfully.

Moving forward, it is important to keep the concept of inferential disclosure in mind for two reasons. First, it leads to a key intuition: It is impossible to publish useful data without incurring some threat to confidentiality. A privacy protection scheme that provably eliminates all inferential disclosures is equivalent to a full encryption of the confidential data and therefore useless for analysis.[3] Second, to be effective against inferential disclosure, certain SDL methods require that statistical agencies also conceal the details of their implementation. For example, with swapping, knowledge of the swap rate would increase inferential disclosure risk by improving the user's knowledge of the full data publication process. We will argue later that researchers, and agencies, should prefer SDL methods whose details can be made publicly available.

### II.C. SDL Methods for Microdata

SUPPRESSION  Suppression is one of the most common forms of SDL. Suppression can be used to eliminate an entire record from the data or to eliminate an entire attribute. Record-level suppression is ignorable under the same assumptions that lead to ignorable missing data models in general. However, if the suppression rule is based on data items deemed to be sensitive, then it is very unlikely that the data were suppressed at random. In that case, knowledge of the suppression rule along with auxiliary information from the underlying microdata is extremely useful in assessing the effect of suppression on any specific application. Sometimes suppression is combined with imputation; this occurs when sensitive information is suppressed and then replaced with an imputed value.

AGGREGATION  Aggregation refers to the coarsening of values a variable can take, or the combination of information from multiple variables. The canonical example is the Census Bureau's practice of aggregating geographic units into Public-Use Microdata Areas (PUMAs). Likewise, data on occupation are often reported in broad aggregates. The aggregation levels are deliberately set in such a way that the number of individuals represented in the data have some combination of attributes that exceeds a certain threshold. Aggregation is what prevents a user from, say, looking up the income of a 42-year-old economist living in Washington, D.C. Other forms of aggregation are quite familiar to empirical researchers, such as topcoding income, and reporting income in bins rather than in levels. These

---

3. Evfimievski, Gehrke, and Srikant (2003) and Dwork (2006) prove it is impossible to deliver full protection against inferential disclosures, using different, but related, formalizations of the posterior probabilities.

methods are well understood by researchers, and their effects on empirical work have been carefully studied. In many cases, it is easy to determine whether aggregation is a problem for a particular research application; in such cases, one possible solution is to obtain access to the confidential, disaggregated data.

NOISE INFUSION  Noise infusion is a method in which the underlying microdata are distorted using either additive or multiplicative noise. The infusion of noise is not generally ignorable. If applied correctly, noise infusion can preserve conditional and unconditional means and covariances, but it always inflates variances and leads to attenuation bias in estimated regression coefficients and correlations among the attributes (Duncan, Elliot, and Salazar-González 2011, p. 113). To assess the effects for any particular application, researchers need to know which variables have been infused with noise along with information about any relevant parameters governing the distribution of noise. If such information is not published, it may be possible to infer the noise distribution from the public-use data if there are multiple releases of information based on the same underlying frame. We illustrate this possibility in our analysis of the public-use Quarterly Workforce Indicators (QWI), Quarterly Census of Employment and Wages (QCEW), and County Business Patterns (CBP) data in section IV.B.

DATA SWAPPING  Data swapping is the practice of switching the values of a selected set of attributes for one data record with the values reported in another record. The goal is to protect the confidentiality of sensitive values while maintaining the validity of the data for specific analyses. To implement swapping, the agency develops an index based on the probability that an individual record can be re-identified.[4] Sensitive records are compared to "nearby" records on the basis of a few variables. If there is a match, the values of some or all of the other variables are swapped. Usually, the geographic identifiers are swapped, thus effectively relocating the records in each other's location.

For example, in Athens, Georgia, there may be only one male household head with 10 children. If that man participates in the ACS and reports his income, it would be possible for anyone to learn his income by simply reading the unswapped ACS. To protect confidentiality, the entire data record can be swapped with the record of another household in a different geographic area with a similar income.

---

4. See Reiter (2005), Skinner and Holmes (1998), and Skinner and Shlomo (2008) for specifics on the risk indexes and Duncan, Elliot, and Salazar-González (2011, p. 114) for a review of historical uses of swapping.

Swapping preserves the marginal distribution of the variables used to match the records at the cost of all joint and conditional distributions involving the swapped variables. The computer science community has frequently criticized this approach to confidentiality protection because it does not meet the "cryptography" standard: an encryption algorithm is provably secure when all details and parameters, except the encryption key, can be made public without compromising the algorithm. SDL algorithms like swapping are not provably effective when too many of their parameters are public. That is why the agencies do not publish them or release more than a few details of their swapping procedures.

The lack of published details is what makes input data swapping so insidious for empirical research. Matching variables, the definition of "nearby," and the rate at which sensitive and nonsensitive records are swapped can all affect the data analyses that use those variables, so parameter confidentiality makes it difficult to analyze the effects of swapping. Furthermore, even restricted-access arrangements that permit use of the confidential data may still require the use of the swapped version, even if other SDL modifications of the data have been removed. Some providers even destroy the unswapped data.

SYNTHETIC MICRODATA  Synthetic microdata involve the publication of a data set with the same structure as the confidential data, in which the published data are drawn from the same data-generating process as the confidential data but some or all of the confidential data have been suppressed and imputed. The confidential data, $Y$, are generated by a model, $p(Y|\theta)$, parameterized by $\theta$. The synthetic microdata are drawn from $p(\tilde{Y}|Y)$, the posterior predictive distribution for the data process given the observed data, which has been estimated by the statistical agency.

When originally proposed by Roderick Little (1993) and Donald Rubin (1993), synthetic data methods mimicked procedures that already existed for missing-data problems. Synthetic data methods impose an explicit cost on the researcher—imputed data replacing actual data—in exchange for an explicit benefit, namely the correct estimation and inference procedures that are available for the synthetic data. The Little–Rubin forms of synthetic data analysis are guaranteed to be SDL-aware. If the researcher's hypothesis is among those for which correct inference procedures are available, then the synthetic data are provably analytically valid. John Abowd and Simon Woodcock (2001), Trivellore Raghunathan, Jerome Reiter, and Rubin (2003), and Reiter (2004) have refined the Little–Rubin methods, allowing them to be applied to complex survey data and combined with

other missing data imputations. They have also shown that the class of hypotheses with provable analytical validity is limited by the models used to estimate $p(\tilde{Y}|Y)$.

Synthetic data can only be used by themselves for certain types of research questions—those for which they are analytically valid. This set of hypotheses depends on the model used to generate the synthetic data. For example, if the confidential data are 10 discrete variables and the synthetic data are generated from a model that includes all possible inter-actions of two of these variables, then any research question involving only two variables can be analyzed in a correct, SDL-aware manner from the synthetic data. The analyst does not need access to the confidential data. But no model involving three or more variables can be analyzed correctly from the synthetic data. Such models require that the analyst have access to the confidential data. When the model used to produce the synthetic data is publicly available, researchers can assess whether a given synthetic data set is appropriate for a specific question.

Synthetic data can also be used as a framework for the development of models, code, and hypotheses. For example, researchers can sometimes develop models using the synthetic data, which are public, and then run those models on the confidential data. These applications form part of a feedback loop in which external researchers help provide improvements to the synthetic data model. We discuss synthetic data and the feedback loop in more detail in section VI.A.

FORMAL PRIVACY MODELS  Formal privacy models emerged from data-base security and cryptography. The idea is to model the publication of data by the statistical agency using a *randomized mechanism* that answers statistical questions after adding noise to the properly computed answer in the confidential data. This is known in SDL as *output distortion.* Breaches of privacy are modeled as a game between users, who try to make inferential disclosures from the published data, and the statistical agency, which tries to limit these disclosures.

Dwork (2006) and Dwork and others (2006) formalized the privacy protection associated with output-distortion SDL in a model called ε-differential privacy. For economists, Ori Heffetz and Katrina Ligett (2014) provide a very accessible introduction. Dwork and Roth (2014), in section 3, use our running example of randomized response to character-ize ε-differential privacy. In ε-differential privacy, the SDL must put an upper bound, ε, on the Bayes factor. In our example, ε = ln (Bayes factor bound) = ln 3 = 1.1. Bounding the Bayes factor implies that the maxi-mum amount the interviewer can learn from a "yes" answer is that the

respondent (in our original example) is three times as likely as a random person in the population to have committed a violent crime.

With formal privacy-protected data publication systems, there are provable limits to the amount of privacy loss that can be experienced in the population even under worst-case outcomes. These systems also have provable accuracy for a specific set of hypotheses. From a researcher perspective, then, formal privacy systems and synthetic data are very similar— only some hypotheses can be studied accurately, and these are determined by the statistical queries answered in the formal privacy model. For example, in a case where the confidential data are, once again, 10 discrete variables, and the formal privacy system publishes a protected version of every two-way marginal table, then, once again, any hypothesis involving only two variables can be studied correctly. Likewise, no hypotheses involving three or more variables can be studied correctly without additional privacy-protected publications. Whether these computations can be safely performed by the formal privacy system depends on whether any privacy budget remains. If the privacy budget has been exhausted by publishing all two-way tables, then no further analysis of the confidential data is permitted.

Synthetic data and formal privacy methods are converging. In the SDL literature, researchers now analyze the confidentiality protection provided by the synthetic data (Kinney and others 2011; Benedetto and Stinson 2015; Machanavajjhala and others 2008). In the formal privacy literature, analysts may choose to publish the privacy-protected output as synthetic data—that is, in a format that allows an analyst to use the protected data as if they were the confidential data (Hardt, Ligett, and McSherry 2012). The analysis of synthetic data produced by a formal privacy system is not automatically SDL-aware. The researcher has to use the published features of the privacy model to correct the estimation and the inference.

### II.D. SDL Methods for Tabular Data

Tabular data present confidentiality risks when the number of entities contributing to a particular cell in a table is small or the influence of a few of the entities on the value of the cell is large, such as for magnitudes like total payroll. A sensitive cell is one for which some function of the cell's microdata falls above or below a threshold set by an agency-specific rule. The two most common methods for handling sensitive cells are forms of randomized rounding, which distorts the cell value and may distort other cells as well, and the more common method of suppression. An alternative to suppression is to build tables after adding noise to the input microdata.

SUPPRESSION  Suppression deletes the values for sensitive cells from the published data. From the outset, it was understood that primary suppression—not publishing easily identified data items—does not protect anything if an agency publishes the rest of the data, including summary statistics (Fellegi 1972). In such a case, users could infer the missing items from what was published. Agencies that rely on suppression for tabular data make *complementary* suppressions to reduce the probability that a user can infer the sensitive items from the published data.

Suppressions introduce a missing-data problem for researchers. Whether that missing-data problem is ignorable or not depends on the nature of the model being analyzed and the manner in which suppression is done. An analysis using geographical variation for identification will benefit from using data where industrial classifications were used for the complementary suppressions, whereas an analysis that uses industrial variation will benefit from using data where the complementary suppressions were made using geographical classifications. Ultimately, the preferences of the agency that chooses the complementary suppression strategy will determine which analyses have higher data quality. As with swap rates, agencies rarely publish details of their methods for choosing complementary suppressions.

INPUT DISTORTION  Input distortion of the microdata is another method for protecting tabular data. Using this method, an agency distorts the value of some or all of the inputs before any publication tables are built, and then computes all, or almost all, of the cells using only the distorted data.

### II.E.  Current Practices in the U.S. Statistical System

The SDL methods in the decentralized U.S. statistical system are varied. The most thorough analysis of this topic is the one published by the Federal Committee on Statistical Methodology (FCSM), which is organized by the chief statistician of the United States in the Office of Management and Budget (Harris-Kojetin and others 2005). We summarize the key features of the FCSM report and, where possible, provide updated information on certain data products used extensively by economists. It is incumbent upon the researcher to read the relevant documentation and, if necessary, contact the data provider to obtain nonconfidential publications detailing how the data were collected and prepared for publication, including which methods of SDL were applied.

The goal of the FCSM report is to characterize best practices for SDL, and it contains a table presenting the methods employed by each agency to protect microdata and tabular data (Harris-Kojetin and others 2005, p. 53). As of 2005, the table shows, almost all federal agencies that published

microdata reported using some form of nonignorable, undiscoverable data perturbation. The Census Bureau's stated policy is "for small populations or rare characteristics, noise may be added to identifying variables, data may be swapped, or an imputation applied to the characteristic" (Harris-Kojetin and others 2005, p. 40). Many other agencies, including the Bureau of Labor Statistics (BLS) and National Science Foundation (NSF), contract with the Census Bureau to conduct surveys and therefore use the same or similar guidelines for SDL. The National Center for Education Statistics (NCES) also reports using ad hoc perturbation of the microdata to prevent matching, including swapping and "suppress and impute" for sensitive data items.

In a recent technical report by Amy Lauger, Billy Wisniewski, and Laura McKenna (2014), the Census Bureau released up-to-date information on its SDL methods. In addition to information about discoverable SDL methods, like geographic thresholds and topcoding, the report describes in more detail how noise is added to microdata to protect confidentiality. Specifically, it states that "noise is added to the age variable for persons in households with 10 or more people," and that "noise is also added to a few other variables to protect small but well-defined populations but we do not disclose those procedures" (Lauger, Wisniewski, and McKenna 2014, p. 2).

This Census Bureau report also confirms that swapping is the primary SDL method used in the ACS and decennial censuses. The swapping method targets records that have high disclosure risk due to some combination of rare attributes, such as racial isolation in a particular location. The records at risk are matched on the basis of an unnamed set of variables and swapped into a different geography. In the past few years, the Census Bureau has changed the set of items it uses to determine whether a record is at risk and should be swapped, and the swap rate has increased slightly. The Census Bureau performed an evaluation of the effects of swapping on the quality of published tabular statistics, but it has not published its evaluation results due to concerns that they might compromise the SDL procedures themselves.

One Census Bureau official whom we interviewed said the rate of swapping is low relative to the rate at which data are edited for other purposes. Furthermore, the official said, swapping is applied to cases that are extreme outliers on some particular combination of variables. Without getting more precise, the official conveyed that swapping, while potentially of considerable concern, may have substantially less effect on economic research than, say, missing-data imputation.

Within the last 10 years the Census Bureau has also begun producing data based on more modern SDL methods. The Quarterly Workforce Indicators are protected using an input noise infusion method that, among other features, eliminates the need for cell suppression in count tables. The Census Bureau also offers synthetic microdata from the linked SIPP/SSA/IRS data, the Longitudinal Business Database, and the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES).[5]

## III. How SDL Affects Common Research Designs

In this section, we demonstrate how to apply the concepts of ignorable and nonignorable SDL in common applied settings. In most cases, SDL is nonignorable, and researchers therefore need to know some properties of the SDL model that was applied to their data. When the SDL model is not known, it may still be *discoverable* in the manner introduced in section I.A.

### III.A. Estimating Population Proportions with Noise Infusion

This example is motivated by the SDL procedure that is used to mask ages in the Census 2000, ACS, and CPS microdata files. Although the misapplication of the procedure has been corrected for Census 2000 and ACS, current versions of the CPS for the mid-2000s may still be affected by the error, and have not been reissued. See the online appendix, section B, for more details.

Suppose the confidential data contain a binary variable (such as gender) and a multicategory discrete variable (such as age). We are interested in estimation and inference for the age-specific gender distribution, where $\beta$, the conditional probability of being male given age, is the parameter of interest. When age has been subjected to SDL, using published age to compute these conditional probabilities will lead to problems. The estimated probability of being male conditional on age is affected by the SDL, even though the gender variable was not itself altered by the SDL.

Using the generalized randomized response structure, suppose that we know the probability that the published age data are unaltered. With probability $\rho$, the observed male/female value comes from the true age category. With the complementary probability, the observed outcome is a

---

5. See for example U.S. Census Bureau (2013a, 2013b, 2015).

binary random variable with expected value $\mu \neq \beta$. For example, $\mu$ might be the average value of the proportion male for all age categories at risk to be changed by the SDL model. In any case, $\mu$ is unknown.

Equation B.16 in the online appendix shows that if we ignore the SDL, the conditional probability estimator and its variance are biased. An SDL-aware estimator for the conditional probability of being male for a given age is $\hat{\beta} = [\bar{z}_1 - (1 - \rho)\,\mu]/\rho$, where $\bar{z}_1$ is the estimated sample proportion of males of the chosen age. The estimator for the conditional proportion of interest $\hat{\beta}$ is confounded by the two SDL parameters, except in the special case that $\rho = 1$, which implies that no SDL was applied to the published age data. If all of the observations have been subjected to SDL, then $\hat{\beta}$ is undefined, and the expected value of $\bar{z}_1$ is just $\mu$. In the starkest possible terms, the estimator in equation B.16 is hopelessly underidentified in the absence of information about $\rho$ and $\mu$.

If $\rho$ and $\mu$ are not known, they may still be discoverable if the analyst has access to estimates of conditional probabilities like $\beta$ from an alternative source. See the online appendix, section B, for more details of the application to the Census 2000 and ACS PUMS that generalizes the analysis in Alexander, Davern, and Stevenson (2010). This procedure can be used to discover the SDL in any data set, for example the CPS, for which alternative reliable published estimates of the gender-specific age distribution are available.

The SDL process is still underidentified if we consider only a single outcome like the gender-age distribution, but there are quite a few other binary outcomes that could also be studied, conditional on age—for example, marital status, race, and ethnicity. The differences between Census 2000 estimates of the proportion married at age 65 and older and their comparable Census 2000 PUMS estimates have exactly the same functional form as online appendix equation B.17 with exactly the same SDL parameters. Since these proportions condition on the same age variable, all the other outcomes that also have an official Census 2000 or ACS published proportion can be used to estimate the unknown SDL parameters. The identifying assumptions are (i) that all proportions are conditioned on the same noisy age variable, and (ii) that the noisy age variable can be reasonably modeled as randomized-response noise. We implement a similar method in section IV.B.

### III.B.  Estimating Regression Models

We next consider the effect of SDL on linear regression models. First, we analyze SDL applied to the dependent variable, assuming that the agency

replaces sensitive values with model-based imputed values. This form of SDL is nonignorable for parameter estimation and inference. Parameter estimates will be attenuated and standard errors will be underestimated. Furthermore, this form of SDL is not discoverable, except when there are two data releases from the same frame that use different, independent SDL processes.

Our analysis draws on the work of Barry Hirsch and Edward Schumacher (2004) and Christopher Bollinger and Hirsch (2006), who study the closely related problem of bias from missing-data imputation in the CPS. Respondents to the CPS commonly fail to provide answers to certain questions. In the published data, the missing values are imputed semi-parametrically, conditional on a set of variables. Hirsch and Schumacher (2004) observe that if union status is not in the conditioning set for the imputation model, the union wage gap will be underestimated when using imputed and non-imputed values in a regression of log wages on union status. This bias is exacerbated by using additional controls. The result occurs because if union status is not in the imputation model's conditioning set, then some union workers are imputed nonunion wages, and some nonunion workers are imputed union wages. Bollinger and Hirsch (2006) show that these results hold very generally.

There are two key differences in our approach. First, assessing bias from missing-data imputation is feasible because the published data include an indicator variable that flags which values were reported and which were imputed. With SDL, the affected records and variables are not flagged. Second, in the SDL application, the published data can be imputed using the distribution of the confidential data. This means that the agency does not have to use an ignorable missing-data model when doing imputations for SDL. When imputing actual missing data, which was the subject of the Bollinger and Hirsch (2006) paper, the agency does assume that the missing data were generated by an ignorable inclusion model. The direct consequence is that the model used to impute the suppressed values can be conditioned on all of the confidential data, including the rule that determines whether an item will be suppressed. More succinctly, the analysis below demonstrates the effect of using an imputation model (or swapping rule) that does not contain a regressor of interest, and thus is not conflated with any bias that could arise from nonrandomness of the suppression rule.

SDL APPLIED TO THE DEPENDENT VARIABLE  The model of interest is the function $E[y_{i1}|y_{i2}] = \alpha + y_{i2}\,\beta$. In the published data, sensitive values of the outcome variable $y_{i1}$ are suppressed and imputed. The variable $\gamma_i$ indicates whether $y_{i1}$ is suppressed and imputed. When $\gamma_i = 1$, the confidential

data are published without modification. When $\gamma_i = 0$, the value for $y_{i1}$ is replaced with an imputed value, $z_{i1}$, which is drawn from $p_{Y_1|X}(y_{i1}|x_i, \gamma_i = 0)$, the conditional distribution of the outcome variable given $x_i$ among suppressed observations. The conditioning information used in the imputation model, $x_i = f_I(y_{i2})$, is a function $f_I$ that maps all of the available conditioning information in $y_{i2}$ into a vector of control variables $x_i$.

The simplest example is a model in which $x_i$ consists of a strict subset of variables in $y_{i2}$. For example, in Hirsch and Schumacher (2004), $y_{i2}$ is a set of conditioning variables that includes an indicator for union membership, and $x_i$ is the same set of conditioning variables but excluding the union membership indicator. Like the suppression model, the features of the imputation model, including the function $f_I$, are known only to the agency and not to the analyst.

The released data are $z_{i1} = y_{i1}$ if $\gamma_i = 1$ and $z_{i1} \sim p_{Y_1|X}(y_{1i}|x_i, \gamma_i = 0)$ otherwise. For the other variables, $z_{2i} = y_{2i}$. The marginal probability that the exact confidential data are published is $\Pr[\gamma_i = 1] = \rho$. So the suppression rate is $(1 - \rho)$, an exact analogue of the rate at which irrelevant data replace good data in randomized response. Finally, note that nothing in this specification requires independence between the decision to suppress, $\gamma_i$, and the data values, $y_{i1}$ and $y_{i2}$.

The effects of statistical disclosure limitation in this context are generically nonignorable except for two unusual cases. If no observations are suppressed ($\rho = 1$), then the SDL is ignorable because it is irrelevant. In the more interesting case, the characteristics, $x_i$, perfectly predict $z_{2i}$, and the SDL model is also ignorable for consistent estimation of $\beta$. This case is interesting because it occurs when the agency conditions on all covariates of interest, $y_{2i}$, when imputing $y_{1i}$, and then releases $y_{2i}$ without any additional SDL. Even in this latter case, while the SDL is ignorable for consistent estimation of $\beta$, it is not ignorable for inference. The SDL model introduces variance that is not included in the standard estimator for the variance of $\hat{\beta}$.

The effects of SDL on estimation and inference could be assessed and corrected if the analyst knew two key properties of the SDL model: (i) the suppression rate, $(1 - \rho) = \Pr[\gamma_i = 0]$; and (ii) the set of characteristics used to impute the suppressed observations, $x_i$. At present, almost nothing is known in the research community about either characteristic of the SDL models used in many data sets. See online appendix, section C.1, for details.

SDL APPLIED TO A SINGLE REGRESSOR  If SDL is applied to a single regressor rather than to the dependent variable, the conclusions of the analysis remain the same, as long as the imputation model does not perfectly predict

the omitted regressor. Curiously, if the regression model only has a single regressor and the conditioning information is the same, the bias from SDL is identical whether the SDL is applied to the regressor or to the dependent variable. If there are multiple regressors, with SDL applied to a single regressor, the SDL introduces bias in all regressors. The model setup and nature of the bias are derived explicitly in the online appendix, section C.2.

### III.C.  Estimating Regression Discontinuity Models

Regression discontinuity (RD) and regression kink (RK) models can be seriously compromised when SDL has been applied to the running variable. To illustrate some of these issues, we consider a design from Guido Imbens and Thomas Lemieux (2008). This analysis is intended to guide economists, who can perform our simplified SDL-aware analysis as part of the specification testing for a general RD.

MODEL SETUP Modeling the unobservable latent outcomes is intrinsic to the RD analysis. We incorporate the usual counterfactual data process inherent in the RD design directly into the data model. As Imbens and Lemieux (2008) note, this is a Rubin Causal Model (Rubin 1974; Holland 1986; Imbens and Rubin 2015). The simplest data model, corresponding to Imbens and Lemieux (2008, pp. 616–19), has three continuous variables and one discrete variable whose conditional distribution is degenerate in the RD design and nondegenerate in the fuzzy RD (FRD) design. The latent data process consists of four variables with the following definitions: $w_i(0)$ = untreated outcome, $w_i(1)$ = treated outcome, $t_i$ = treatment indicator, and $r_i$ = RD running variable. The confidential data vector has the experimental design structure, $Y = (w_i^*, t_i, r_i)$ where $w_i^* = w_i(t_i)$.

Our interest centers on the conditional expectations in the population data model $E[w_i(0)|r_i] = f_1(r_i)$ and $E[w_i(1)|r_i] = f_2(r_i)$, where $f_1(r_i)$ and $f_2(r_i)$ are continuous functions of the running variable, $r_i$. The parameter of interest is the average treatment effect at $\tau$:

$$\theta_{RD} = \lim_{r_i \downarrow \tau} E[w_i(1)|r_i = \tau] - \lim_{r_i \uparrow \tau} E[w_i(0)|r_i = \tau]$$

$$= \lim_{r_i \downarrow \tau} f_2(r_i) - \lim_{r_i \uparrow \tau} f_1(r_i).$$

NONIGNORABLE SDL IN THE RUNNING VARIABLE We focus on the setting where SDL is only applied to the RD running variable and its associated indicator. The published data vector is $Z = (w_i^*, t_i, z_i)$. The published running variable is sampled from a distribution that depends on the true value: $z_i \sim p_{Z|R}(z_i|r_i)$. We assume the distribution $p_{Z|R}(z_i|r_i)$ is the randomized

response mixture model, a generalization of simple randomized response described in the online appendix, section D.1. The SDL process depends on two parameters: $\rho$, the probability that the confidential value of the running variable is released without added noise, and $\delta$, the standard deviation of a mean zero noise term added to the running variable when subjected to SDL.

If the agency publishes its SDL values $\rho = \rho_0$ and $\delta = \delta_0$ and the true RD is strict, then the analyst can correct the strict RD estimator directly using

$$
(1) \qquad \hat{\theta}_{SRD} = \frac{\lim_{z_i \downarrow \tau} \hat{f}_2(z_i) - \lim_{z_i \uparrow \tau} \hat{f}_1(z_i)}{\rho_0}.
$$

Clearly, this implies that the uncorrected estimate is attenuated toward zero. Intuitively, the introduction of noise into the running variable converts the strict RD to a fuzzy RD, with $E[t_i|z_i, \rho_0, \delta_0]$ playing the role of the "compliance status" function. For details, see the online appendix, section D.2.

When the true RD is strict, the SDL is discoverable from the compliance function even if the agency has not released the SDL parameters. The researcher can use the fact that the compliance function $g(z_i) = \rho \mathbf{1}[z_i \geq \tau] + (1 - \rho)\Phi\left(\dfrac{z_i - \tau}{\delta}\right)$. The fuzzy RD estimator is

$$
\hat{\theta}_{FRD} = \frac{\lim_{z_i \downarrow \tau} \hat{f}_2(z_i) - \lim_{z_i \uparrow \tau} \hat{f}_1(z_i)}{\lim_{z_i \downarrow \tau} \hat{g}(z_i) - \lim_{z_i \uparrow \tau} \hat{g}(z_i)}.
$$

When the noise addition is independent of the outcome variables (as is the case here), the change in the probability of treatment at the discontinuity point, $\tau$, is equal to the share of undistorted observations, $\rho_0$. When $\rho = 1$, there has been no SDL, and both estimators yield the conventional sharp RD estimate. A similar analysis shows that a sharp RK design becomes a fuzzy RK design (Card and others 2012) in the presence of SDL. As in the case of linear regression, it is still necessary to model the extra variability from the SDL to get correct estimates of the variance of the estimated RD parameter.

IMPLICATIONS OF SDL IN THE RUNNING VARIABLE FOR FUZZY RD MODELS  If generalized randomized-response SDL is applied to the running variable, then the SDL is ignorable for parameter estimation when using a fuzzy RD design. The FRD compliance function must be augmented with the

contribution from SDL. When the running variable is distorted with normally distributed noise, as we have assumed, there is no point mass anywhere, and hence no discontinuity in the probability of treatment at the discontinuity that is due to the SDL. The claim that the SDL is ignorable for estimation of the treatment effect in the fuzzy RD design follows because the only discontinuity in the estimated compliance function is entirely due to the discontinuity in the true running variable. (See the online appendix, section D.2.1, for details.) Imbens and Lemieux (2008) show that the instrumental variable (IV) estimator that uses the RD as an exclusion restriction is formally equivalent to the fuzzy RD estimator, so the SDL is also ignorable for consistent estimation in this case as well.

Whether or not the SDL is ignorable for consistent estimation, it is never ignorable for inference. The estimated standard errors of the RD and FRD treatment effects must be adjusted.

In some applications, the treatment indicator is not observed and must be proxied by the discontinuity point, around which the RD is strict. If the treatment indicator is not observed and SDL has been applied to the running variable, only the sharp RD estimator is available, and it will be attenuated by a factor $\rho$. Nothing can be done in this setting without auxiliary information about the SDL model.

NONIGNORABLE SDL IN OTHER PARTS OF THE RD DESIGN  When SDL is applied to the dependent variable rather than the running variable, the situation is more complicated. We refer to our analysis of regression models in section III.B. SDL applied to the dependent variable will lead to attenuation of the estimated treatment effect unless all relevant variables, including the running variable and its interaction with the discontinuity point, are included in the SDL model for the dependent variable. Hence, SDL applied to the dependent variable is more likely to cause problems for RD than for conventional linear regression models, since the variation around the discontinuity point is unlikely to be included in the agency's imputation or swapping algorithms.

CONSEQUENCES OF DATA COARSENING FOR SDL  The ignorability of SDL in some circumstances was anticipated in the work of Daniel Heitjan and Rubin (1991), which considers the problem of inference when the published data are coarsened. Their application was to reporting errors where, for instance, individuals round their hours to salient, whole numbers. The same model is relevant to those types of microdata SDL that aggregate attribute categories, like occupations or geographies, and to topcoding.

David Lee and David Card (2008) consider the consequences of microdata coarsening for RD designs. For example, if ages are coarsened into years, the

RD design in which age is the running variable will group observations near the boundary with those further from the boundary, violating the required assumption that the running variable is continuous around the treatment threshold. Once again, depending on the type of RD design, when SDL is accomplished through coarsening of the running variable, it is not ignorable. An analysis that uses the coarsened running variable with a standard RD estimator may be biased and understate standard errors. As in Heitjan and Rubin (1991), Lee and Card (2008) establish conditions under which a grouped-data estimator provides a valid way to handle coarsened data. This method is agnostic about the cause of the grouping and is therefore SDL-aware by construction.

### III.D.  Estimating Instrumental Variable Models

We consider simple instrumental variable models with a single endogenous explanatory variable, a single instrument, and no additional regressors. Except where indicated, the intuition for these examples carries through to a more general setting with multiple instruments and controls.

The confidential data model of interest is the standard IV system

$$y_i = \kappa + \gamma t_i + \varepsilon_i$$

$$t_i = \phi + \delta z_i + \eta_i$$

where $y_i$ is the outcome of interest, $t_i$ is a scalar variable that may be correlated with the structural residual $\varepsilon_i$, and $z_i$ is a scalar variable that can serve as an instrument. That is, $z_i$ is uncorrelated with $\varepsilon_i$ and $\delta \neq 0$. We assume the SDL described in section III.B is applied to either the dependent variable, the endogenous regressor, or the instrument.

With this simplified setup, the IV estimator $\hat{\gamma}_{IV} = \hat{\beta}_{RF}/\hat{\delta}$, where $\hat{\beta}_{RF}$, is the parameter estimate from the reduced form equation $y_i = \alpha + \beta z_i + v_i$. We apply the results in section III.B. First, if SDL is applied to the dependent variable, then the point estimate of $\gamma$ will be attenuated. This is an immediate consequence of the fact that plim $\hat{\beta} \leq \beta$, while plim $\hat{\delta} = \delta$. Second, by parallel reasoning, if SDL is applied to the endogenous regressor, then the point estimate of $\gamma$ will be exaggerated. In this case, plim $\hat{\beta} = \beta$, but plim $\hat{\delta} \leq \delta$. This result implies that IV models may overstate the coefficient of interest when SDL is applied to the endogenous regressor. It is also not possible to use IV to correct for SDL in this case.

Finally, somewhat surprisingly, SDL is ignorable when applied to the instrument. In this particular model, with a single instrument and no regressors, the attenuation term is the same in the first-stage and reduced form,

and therefore cancels out of the ratio $\hat{\beta}_{RF}/\hat{\delta}$. We caution, however, that this ignorability does not extend to the case where there are additional exogenous regressors. In summary, our analysis suggests that blank-and-impute SDL is generally nonignorable for instrumental variables estimation and inference.

## IV. Analysis of Official Tables

Tabular or aggregate data are the primary public output of most official statistical systems. Most agencies offer a technical manual that provides an extensive description of how the microdata inputs were transformed into the publication tables. These manuals rarely, if ever, include an assessment of the effects of the SDL, and we could find no examples of manuals that did among the federal statistical agencies. When an agency releases measures of precision for aggregate data, these measures do not include variation due to SDL.

There are three key forms of SDL applied to tabular summaries. All federal agencies rely on primary and complementary suppression as the main SDL method. When an alternative SDL method is used, the most common ones add noise to the underlying input microdata or to the prerelease tabulated estimates. For household-based inputs, most agencies also perform some form of swapping before preparing tabular summaries. For business-based inputs, we are not aware of any SDL system that uses swapping.

### IV.A. Directly Tabulating Published Microdata

An alternative to using published tabulations is to tabulate from published microdata files. This is usually not an option for business data, which form the bulk of our examples in this section, but it may be an option for household data. We explore some of the pitfalls of doing custom tabulations in the online appendix, section E.3. Researchers should use caution when making tabulations from published microdata if the subpopulations being studied are often suppressed in the official tables. The presence of suppression usually signals a data quality problem.

### IV.B. Suppression versus Noise Infusion

WHEN SUPPRESSION IS NONIGNORABLE  Tabular suppression rules identify cells that are too heavily influenced by a few observations. The consequences for research are profound when those few observations are the focus of a particular study or the cause of a very inconvenient complementary suppression. It is not surprising that detailed data about the upper

0.25 percent of the income distribution are almost all suppressed by the Statistics of Income Division of the IRS. If a study focuses on unusual subpopulations, dealing with suppression is a normal part of the research design.

The most common form of suppression bias occurs when an analyst is assembling data at a given aggregation level, such as county level by four-digit NAICS[6] industry group from the BLS's Census of Employment and Wages frame. Between 60 and 80 percent of the published cells will have missing data. These data cannot reasonably be missing at random (ignorably missing) because the rule used to determine if those data could be published depends upon the values of the missing data. The problem compounds as covariates from other sources are added to the analysis.

Formally, SDL suppression is never ignorable. The probability that a cell is suppressed depends on the values of its component microdata records. Surprisingly, there is considerable resistance to replacing suppression with SDL methods that infuse deliberate noise. Noise-infusion SDL, as applied in the QWI, allows for the elimination of cell suppression and therefore eliminates bias from missing data. The trade-off is an increase in variance of all table entries, including those that would not be suppressed.

Perhaps the resistance to replacing suppression with noise-infusion arises because the bias from suppression is buried in a missing-data problem that most applied studies address with ad hoc methods: (i) analyze the published data as though the suppressions were ignorable, or (ii) do the analysis at a more aggregated level (say, NAICS subsector rather than NAICS industry group). These approaches are generally not as good as what could be accomplished with the same data if the cause were acknowledged and addressed.

A better solution, which is still ad hoc, is to use the frame variable to allocate the values of higher-level aggregates into the missing lower-level observations for the same variable. For example, in the QWI the frame variable is quarterly payroll—it is never suppressed at any level of aggregation—and in the QCEW and CBP the frame variable is the number of establishments, which is also never suppressed in these publications. The analyst can proportionally allocate the three-digit industrial aggregate employment, say, using the four-digit proportions of the frame

6. North American Industry Classification System.

variable as weights. This can be done in a sophisticated manner so that none of the observed original data are overwritten or contradicted by this imputation. For example, it can be done by only imputing the values of the four-digit employment that were actually suppressed and respecting the published three-digit employment totals for the sum of all four-digit industries within that total. This solution at least acknowledges that the suppression bias is nonignorable. The values for the higher-level aggregates contain some information about the suppressed values. Allocations based on the frame variable assume that the distribution of every variable with missing data across the entire population is the same as the distribution of the frame variable.

The analyst can do better still. The best solution for any given analysis is to combine the model of interest with a model for the suppressed data. Bayesian hierarchical models, like the ones we used in this paper, work well. Software tools for specifying and implementing such models are readily available. The complete model will properly account for the nonrandom pattern of the missing data, will incorporate prior information about the suppression rule that can be used for identification, and account for the additional uncertainty introduced by suppression. See Scott Holan and others (2010) for a specific application to BLS data.

WHEN NOISE INFUSION MAKES THE SDL NONIGNORABLE  Applying SDL by input noise infusion dramatically reduces the amount of suppression in the publication data. Since we are going to illustrate many of the features of these systems in the example in section V, we devote our attention here to the basic nonignorable features of input noise infusion.

Input noise infusion models were first proposed by Timothy Evans, Laura Zayatz, and John Slanta (1998). The noise models they proposed are constructed so that the expectation of the noisy aggregate, given the confidential aggregate, equals the confidential aggregate. This is the sense in which these measures are unbiased. In addition, as the number of entities in a cell (usually business establishments) gets large, the variance of the aggregate that is due to noise infusion vanishes. This is the sense in which these measures add variance to the published data in exchange for reducing suppression bias. Finally, the noise itself is usually generated from an independent, identically distributed random variable, so the joint distribution of the confidential data and the input noise factors into two independent distributions. Thus, SDL using input noise infusion can sometimes be ignorable for estimating the parameter of interest, but it will generally not be ignorable when trying to form a confidence interval around that estimate.

Because the noise process affects the posterior distribution of most parameters of interest, it is generally not ignorable.

Fortunately, agencies have been much more open about the processes used to produce publication tables from noise-infused inputs. A data-quality variable generally indicates whether the published value suffers from substantial infused noise. These flags are based on the absolute percentage error in the published value compared to the confidential value. It turns out, as we will see below, that they also sometimes release enough information to estimate the variance of the noise process itself, which is the SDL parameter that plays the role of the randomized-response "true data" probability. When the variance of the noise-infusion process goes to zero, the SDL becomes ignorable for all analyses, if no other SDL replaces it.

## V.  SDL Discovery in Published Tables

In this section, we show that it is possible to use information from three data sets released from very similar frames to conduct complete SDL-aware analyses. These data sets are the QWI, the QCEW, and the CBP. The key insight is that each data set applies a different SDL method to the same confidential microdata. The variation across the published data facilitates discovery of the SDL process. First, it is possible to directly infer a key unpublished variance term from the QWI noise infusion model. This variance term can then be used to correct SDL-generated estimation bias. Second, we argue that the QCEW and CBP data can be used as instruments to correct SDL-induced measurement error in analysis based on the QWI.

### V.A.  Overview of the QWI, QCEW, and CBP Data Sets

The QWI is a collection of 32 employment and earnings statistics produced by the Longitudinal Employer-Household Dynamics program at the U.S. Census Bureau. It is based on state Unemployment Insurance (UI) system records integrated with information on worker and workplace characteristics. Workplace characteristics are linked from the QCEW microdata. The frame for employers and workplaces is the universe of QCEW records, including both the employer report and the separate workplace reports. A QCEW workplace is an establishment in the QWI data. Essentially, the same QCEW inputs are used by the BLS to publish its Census of Employment and Wages (CEW) quarterly series on employment and total payroll. (In what follows, the acronym QCEW is reserved for the

inputs and publications of the BLS in the CEW series.) CBP data sets are also published by the Census Bureau from inputs based on its employer Business Register.

While the QWI, QCEW, and CBP use closely related sources to publish statistics by employer characteristics, they apply different methods for SDL. The QWI and CBP distort the establishment-level microdata using a multiplicative noise model and publish the aggregated totals. The QCEW aggregates the undistorted confidential establishment-level microdata and then suppresses sensitive cells with enough complementary suppressions of nonsensitive cells to allow publication of most table margins.

### V.B. Published Aggregates from the QWI, QCEW, and CBP

We give just enough detail here so that the reader can see how the Census Bureau and BLS form the aggregates for the quarterly payroll variables that we will use to illustrate the consequences of universal noise infusion for SDL. (More details are in the online appendix, section F.)

Tabular aggregates are formed over a classification $k = 1, \ldots, K$ that partitions the universe of establishments into $K$ mutually exclusive and exhaustive cells $\Omega_{(k)t}$. These partitions have detailed geographic and industrial dimensions. For all three data sources, geography is coded using FIPS[7] county codes. Industrial classifications are NAICS sectors, subsectors, and industry groups. The tabular magnitudes are computed by aggregating the values over the establishments in the group $k$. For the QWI, in the absence of SDL, the total quarterly payroll $W_{jt}$ for establishment $j$ in group $k$ and quarter $t$ would be estimated by[8]

$$(2) \qquad\qquad W_{(k)t} = \sum_{j \in \Omega_{(k)t}} W_{jt}.$$

For the QCEW, an identical formula uses total quarterly payroll, as measured by $W_{jt}^{(QCEW)}$ and for CBP, the quarterly payroll variable would be $W_{jt}^{(CBP)}$. Published aggregates from the QWI are computed using multiplicative noise factors $\delta_j$ that have mean zero and constant variance. (More details are in the online appendix, section G.) The published quarterly payroll is computed as

$$(3) \qquad\qquad W_{(k)t}^{*} = \sum_{j \in \Omega_{(k)t}} \delta_j W_{jt},$$

---

7. Federal Information Processing Standard.
8. We abstract from the weight that QWI uses to benchmark certain state-level aggregates. Formulas including weights are in the online appendix, section H.

where we have adopted the convention of tagging the post-SDL value with an asterisk. The same noise factor is used to aggregate total quarterly payroll and all other QWI variables. Total quarterly payroll is never suppressed in the QWI. The number of establishments in a cell is not published. If, and only if, a cell has a published value of $W^*$, then there is at least one establishment in that cell.

The published QCEW payroll aggregate is exactly the output of equation 2 using QCEW inputs. The published QCEW total quarterly payroll might be missing due to suppression. The QCEW data use item-specific suppression. Payroll might be suppressed when employment is not, and vice versa.

The CBP total quarterly payroll is exactly the output of equation 3 with CBP-specific inputs, including the noise factor. As with the QWI data, the same noise factor is used for all the input variables from a particular establishment. The published CBP aggregates have some SDL suppressions and can therefore be missing. The number of establishments in a cell is never suppressed, nor is the size distribution of employers.

### V.C. Regression Models with Nonignorable SDL

The noise infusion in QWI may be nonignorable. Univariate regression of a variable from another data set onto a QWI aggregate provides a simple illustration, which we summarize here. (See the online appendix, section E.4, for details.)

The model of interest is appendix equation E.26, the regression of a county-level outcome $Y_{(k)t}$ from a non-QWI source on QWI quarterly payroll in the county $W^*$. The dependent variable can be subjected to SDL as long as it is independent of the QWI SDL, as would be the case if the dependent variable were computed by the BLS or the Bureau of Economic Analysis (BEA). The published aggregate data are the $[Y_{(k)t}, W^*_{(k)t}]$. The undistorted values, $W_{(k)t}$, are confidential.

The probability limit of the ordinary least squares (OLS) estimator for the regression coefficient on $\beta$ based on using the published data is appendix equation E.27, and the asymptotic bias ratio is appendix equation E.28. The bias due to SDL depends on the product of two factors: the variance of the noise-infusion process and the expected Herfindahl index for payroll within aggregate $k$, as derived in the online appendix, section E.5. If either of these factors is zero, there is no bias in estimation. But the expected Herfindahl index is data, so we cannot make prior restrictions on that component. This leaves only the SDL noise variance. Clearly, the noise infusion is nonignorable in this setting.

One option is to correct the bias analytically. If the noise variance is known or can be estimated, the bias can be corrected directly. An unbiased estimator for $E[W_{(k)t}]^2$ is available from $E[W^*_{(k)t}]^2$ once the variance of the multiplicative noise factor, $V[\delta_j]$, is known, after which it only remains to recover $V[W_{(k)t}]$ from the definition of $V[W^*_{(k)t}]$.

The second possibility is to find instruments. Any instrument, $Z_{(k)t}$, correlated with $W_{(k)t}$ and uncorrelated with the SDL noise infusion process, will work, as shown in appendix equation E.29. In the QWI setting, there are three natural candidates for such instruments: (i) data from the QCEW for the same cell; (ii) data from CBP from the same cell; and (iii) data from neighboring cells (geographies or industries) in the QWI.

Data from QCEW for the same cell are based on the same administrative record system. QWI tabulates its measures from the UI wage records. QCEW tabulates from the associated ES-202 workplace report. The total payroll measure has an identical statutory definition on both administrative record systems for the state's Unemployment Insurance. Data for CBP are tabulated from the Census Bureau's employer Business Register. Payroll and employment come from the employer federal tax filings, and the payroll measured from this IRS source has a very similar statutory definition as compared to the definition used by QWI and QCEW. Finally, QWI data from nearby geographies or industries (depending on the aggregate represented by $k$) should be correlated with the QWI variable in the regression because they are based on the same administrative record system reports.

By construction, all of these instruments are uncorrelated with the SDL-induced noise in the right-hand side of equation E.26. In the case of QCEW or CBP data, any SDL-induced noise (CBP) or suppression bias (QCEW and CBP) in the instrument is independent of the noise in QWI. However, if many of the cells in the tabulation of the instrument are suppressed, that will affect the validity of the instrument, as we analyzed in section IV.B. When there are many suppressions in QCEW or CBP for the partition under study, data from the neighboring QWI cells can be used to complete the set of instruments.

Perhaps surprisingly, the input noise infusion to the QWI does not bias parameter estimates if the dependent and independent variables all come from QWI. Once drawn, the establishment-level noise factors are the same across variables and over time. Therefore, the variance from noise infusion affects all variables in exactly the same manner, factors out of the OLS moment equations, and then cancels. The same feature of the QWI also leads the time-series properties of the data to be preserved after noise infusion. We note that this feature is unique to the QWI method of noise

infusion, where the noise process is fixed over time for each cross-sectional unit. It does not hold for other forms of noise infusion, such as the one used by CBP.

### V.D.  Estimating the Variance Contribution of SDL for the QWI

It is possible to recover the variance of the noise factor $V[\delta_j]$, which is needed to correct directly for bias in the univariate and multivariate regression examples using the QWI. The details of this estimation process are presented in the online appendix, section E.5.

Our leverage in this analysis comes from the fact that QWI and QCEW use identical frames (QCEW establishments). Hence, we can use $W_{(k)t}^{(QCEW)}$ as the instrument for $W_{(k)t}$, as long as it has not been suppressed too often. Furthermore, we can use $W_{(k)t}^{(QCEW)}$, which is published at the county level as an instrument for any subcategory of QWI payroll, for example payroll of females ages 55–64, even though no exact analogue is published in QCEW.

Although the data come from a different administrative record system, the concepts underlying the CBP payroll variable are very similar to both the QWI and QCEW inputs. The SDL system used for CBP data is very similar to the one used for QWI, but the random noise in CBP is independent of the random noise in QWI. Therefore, CBP data can also be used as instruments, and they are suppressed far less often than QCEW data. The formulas for recovering both systems' SDL parameters are in the online appendix, section E.5.

### V.E.  Empirical Results

Table 1 presents the estimates of the equation used to recover the SDL parameters fitted using matched QWI and QCEW data for the first quarters of 2006 through 2011 by ordinary least squares. Table 2 fits the same functions using mixed-effect models.[9] The equations are fitted for state-level aggregations, where the error in both the employment and payroll magnitudes is mitigated by the benchmarking, county-level aggregations, where the agreement in the workplace codes for county is most likely to be strong, and county by NAICS sector-level aggregations, where there is greater scope for differences between the coding of the microdata in QWI and QCEW.

Both tables give very similar estimates for $V[\delta]$ whether we use payroll or employment as the basis. This suggests that the bias in estimating $V[\delta]$

---

9.  By the construction of the noise-infusion process for QWI, the design of the random effects is orthogonal to $\ln N_{(k)t}$.

**Table 1.** Estimated Variance of QWI Establishment Noise Factor (δ)[a]

| | County-sector[b] | | County | | State | |
|---|---|---|---|---|---|---|
| | Employment (ln) (1) | Payroll (ln) (2) | Employment (ln) (3) | Payroll (ln) (4) | Employment (ln) (5) | Payroll (ln) (6) |
| Number of establishments (ln) | -0.281 | -0.211 | -0.209 | -0.155 | -0.144 | 0.205 |
| | (.0016) | (.0017) | (.0061) | (.0062) | (.0679) | (.0987) |
| Constant | -1.527 | -1.610 | -2.027 | -1.962 | -4.679 | -6.747 |
| | (.0153) | (.0070) | (.0408) | (.0422) | (.7885) | (1.159) |
| No. of observations | 228,770 | 236,925 | 18,000 | 18,057 | 282 | 282 |
| $R^2$ | 0.1246 | 0.0582 | 0.0604 | 0.0324 | 0.0138 | 0.0196 |
| Var. fuzz ($V[\delta]$) | 0.046 | 0.040 | 0.017 | 0.020 | 0.0001 | 0.000 |

Source: QCEW and QWI data for Q1 for years 2006–11.

a. Each column reports estimates of a bivariate regression of the log coefficient of variation between QCEW and QWI employment (payroll) onto the natural logarithm of the number of establishments (reported in QCEW). The variance of the QWI noise factor is estimated as $V[\delta] = \exp(-2 \times Constant)$.

b. Estimates from data disaggregated by county and NAICS major sector.

**Table 2.** Estimated Variance of QWI Establishment Noise Factor (δ)—Mixed Models[a]

| | County-sector | | County | | State | |
|---|---|---|---|---|---|---|
| | Employment (ln) (1) | Payroll (ln) (2) | Employment (ln) (3) | Payroll (ln) (4) | Employment (ln) (5) | Payroll (ln) (6) |
| Number of establishments (ln) | −0.321 | −0.219 | −0.224 | −0.153 | −0.164 | 0.254 |
| | (.0035) | (.0030) | (.0166) | (.0117) | (.1467) | (.1688) |
| Constant | −1.405 | −1.578 | −1.971 | −1.979 | −4.455 | −7.270 |
| | (.0126) | (.0119) | (.1179) | (.0786) | (1.692) | (1.961) |
| No. of observations | 228,770 | 236,925 | 18,000 | 18,057 | 282 | 282 |
| Var. fuzz (V[δ]) | 0.060 | 0.043 | 0.019 | 0.019 | 0.0001 | 0.000 |

Source: QCEW and QWI data for Q1 for years 2006–11.

a. Each column reports estimates of a mixed-effects model of the log coefficient of variation between QCEW and QWI employment (payroll) that includes fixed effects for the natural logarithm of the number of establishments (reported in QCEW) and random slopes and intercepts at the county-sector, county, and state level, respectively.

from using proxies for the Herfindahl index is either minimal or uncorrelated between employment and payroll. Either way, we are able to estimate with reasonable precision the range of possibilities for $V[\delta]$, and these indicate that the noise infusion does not create a very substantial bias or inflate estimated variances substantially.

## VI. The Frontiers of SDL

In this section we discuss the relationship between synthetic data and validation servers, the nature and limits of formal privacy systems, and the analysis of confidential data in enclaves.

### VI.A. Analysis of Synthetic Data

We defined synthetic data in section II. Here we discuss the tight relationship between synthetic data systems and validation servers, a method of improving the accuracy of synthetic data that links the user community and the data providers directly. In a synthetic data feedback loop, the agency releases synthetic microdata to the research community. Researchers analyze the synthetic data as if they were public-use versions of the confidential data using SDL-aware analysis software. When the analysis of the synthetic data is complete, the researchers may request a validation, which is performed by the data providers on the actual confidential data. The results of the validation are subjected to conventional SDL and then released to the researcher as public-use data. The data provider then inventories these analyses and uses them to improve the analytical validity of the synthetic data in the next release by testing new versions of the synthetic data on the models in its inventory.

The Census Bureau has two active feedback-loop, synthetic-data systems: the Survey of Income and Program Participation (SIPP) Synthetic Beta (SSB) and the Synthetic Longitudinal Business Database (SynLBD).[10] The SSB provides synthetic data for all panels of the SIPP linked to longitudinal W-2 data. SynLBD is a synthetic version of selected variables and all observations from the confidential Longitudinal Business Database, the research version of the employer Business Register, longitudinally linked.

A recent paper by Marianne Bertrand, Emir Kamenica, and Jessica Pan (2015) provides an excellent illustration of the advantages of using

---

10. Information about the SIPP database can be found here: https://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file. Information about the SynLBD database can be found here: https://www2.vrdc.cornell.edu/news/data/lbd-synthetic-data/

synthetic data that are part of a feedback loop. The authors use the administrative record values for married couples' individual W-2 earnings to compute the proportion of household income that was due to each partner. They hypothesize that there should be a regression discontinuity at 50 percent because of their model prediction that women should prefer to marry men with higher incomes than their own. The SSB data have undergone extensive SDL and, for this model, the effects of this SDL on the RD running variable was extensive, nonignorable, and had a stated "suppress and impute rate" of 100 percent. Analyses from synthetic data show no causal effect. However, analyses from the validation estimation on the confidential data, where the earnings variables have not been subjected to any SDL but are imputed when missing, show a clear discontinuity. The validated estimates are reported in the published paper. Any researcher anywhere in the world can use the SSB and SynLBD by following the instructions on the Cornell University-based server that is used as the interface for analyses that are part of the feedback process.[11]

While writing this paper, we discovered why the analysis of the linked SIPP-IRS data by Bertrand, Kamenica, and Pan (2015) showed no causal effect when the synthetic data were used. The reason can be seen by examining equation 1 when the running variable has been modified for every observation, as is the case in the SSB. The regression-discontinuity effect is not identified in the synthetic data, and it will not generally be identified for any RD design that uses the many exact earnings and date variables in the SSB. If only the SSB were available with no access to validation, RD and FRD analyses using these data would be pointless. However, because the SSB offers validation using the underlying confidential data and traditional SDL on the output coefficients, an analyst can do a specification search for the response functions $f_1$ and $f_2$ using the SSB, then submit the entire protocol from the specification search for validation. The validated estimate of the RD or FRD treatment effect provides the researcher's first evidence on that effect. Thus, the use of the feedback mechanism for the synthetic data protected the research design from pretest estimation and false-discovery bias for the inferences on the causal RD effect, an incredible silver lining.

We have already noted that the Survey of Consumer Finances (SCF) uses synthetic data for SDL, based on the same model that is used for edit and imputation of item missing data. The statutory custodian for the SCF is the Federal Reserve Board of Governors. The Fed maintains a very limited

11. The Cornell-based server is located here: http://www2.vrdc.cornell.edu/news/synthetic-data-server/step-1-requesting-access-to-sds/

feedback loop that is described in the codebook (Federal Reserve Board of Governors 2013).

### VI.B.  Formal Privacy Systems

A researcher is much more likely to encounter a formal privacy system for SDL when interacting with a private data provider. Differential privacy was invented at Microsoft. As early as 2009, Microsoft had in place a system, Privacy Integrated Queries (PINQ), that allowed researchers to analyze its internal data files (such as search logs) with a fixed privacy budget using only analysis tools that were differentially private at every step of the process, including data editing (McSherry 2009). These tools ensure that every statistic seen by the researcher, and therefore available for publication, satisfies $\varepsilon$-differential privacy. When the researcher exhausts $\varepsilon$, no further access to the data is provided.

PINQ computes contingency tables, linear regressions, classification models, and other statistical analyses using provably private algorithms. Its developer recognized that a strong privacy guarantee comes at the expense of substantial accuracy. It was up to the analyst to decide how to mitigate that loss of accuracy. The analyst could spend most of the privacy budget to get some very accurate statistics—ones for which the inferences were not substantially altered as compared to the same inference based on the confidential data. But then the analysis was over, and the analyst could not formulate follow-up hypotheses because there was no remaining privacy budget. Alternatively, the analyst could use only a small portion of the privacy budget doing many specification searches, each one of which was highly inaccurate as compared to the same estimation using the confidential data, then use the remainder of the privacy budget to compute an accurate statistic for the chosen specification.

The literature on formal privacy models is still primarily theoretical. At present, there are serious concerns about the computational feasibility of applying formal privacy methods to large, high-dimensional data, as well as their analytical validity for nontrivial research questions. However, these methods make clear the cost in terms of loss of accuracy that is inherent in protecting privacy by distorting the analysis of the confidential data. The formal methods also allow setting a privacy budget that can be allocated across competing uses of the same underlying data.

Economists should have no trouble thinking about how to spend a privacy budget optimally during a data analysis. But they might also wonder

how any real empirical analysis can survive the rigors of never seeing the actual data. That is a legitimate worry, and one that the formal privacy community takes very seriously. For a glimpse of one possible future, see the work of Dwork (2014), who calls for all custodians of private data to publish the rate at which their data publication activities generate privacy losses and to pay a fine for nonprivate uses (infinite privacy loss, $\varepsilon = \infty$). Public and private data providers will have an increasingly difficult time explaining why they are unwilling to comply with this call when others begin to do so. The resulting public policy debate is very unlikely to result in less SDL applied to the inputs or outputs of economic data analyses.

### VI.C.  Analysis of Confidential Data in Enclaves

Because this paper is about the analysis of public-use data when the publisher has used statistical disclosure limitation, we have not discussed restricted access to the underlying confidential data. Restricted access to the confidential data also involves SDL. First, some agencies do not remove all of the SDL from the confidential files they allow researchers to use in enclaves. Second, the output of the researcher's analysis of the confidential data is considered a custom tabulation from the agency's perspective. The output is subjected to the same SDL methods that any other custom tabulation would require.

## VII.  Discussion

Unlike many other aspects of the processes by which data are produced, SDL is poorly understood and seldom discussed among economists. SDL is applied widely to the data most commonly used by economists, and the pressure on data custodians to protect privacy will only get stronger with time. We offer suggestions to researchers, journal editors, and statistical agencies to facilitate and advance SDL-aware economic research.

### VII.A.  Suggestions for Researchers

Over the decades since SDL was invented, research methods have changed dramatically—most notably in the applied microeconomists' adoption of techniques that require both enormous amounts of data and very precise model-identifying information. The combination of these two requirements has led to much more extensive use of confidential data with the publication of only summary results. Studies carried out this way have very limited potential for replication or reuse of the confidential data. Grant funding agencies have insisted that the researchers they fund prepare a data

management plan for the curation of the data developed and analyzed using their funds, yet very few statistical agencies or private firms will surrender a copy of the confidential data for secure curation to allow research teams to comply with this requirement. Consequently, only the public portion of this scientific work can be curated and reused. But all such public data have been subjected to very substantial SDL, almost all of it in the form of suppression—none of the original confidential data and very little of the intermediate work product can be published.

Suppression on this scale leads to potentially massive biases and very limited data releases. To address this problem, over these same decades statisticians and computer scientists have worked to produce SDL methods that permit the publication of more data, including detailed microdata with large samples and precise model-identifying variables. Yet only a handful of applied economists are active in the SDL and data privacy communities. What Arthur Kennickell accomplished by integrating the editing, imputation, and SDL components of the Survey of Consumer Finances in 1995 and orchestrating the release of those microdata in a format that required SDL-aware analysis methods was not accomplished again until 2007, when the Census Bureau released synthetic microdata for the Survey of Income and Program Participation. We believe that the reason economists have been reticent about exploring alternatives to suppression is that they have not fully understood how pernicious suppression bias actually is.

Statistical agencies do understand this, and the SDL and privacy-preserving methods they have adopted are designed to control suppression bias by introducing some deliberate variance. Economists tend to argue that the deliberate infusion of unrelated noise is a form of measurement error that infects all of the analyses. That is true, as we have shown, but it is an incomplete picture. Suppression too creates massive amounts of unseen bias—the direct consequence of not being able to analyze the data that are not released. Economists should recognize that the publication of altered data with more limited suppression instead of just the unsuppressed unaltered data could be a technologically superior solution to the SDL problem. We challenge more economists to become directly involved in the creation and use of SDL and privacy-preserving methods that are more useful to the discipline than the ones developed to serve the general user communities of statistical agencies and Internet companies.

In the meantime, what can productively be done? Economic researchers who use anything other than the most aggregated data should become more familiar with the methods used to produce those data: population frames, sampling, edit, imputation, and publication formulas, in addition

to SDL. This will help reduce the tendency to think of SDL as the only source of bias and variation. For students, these topics are usually covered in courses called "Survey Methodology," but they belong in econometrics and economic measurement courses too.

### VII.B.  Suggestions for Journals, Editors and Referees

Journals should insist that authors document the entire production process for the inputs and output of their analyses. The current standards are incomplete because they focus on the reproducibility of the published results from uncurated inputs. Economists do not even have a standard for citing data. A proper data citation identifies the provenance of the exact file used as the starting point for the analysis. Requiring proper citation of curated data inputs provides an incentive for those who perform such activities, just as proper software citation has provided an incentive to create and maintain curated software distribution systems. Discussions of the consequences of frame definitions, sampling, edit, imputation, publication formulas, and SDL that were applied to the inputs are also important for any econometric analysis. If authors cannot cite sources that document each of these components, they should be required to include the information in an archival appendix.

We make these points because we also want the journals to require documentation of the SDL procedures that were applied to the inputs and outputs of the analyses, although we do not think it is appropriate to single out SDL for special attention. The other aspects of data publication we discuss here also have implications for interpreting and reproducing the published results. If scientific journals added their voices to the calls for better documentation of all data publication methods, it would be easier to press statistical agencies to release more details of their SDL methods.

### VII.C.  Suggestions for Statistical Agencies and Other Data Providers

We think that the analysis in this paper should be considered a prima facie case for releasing more information about the actual parameters used in SDL methods and for favoring SDL methods that are amenable to SDL-aware statistical analysis. By framing our arguments using methods already widely adopted to assess the effects of data quality issues, we hope to show that the users are also entitled to better information about specific SDL methods. We have also shown that if certain SDL methods are used, only very basic summary parameters need to be released. These can even be released as probability distributions, if desired.

We stress that we are not singling out SDL for special attention. Very specific information about the sample design is released in the form of the sampling frames used, detailed stratification structures, sampling rates, design weights, response rates, cluster information, replicate weights, and so on. Very specific information is released about items that have been edited, imputed or otherwise altered to address data quality concerns. But virtually nothing—nothing *specific*—is released about SDL parameters. This imbalance fuels the view that the SDL methods may have unduly influenced a particular analysis. In addition, it is critical to know which SDL methods have been permanently applied to the data, so that they must be considered even when restricted access is granted to the confidential data files.

Our remarks are not directed exclusively to government statistical agencies; they apply with equal force to Amazon, Facebook, Google, Microsoft, Netflix, Yahoo, and other Internet giants as they begin to release data products like Google Trends for use by the research community.

## VIII.  Conclusion

Although SDL is an important component of the data publication process, it need not be more mysterious or inherently problematic than other widely used and well understood methods for sampling, editing, and imputation, all of which affect the quality of analyses that economists perform on published data. Enough is known about current SDL methods to permit modeling their consequences for estimation of means, quantiles, proportions, moments, regression models, instrumental variables models, regression discontinuity designs, and regression kink models. We have defined ignorable SDL methods in a model-dependent manner that is exactly parallel to the way ignorability is defined for missing-data models. We have shown that an SDL process is ignorable if one can apply the methods that would be appropriate for the confidential data directly to the published data and reach the same conclusions.

Most SDL systems are not ignorable. This is hardly surprising, since the main justification for using SDL is limiting the ability of the analyst to draw conclusions about unusual data elements such as re-identifying a respondent or a sensitive attribute. The same tools that help assess the influence of experimental design and missing data on model conclusions can be used to make any data analysis SDL-aware. One such system, the multiple imputation model used for SDL by the Survey of Consumer Finances, has operated quite successfully for two decades. Other systems, most notably the synthetic data systems with feedback loops operated by the Census

Bureau, are quite new but permit fully SDL-aware analyses of important household and business microdata sources.

Finally, we have shown that the methods we developed here can be used effectively on real data and that the consequences of SDL for data analysis are limited, at least for the models we considered here. When methods that add noise are used, there is less bias than for equivalent analyses that use data subjected to suppression. The extra variability that the noise-infusion methods generate is of a manageable magnitude.

We use these findings to press for two actions: (i) publication of more SDL details by the statistical agencies so that it is easier to assess whether or not SDL matters in a particular analysis and (ii) less trepidation by our research colleagues in using data that have been published with extensive SDL. There is no reason to treat the use of SDL as significantly more challenging than the analysis of quasi-experimental data or an analysis with substantial nonignorable missing data.

# References

Abowd, John M., and Simon D. Woodcock. 2001. "Disclosure Limitation in Longitudinal Linked Data." In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies,* edited by Pat Doyle, Julia Lane, Jules Theeuwes, and Laura Zayatz. Amsterdam: North Holland.

Alexander, J. Trent, Michael Davern, and Betsey Stevenson. 2010. "Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications." *Public Opinion Quarterly* 74, no. 3: 551–69.

Anderson, Margo, and William Seltzer. 2007. "Challenges to the Confidentiality of U.S. Federal Statistics, 1910–1965." *Journal of Official Statistics* 23, no. 1: 1–34.

———. 2009. "Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues." *Journal of Privacy and Confidentiality* 1, no. 1: 7–52.

Benedetto, Gary, and Martha Stinson. 2015. "Disclosure Review Board Memo: Second Request for Release of SIPP Synthetic Beta Version 6.0." U.S. Census Bureau, Survey Improvement Research Branch, Social, Economic, and Housing Statistics Division (SEHSD). http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/DRBMemoTablesVersion2SSBv6_0.pdf

Bertrand, Marianne, Emir Kamenica, and Jessica Pan. 2015. "Gender Identity and Relative Income within Households." *Quarterly Journal of Economics* 130, no. 2: 571–614.

Bollinger, Christopher R., and Barry T. Hirsch. 2006. "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching." *Journal of Labor Economics* 24, no. 3: 483–520.

Burkhauser, Richard V., Shuaizhang Feng, Stephen P. Jenkins, and Jeff Larrimore. 2012. "Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data." *Review of Economics and Statistics* 94, no. 2: 371–88.

Card, David, David Lee, Zhuan Pei, and Andrea Weber. 2012. "Nonlinear Policy Rules and the Identification and Estimation of Causal Effects in a Generalized Regression Kink Design." Working Paper no. 18564. Cambridge, Mass.: National Bureau of Economic Research.

Dalenius, Tore. 1977. "Towards a Methodology for Statistical Disclosure Control." *Statistik Tidskrift* 15: 429–44.

Duncan, George T., Mark Elliot, and Juan-José Salazar-González. 2011. *Statistical Confidentiality: Principles and Practice.* New York: Springer.

Duncan, George T., and Stephen E. Fienberg. 1999. "Obtaining Information While Preserving Privacy: A Markov Perturbation Method for Tabular Data." Presented at Eurostat conference *Statistical Data Protection '98 (SDP'98).* Available at http://www.heinz.cmu.edu/research/21full.pdf

Duncan, George T., Thomas B. Jabine, and Virginia A. de Wolf, eds. 1993. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Washington: National Academies Press.

Duncan, George T., and Diane Lambert. 1986. "Disclosure-Limited Data Dissemination." *Journal of the American Statistical Association* 81, no. 393: 10–18.

Dwork, Cynthia. 2006. "Differential Privacy." In *Automata, Languages and Programming: 33rd International Colloquium, Proceedings, Part II,* edited by Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener. Berlin and Heidelberg: Springer.

———. 2014. "Differential Privacy: A Cryptographic Approach to Private Data Analysis." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement,* edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. Cambridge University Press.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography: Third Theory of Cryptography Conference, Proceedings,* edited by Shai Halevi and Tal Rabin. Berlin and Heidelberg: Springer.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends in Theoretical Computer Science* 9, nos. 3–4: 211–407.

Evans, Timothy, Laura Zayatz, and John Slanta. 1998. "Using Noise for Disclosure Limitation for Establishment Tabular Data." *Journal of Official Statistics* 14, no. 4: 537–51.

Evfimievski, Alexandre, Johannes Gehrke, and Ramakrishnan Srikant. 2003. "Limiting Privacy Breaches in Privacy Preserving Data Mining." In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS).* New York: Association for Computing Machinery. http://www.cs.cornell.edu/johannes/papers/2003/pods03-privacy.pdf

Federal Reserve Board of Governors. 2013. *Codebook for 2013 Survey of Consumer Finances.* Washington.

Fellegi, I. P. 1972. "On the Question of Statistical Confidentiality." *Journal of the American Statistical Association* 67, no. 337: 7–18.

Goldwasser, Shafi, and Silvio Micali. 1982. "Probabilistic Encryption and How to Play Mental Poker Keeping Secret All Partial Information." In *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing (STOC).* New York: Association for Computing Machinery. https://www.cs.purdue.edu/homes/ninghui/readings/Qual2/Goldwasser-Micali82.pdf

Hardt, Moritz, Katrina Ligett, and Frank McSherry. 2012. "A Simple and Practical Algorithm for Differentially Private Data Release." In *Advances in Neural Information Processing Systems* 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Red Hook, N.Y.: Curran Associates.

Harris-Kojetin, Brian A., Wendy L. Alvey, Lynda Carlson, Steven B. Cohen, and others. 2005. "Report on Statistical Disclosure Limitation Methodology." Statistical Policy Working Paper no. 22, Federal Committee on Statistical Methodology. https://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf

Heffetz, Ori, and Katrina Ligett. 2014. "Privacy and Data-Based Research." *Journal of Economic Perspectives* 28, no. 2: 75–98.

Heitjan, Daniel F., and Donald B. Rubin. 1991. "Ignorability and Coarse Data." *Annals of Statistics* 19, no. 4: 2244–53.

Hirsch, Barry T., and Edward J. Schumacher. 2004. "Match Bias in Wage Gap Estimates due to Earnings Imputation." *Journal of Labor Economics* 22, no. 3: 689–722.

Holan, Scott H., Daniell Toth, Marco A. R. Ferreira, and Alan F. Karr. 2010. "Bayesian Multiscale Multiple Imputation with Implications for Data Confidentiality." *Journal of the American Statistical Association* 105, no. 490: 564–77.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, no. 396: 945–60.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142, no. 2: 615–35.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction.* Cambridge University Press.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *American Statistician* 60, no. 3: 224–32.

Kennickell, Arthur B. 1997. "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances." In *Record Linkage Techniques,* edited by Wendy Alvey and Bettye Jamerson. Arlington, Va.: Federal Committee on Statistical Methodology.

Kennickell, Arthur, and Julia Lane. 2006. "Measuring the Impact of Data Protection Techniques on Data Utility: Evidence from the Survey of Consumer Finances." In *Privacy in Statistical Databases: CENEX-SDC Project International Conference, Proceedings,* edited by Josep Domingo-Ferrer and Luisa Franconi. Berlin and Heidelberg: Springer.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, and others. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." *International Statistical Review* 79, no. 3: 362–84.

Larrimore, Jeff, Richard V. Burkhauser, Shuaizhang Feng, and Laura Zayatz. 2008. "Consistent Cell Means for Topcoded Incomes in the Public Use March CPS (1976–2007)." *Journal of Economic and Social Measurement* 33, no. 2: 89–128.

Lauger, Amy, Billy Wisniewski, and Laura McKenna. 2014. "Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research." Research Report Series (Disclosure Avoidance) no. 2014-02. Washington: Center for Disclosure Avoidance Research, U.S. Census Bureau.

Lee, David S., and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142, no. 2: 655–74.

Little, Roderick J. A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9, no. 2: 407–26.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering.* Red Hook, N.Y.: Curran Associates.

McSherry, Frank 2009. "Privacy Integrated Queries: An Extensible Platform for Privacy Preserving Data Analysis." In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data.* New York: Association for

Computing Machinery. http://research.microsoft.com/pubs/80218/sigmod115-mcsherry.pdf

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-Anonymization of Large Sparse Datasets." In *Proceedings of the 2008 IEEE Symposium on Security and Privacy.* Red Hook, N.Y.: Curran Associates.

Ohm, Paul 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57: 1701.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118, no. 1: 1–41.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19, no. 1: 1–16.

Reiter, Jerome P. 2004. "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation." *Survey Methodology* 30, no. 2: 235–42.

———. 2005. "Estimating Risks of Identification Disclosure in Microdata." *Journal of the American Statistical Association* 100, no. 472: 1103–12.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66, no. 5: 688–701.

———. 1993. "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics* 9, no. 2: 461–68.

Skinner, C.J., and D.J. Holmes. 1998. "Estimating the Re-Identification Risk per Record in Microdata." *Journal of Official Statistics* 14, no. 4: 361–72.

Skinner, Chris, and Natalie Shlomo. 2008. "Assessing Identification Risk in Survey Microdata Using Log-Linear Models." *Journal of the American Statistical Association* 103, no. 483: 989–1001.

Sweeney, L. 2000. "Uniqueness of Simple Demographics in the U.S. Population." Technical report no. LIDAP-WP4. Laboratory for International Data Privacy, Carnegie Mellon University.

U.S. Census Bureau. 2013a. SIPP Synthetic Beta: Version 6.0 [computer file], Washington; Cornell University, Synthetic Data Server [distributor], Ithaca, N.Y.

U.S. Census Bureau. 2013b. Synthetic Longitudinal Business Database: Version 2.0 [computer file], Washington; Cornell University, Synthetic Data Server [distributor], Ithaca, N.Y.

U.S. Census Bureau. 2015. LEHD Origin-Destination Employment Statistics (LODES), Washington; U.S. Census Bureau [distributor].

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60, no. 309: 63–69.

Yakowitz, Jane. 2011. "Tragedy of the Data Commons." *Harvard Journal of Law and Technology* 25, no. 1.