ECONOMIC ANALYSIS AND STATISTICAL DISCLOSURE LIMITATION

John M. Abowd
Department of Economics
Labor Dynamics Institute/ILR
Cornell University
john.abowd@cornell.edu

Ian M. Schmutte
Department of Economics
Terry College of Business
University of Georgia
schmutte@uga.edu

August 11, 2015

REVISION DRAFT – PLEASE REQUEST FINAL VERSION FOR CITATION

We acknowledge direct support from Alfred P. Sloan Foundation Grant G-2015-13903 and, of course, the Brookings Institution. Abowd acknowledges direct support from NSF Grants BCS-0941226, TC-1012593 and SES-1131848. This paper was written while Abowd was visiting the Center for Labor Economics at UC-Berkeley. We are grateful for helpful comments from David Card, Cynthia Dwork, Caroline Hoxby, Tom Louis, Laura McKenna, David Romer, Betsey Stevenson, Lars Vilhuber, and Justin Wolfers. Appendices containing technical derivations, details of the data analysis and the replication archive are available at http://digitalcommons.ilr.cornell.edu/ldi/24/.

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.

1 Introduction

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 upon the specific hypotheses under study. In this paper, we describe how SDL actually works. We provide tools to evaluate the effects of SDL on economic modeling. We also provide 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 in Alexander, Davern and Stevenson (2010), ADS hereafter. ADS show that the percentage of men and women by age in public-use microdata samples 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 ADS, shortly after those announcements, 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

[1] See Technical Appendix Section B.1.

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 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 knowledge of these parameters, it is not possible to determine whether the magnitude of nonignorable SDL is substantial. As the analysis in ADS 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 5.2.2.

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.

Kennickell and 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 et al. 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 an awareness of how the data publication technology, including its SDL aspects, might affect our particular analyses. Furthermore, we should engage with data providers to help ensure 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.

2 Conceptual Framework and Motivating Examples

2.1 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 Technical 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 do correct inference 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. Implicitly assuming ignorable SDL is common in applied economic research, and our definition is an explicit extension of the related concept of *ignorable missing data*.

If the data analyst can't recover the estimate of interest without the parameters of the SDL model, then we have *nonignorable* SDL. In this case, the analyst needs to perform an *SDL-aware analysis*. However, it is only possible to perform an SDL-aware analysis if either (1) the data provider publishes sufficient details of the SDL model applied to the confidential data, or (2) 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*.

2.2 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 without revealing the true answer to that question 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 that the ignorability of SDL is not just a function of the SDL method, but also of the estimand of interest.

2.2.1 Randomized Response

Warner (1965) proposed a survey technique in which the respondent is presented with one of two questions that can both be answered "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," then destroys the question. With a certain probability, the question is sensitive (e.g., "Have you ever committed a violent crime?"), and, with complementary probability, the question is innocuous (e.g., "Is your birthday between July 1st and December 31st?"). The interviewer records only the "yes" or "no" answer, and never sees the true question.

If we run this single-question survey on a sample of 100 people chosen randomly, then the estimated proportion of "yes" answers has expected value equal to the probability that the respondent was asked the sensitive question times the population probability of committing 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%, then to recover the implied estimate for the population probability of having committed a violent crime, we need to know the probability that the sensitive question was asked. The standard error of the estimated proportion of "yes" answers is 4.4%, but the standard error for the estimated population proportion of having committed a violent crime is 4.4% divided by the probability that the respondent was asked the sensitive question.

Why is this 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 3.2.1 for a complete discussion. This form of SDL is obviously not ignorable. The data analyst doesn't care about the 26%. The analyst 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%, and its standard error is 8.8%. Notice that the SDL affected both the mean and the standard error of the estimate.

2.2.2 Consequences of Topcoding for Quantile Estimation

Burkhauser et al. (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 CPS. Because income is a sensitive data item, and large incomes can be particularly revealing in combination with other information, the Census Bureau and BLS censor incomes above a certain threshold in their public-use files. The topcoding of income protects privacy, but also limits what can be done with the data.

Burkhauser et al. (2012) report that the income topcode results in 4.6% 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 Piketty and Saez (2003) analyzed uncensored income data based on IRS tax filings. They showed that trends in income inequality look quite different in the administrative record data than in the CPS. Burkhauser et al. (2012) showed, using restricted-access CPS data, 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, then they would have probability distribution function $p_Y(Y)$ and cumulative distribution function $\mathrm{F}_Y[Y]$. For studying income inequality, interest centers on the quantiles of F_Y, defined by the inverse CDF function Q_Y. For the purpose of drawing inferences about the quantiles of the income distribution, topcoding is irrelevant for all quantiles that fall below the topcoding threshold, T. We say topcoding 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, assessment of the effect of SDL requires knowledge of the mechanism. If the value of the topcode threshold, T, were not published, then 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 inspection of the published data. In this case, we say the topcode is a *discoverable* form of SDL.

The work of Larrimore et al. (2008) also illustrates the potential for researchers, when armed with information about SDL methods and access to the confidential data, to improve analysis with minimal change to the risk of harmful or unlawful data disclosure. Larrimore et al. (2008) publish new data for 1976–2006 that contain the mean value of incomes above the topcode value within cells disaggregated by race, gender and employment status. They do so for 24 separate income series. They show that these cell means can be used with the public-use CPS microdata for analysis of the income distribution 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 by inspecting the data.

3 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 et al. 1993, p. 3). This principle embodies two distinct ideas. First, individuals have a property right of privacy covering their personal information. Second, once these 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 U.S., the Federal Committee on Statistical Methodology has produced a very thorough summary of the objectives and practices of SDL in Harris-Kojetin et al. (2005).

The constant evolution of information technology makes translating the principle of confidentiality into policy and practice challenging. The statutes that govern how statistical agencies approach SDL explicitly prohibit any breach of confidentiality.[2] However,

[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, cited 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

statisticians and computer scientists have formally proven 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.

3.1 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. Although 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), and one is hard-pressed to find a report of, say, the American Community Survey being "hacked," 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 2009; 2007). 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 anti-trust proceedings, or be used by their 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, Sweeney (2000) showed how then publicly-available hospital records might be linked to

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 cited August 6, 2015).

survey data to compromise confidentiality. Narayanan and Shmatikov (2008) showed that supposedly anonymous user data published by Netflix could be re-identified. Although no harm was documented in these cases, they highlight the potential for harm in the world of big data.

Ohm (2010) argues there may be, for every individual, 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 firms might do so, despite the absence of documented instances of such behavior. An alternative perspective is offered by 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 socio-economic data.

3.2 Concepts and Methods of SDL

Modern SDL methods are designed to allow publication of high-quality statistical information while protecting confidentiality. Many applied researchers may have an incomplete awareness of, and knowledge about, the ways in which SDL distorts published data. We provide a basic 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 books on SDL and formal privacy models (Duncan et al. 2011; Dwork and Roth 2014).

3.2.1 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 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 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 Dalenius (1977) and formalized by Duncan and Lambert (1986) in statistics, and by Goldwasser and 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 2.2.1 provides intuition about inferential disclosure. The probability that the respondent will answer "yes" given that the truth is "yes" is 75%. The probability that the respondent will answer "yes" given that the truth is "no" is 25%. 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 committing a violent crime. The ratio of these two probabilities is the Bayes factor–the ratio of the posterior odds that the truth is "yes" v. "no" given the survey answer "yes" to the prior odds of "yes" v. "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 interview 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.

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

3.3 SDL Methods for Microdata

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. 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 this case, knowledge of the suppression rule along with auxiliary information from the underlying microdata are extremely useful in assessing the effect of suppression on any specific application. Sometimes, suppression is combined with imputation. In that case, the sensitive information is suppressed, and then replaced with an imputed value.

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 in the data that have some combination of attributes 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 top-coding income, and reporting income in bins rather than in levels. These 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 those cases, one possible solution is to obtain access to the confidential, disaggregated data.

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 et al. 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 QWI, QCEW and CBP data in Section 5.2.2.

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 he participates in the ACS, and reports 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.

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. Agencies do not release many details of their swapping procedures. 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's why the agencies don't publish them.

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 affect data analyses that use those variables. The parameter confidentiality makes analyzing the effects of swapping difficult. 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 is the publication of a dataset with the same structure as the confidential data where 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 θ. 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.

[4] See Reiter (2005); Skinner and Holmes (1998); Skinner and Shlomo (2008) for specifics on the risk indices and Duncan et al. (2011, p. 114) for a review of historical uses of swapping.

When originally proposed by Little (1993) and 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 replace actual data–in exchange for an explicit benefit–correct estimation and inference procedures are available for the synthetic data. The Little and Rubin-style synthetic data analyses 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. Abowd and Woodcock (2001), Raghunathan et al. (2003) and Reiter (2004) refined the Rubin and Little methods, allowing them to be applied to complex survey data and combined with other missing data imputations. They also showed that the class of hypotheses with provable analytical validity was 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 interactions 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. In this case, the analyst requires access to the confidential data. When the model used to produce the synthetic data is publicly available, researchers can assess whether a given synthetic dataset 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 have these models run 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 7.1.

Formal privacy models emerged from database security and cryptography. The idea is to model the publication of data by the statistical agency via 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 et al. (2006) formalized the privacy protection associated with output-distortion SDL in a model called ε -differential privacy. For economists, Heffetz and Ligett (2014) is a very accessible introduction. Dwork and Roth (2014), in Chapter 3, use our running example of randomized response to characterize ε-differential privacy. In ε-differential privacy, the SDL must put an upper bound, ε, on the Bayes factor. In our example, $\varepsilon = \ln$ (Bayes factor bound) $= \ln 3 = 1.1$. Bounding the Bayes factor implies that the *maximum* amount the interviewer can learn from a "yes" answer is that the respondent is three times more likely than a random person in the population to have committed a violent crime.

Formal privacy-protected data publication systems have provable limits on the amount of privacy loss experienced by the worst-case outcome in the population. They 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, if 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. 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 upon whether any privacy budget remains. If the privacy budget was 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 et al. 2011; Benedetto and Stinson 2015; Machanavajjhala et al. 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 et al. 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 inference.

3.4 SDL Methods for Tabular Data

Tabular data present confidentiality risks when the number of entities contributing to a particular cell in the table is small or the influence of a few of the entities on the value of the cell is large (for magnitudes like total payroll, for example). A sensitive cell is one for which some function of the cell's microdata falls above or below a threshold set according to 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 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 the agency publishes the rest of the data, including summary statistics (Fellegi 1972). 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. It is not sufficient to suppress only the sensitive values. Enough other cells must be suppressed so that it is not possible to retrieve the sensitive values using all the published tables.

Suppressions introduce a missing data problem for researchers. Whether that missing data problem is ignorable or not will depend 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 one that uses industrial variation will benefit from using data where the complementary suppressions were over geographical classifications. Ultimately, the preferences of the agency that chooses the complementary suppression strategy will determine which analyses have higher data quality. Like swap rates, agencies rarely publish details of their methods for choosing complementary suppressions.

Input distortion of the microdata is another method for protecting tabular data. The agency distorts the value of some or all of the inputs before any publication tables are built. Then, it computes all, or almost all, of the cells using only the distorted data.

3.5 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 reported by the Federal Committee on Statistical Methodology (FCSM), which is organized by the Chief Statistician of the United States in the Office of Budget and Management, in Harris-Kojetin et al. (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. The table on page 53 presents the methods employed by each agency to protect microdata and tabular data. As of 2005, 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." 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 (Lauger et al. 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."

This report also confirms that swapping is the primary SDL method used in the ACS and decennial censuses. The swapping method targets records with high disclosure risk due to some combinations 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 the evaluation results have not been published because of concerns that they might compromise the SDL procedures themselves.

We interviewed one Census Bureau official who articulated that the rate of swapping is low relative to the rate at which data are edited for other purposes. Furthermore, 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 ten years the Census Bureau has also begun producing data based on more modern SDL methods. The Quarterly Workforce Indicators (QWI) 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 LEHD Origin-Destination Employment Statistics (LODES).[5]

4 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, as a result, researchers 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 2.1.

4.1 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 re-issued. See the Technical Appendix Section B for more details.

Suppose the confidential data contain a binary variable (gender) and a multicategory discrete variable (age). We are interested in estimation and inference for the age-specific gender distribution, where β, the conditional probability of being male given age, is the parameter of interest. When age has been subjected to SDL, the problem arises from using published age to compute these conditional probabilities. The estimated probabilities are affected by the SDL even though the gender variable was not 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 ρ, the observed male/female value comes from the true age category. With the complementary probability, the observed outcome is a binary random variable with expected value $\mu \neq \beta$. For example, μ 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, μ is unknown.

Equation (B.16) in the Technical Appendix shows that if we ignore the SDL, the conditional probability estimator and its variance are biased. An SDL-aware estimator for

[5] See, e.g. U.S. Census Bureau (2013a;b; 2015).

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 observations have been subjected SDL, then $\hat{\beta}$ is undefined, and the expected value of $\bar{z}_1$ is just μ. In the starkest possible terms, the estimator in equation (B.16) is hopelessly underidentified in the absence of information about ρ and μ.

If ρ and μ are not known, they may still be discoverable if the analyst has access to estimates of conditional probabilities like β from an alternative source. See the Technical Appendix Section B for more details of the application to the Census 2000 and ACS PUMS that generalizes the analysis in ADS. 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 ages 65 and greater and their comparable Census 2000 PUMS estimates have exactly the same functional form as appendix equation (B.17) with exactly the same SDL parameters. Since these proportions condition on the same age variable, all of 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: (1) all proportions are conditioned on the same noisy age variable, and (2) the noisy age variable can be reasonably modeled as randomized-response noise. We implement a similar method in Section 5.2.2.

4.2 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 Hirsch and Schumacher (2004) and 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) observed 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 non-union wages, and some non-union workers are imputed union wages. Bollinger and Hirsch (2006) showed 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 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 this is not conflated with any bias that could arise from non-randomness of the suppression rule.

4.2.1 SDL Applied to the Dependent Variable

The model of interest is the function $\mathrm{E}\left[y_{i1}\left|y_{i2}\right.\right] = \alpha + y_{i2}\beta$. In the published data, sensitive values of the outcome variable y_{i1} are suppressed and imputed. The variable γ_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 $p_{Y_1|X}(y_{i1}\left|x_i, \gamma_i = 0\right.)$, the conditional distribution of the outcome variable given x_i among suppressed observations. The conditioning information used in the imputation model, $x_i = f_I\left(y_{i2}\right)$, 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, γ_i and the data values, y_{i1}, 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 β. 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 β, it is not ignorable for inference. The SDL model introduces variance that is not be 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: (1) the suppression rate, $(1 - \rho) = \Pr[\gamma_i = 0]$; and (2) 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 Technical Appendix Section C.1 for details.

4.2.2 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 Technical Appendix Section C.2.

4.3 Estimating Regression Discontinuity Models

Regression discontinuity (RD) and regression kink (RK) models can be seriously compromised when SDL has been applied on the running variable. To illustrate some of these issues we consider a design from Imbens and 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.

4.3.1 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 note, this is a Rubin Causal Model (Rubin 1974; Holland 1986; Imbens and Rubin 2015). The simplest data model, corresponding to Imbens and Lemieux (pp. 616-619), 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 $\mathrm{E}\left[w_i(0)\,|r_i\right] = f_1\left(r_i\right)$ and $\mathrm{E}\left[w_i(1)\,|r_i\right] = f_2\left(r_i\right)$, where $f_1\left(r_i\right)$ and $f_2\left(r_i\right)$ are continuous functions of the running variable, r_i. The parameter of interest is the average treatment effect at τ

$$\begin{aligned}\theta_{RD} &= \lim_{r_i \downarrow \tau} \mathrm{E}\left[w_i(1)\,|r_i = \tau\right] - \lim_{r_i \uparrow \tau} \mathrm{E}\left[w_i(0)\,|r_i = \tau\right] \\ &= \lim_{r_i \downarrow \tau} f_2\left(r_i\right) - \lim_{r_i \uparrow \tau} f_1\left(r_i\right).\end{aligned}$$

4.3.2 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: $z_i \sim p_{Z|R}\left(z_i\,|r_i\right)$. We assume the distribution $p_{Z|R}\left(z_i\,|r_i\right)$ is the randomized response mixture model, a generalization of simple randomized response described in Technical Appendix Section D.1. The SDL process depends on two parameters: ρ, the probability that the confidential value of the running variable is released without added noise, and δ, 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

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

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 $\mathrm{E}\left[t_i\left|z_i, \rho_0, \delta_0\right.\right]$ playing the role of the "compliance status" function. For details, see Technical Appendix 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 $g\left(z_i\right) = \rho\, 1\left[z_i \geq \tau\right] + \left(1-\rho\right)\Phi\left(\frac{z_i-\tau}{\delta}\right)$. The fuzzy RD estimator is

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

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, τ, is equal to the share of undistorted observations, ρ_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 et al. 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.

4.3.3 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 Technical Appendix Section D.2.1 for details. Imbens and Lemieux (2008) show that the 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 ρ. Nothing can be done in this setting without auxiliary information about the SDL model.

4.3.4 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 4.2. 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.

4.3.5 Consequences of Data Coarsening for SDL

The ignorability of SDL in some circumstances was anticipated in 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 hours to salient, whole numbers. The same model is relevant to those types of microdata SDL that aggregate attribute categories, like occupations or geographies, or topcoding.

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

4.4 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

$$\begin{aligned} y_i &= \kappa + \gamma t_i + \varepsilon_i \\ t_i &= \phi + \delta z_i + \eta_i, \end{aligned}$$

where y_i is the outcome of interest, t_i is a scalar variable that may be correlated with the structural residual, ε, and z_i is a scalar variable that can serve as an instrument. That is, z is uncorrelated with ε and $\delta \neq 0$. We assume the SDL described in Section 4.2 is applied to either the dependent variable, the endogenous regressor, or the instrument.

With this simplified setup, the IV estimator is $\hat{\gamma}_{IV} = \hat{\beta}_{RF}/\hat{\delta}$, where $\hat{\beta}_{RF}$ is the parameter estimate from the reduced form equation $y_i = \alpha + \beta z_i + \nu_i$. We apply the results in Section 4.2. First, if SDL is applied to the dependent variable, then the point estimate of γ will be attenuated. This is an immediate consequence of the fact that $\operatorname{plim} \hat{\beta} \leq \beta$, while $\operatorname{plim} \hat{\delta} = \delta$. Second, by parallel reasoning, if SDL is applied to the endogenous regressor, then the point estimate of γ will be exaggerated. In this case, $\operatorname{plim} \hat{\beta} = \beta$, but $\operatorname{plim} \hat{\delta} \leq \delta$. 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.

5 Analysis of Official Tables

Tabular or aggregate data are the primary public output of most official statistical systems. For most agencies, a technical manual 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. We could find no examples 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 American 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 pre-release 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.

5.1 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 may be an option for household data. We explore some of the pitfalls of doing custom tabulations in Technical 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.

5.2 Suppression Versus Noise Infusion

5.2.1 When Suppression is Nonignorable

Tabular suppression rules identify cells that are influenced too much 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. So, it is not surprising that detailed data about the upper 0.25% of the income distribution are almost all suppressed by the Statistics of Income Division of 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, say county by NAICS industry group (4-digit) 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 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, so bias from missing data is eliminated. 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 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, 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 non-random 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 Holan et al. (2010) for a specific application to BLS data.

5.2.2 When Noise Infusion Makes the SDL Nonignorable

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 6, we devote our attention here to the basic nonignorable features of input noise infusion.

Input noise infusion models were first proposed by Evans et al. (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. Hence, 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 estimation of the parameter of interest, but will generally not be ignorable when trying to form a confidence interval around that estimate. The noise process affects the posterior distribution of most parameters of interest. It is, therefore, not ignorable in general.

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.

6 SDL Discovery in Published Tables

In this section, we show that it is possible to use information from three datasets released from very similar frames to conduct complete SDL-aware analyses. These datasets are the Quarterly Workforce Indicators (QWI), the Quarterly Census of Employment and Wages (QCEW), and the County Business Patterns (CBP). The key insight is that each dataset 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 the QCEW and CBP data can be used as instruments to correct SDL-induced measurement error in analysis based on the QWI.

6.1 Overview of the QWI, QCEW, and CBP

The Quarterly Workforce Indicators (QWI) are a collection of 32 employment and earnings statistics produced by the Longitudinal Employer-Household Dynamics (LEHD) program at the U.S. Census Bureau. They are based on state Unemployment Insurance system records, integrated with information on the characteristics of workers and the characteristics of the workplace. Characteristics of the workplace are linked from the Quarterly Census of Employment and Wages 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 quarterly series on employment and total payroll. In what follows below, the acronym QCEW is reserved for the inputs and publications of the BLS in the CEW series. County Business Patterns (CBP) are also published by the Census Bureau from inputs based on its employer Business Register.

The QWI, QCEW and CBP use highly related sources to publish statistics by employer characteristics, but 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 aggregate the undistorted confidential establishment-level microdata, and then suppress sensitive cells with enough complementary suppressions of nonsensitive cells to allow publication of most table margins.

6.2 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 Data 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 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[6]

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

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 δ_j that have mean zero and constant variance. Details are in the Data Appendix Section G. The published quarterly payroll is computed as

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

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 $W^*_{(k)t}$, 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

[6] We abstract from the weight that QWI uses to benchmark certain state-level aggregates. Formulas including weights are in the Data Appendix Section H.

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.

6.3 Regression Models with Nonignorable SDL

The noise infusion in QWI may be nonignorable. Univariate regression of a variable from another dataset onto a QWI aggregate provides a simple illustration, which we summarize here. See Technical Appendix Section E.4 for details.

The model of interest is 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 $W^*_{(k)t}$. The dependent variable can be subjected to SDL as long as it is independent of the QWI SDL, as will be the case for example if the dependent variable was computed by the BLS or 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 OLS estimator for the regression coefficient on β based 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 Data 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 $\text{Var}\left[\delta_j\right]$ is known, after which it only remains to recover $\text{Var}\left[W_{(k)t}\right]$ from the definition of $\text{Var}\left[W^*_{(k)t}\right]$.

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: (1) data from the QCEW for the same cell; (2) data from CBP from the same cell; and (3) 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. 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 upon 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 5.2.1. 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, and factors out of the OLS moment equations, 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.

6.4 Estimating the Variance Contribution of SDL for the QWI

It is possible to recover the variance of the noise factor $\mathrm{Var}\left[\delta_j\right]$, 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 in Technical 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 Data Appendix E.5.

6.5 Empirical Results

Table 1 presents the estimates of the equation used to recover the SDL parameters fit 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.[7] The equations are fit 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 $\mathbb{V}\left[\delta\right]$ whether we use payroll or employment as the basis. This suggests that the bias in estimating $\mathbb{V}\left[\delta\right]$ 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 $\mathbb{V}\left[\delta\right]$, and these indicate that the noise infusion does not create a very substantial bias or inflate estimated variances substantially.

7 The Frontiers of SDL

7.1 Analysis of Synthetic Data

We defined synthetic data in Section 3. 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

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

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 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 Synthetic Beta (SSB) (U.S. Census Bureau 2013a) and the Synthetic Longitudinal Business Database (SynLBD) (U.S. Census Bureau 2013b). 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 Bertrand et al. (2015) provides an excellent illustration of the advantages of using synthetic data that are part of a feedback loop. The authors used 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 hypothesized that there should be a regression discontinuity at 50% 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%. Analyses from synthetic data showed 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, showed 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-based server that is used as the interface for analyses that are part of the feedback process.[8]

In the process of writing this paper, we discovered why the Bertrand et al. analysis of the linked SIPP-IRS data showed no causal effect when using the synthetic data. 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 will not, in general, be identified for any RD design that uses the many exact earnings and date variables in the SSB. If only the SSB

[8] http://www2.vrdc.cornell.edu/news/synthetic-data-server/step-1-requesting-access-to-sds/

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. That's an incredible silver lining.

We have already noted that the Survey of Consumer Finances 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 feedback loop that is described in the codebook (Federal Reserve Board of Governors 2013).

7.2 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, PINQ, that allowed researchers to analyze its internal data files (search logs, etc.) 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 ε-differential privacy. When the researcher exhausts ε, 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 non-trivial research questions. However, they facilitate a formal expression of ideas that are inherent, but implicit, in SDL. They make clear the cost in terms of loss of accuracy that is inherent in protecting privacy by distorting the analysis of the confidential data. They also allow setting of 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's a legitimate worry, and one that the formal privacy community takes very seriously. For a glimpse of one possible future, see Dwork (2014)–paying particular attention to the call 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 non-private 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.

7.3 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.

8 Discussion

8.1 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 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. These studies have very limited potential for replication or reuse of the confidential data. Grant funding agencies have insisted that researchers whom they fund prepare a data management plan for the curation of the data developed and analyzed using grant funds. Very few statistical agencies or private firms will allow research teams to comply with this requirement by surrendering a copy of the confidential data for secure curation. 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 most of the intermediate work product cannot be published.

Suppression on this scale leads to potentially massive biases and very limited data releases. To address this problem, statisticians and computer scientists have worked over these same decades 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 wasn't accomplished again until the Census Bureau released synthetic microdata for the Survey of Income and Program Participation in 2007. We believe that the reason for economists' reticence to explore alternatives to suppression is an incomplete understanding of how pernicious suppression bias really 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's true, as we have shown, but incomplete. 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 is a potentially 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. These topics are usually covered in courses called "Survey Methodology," but they belong in econometrics and economic measurement courses too.

8.2 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 don't 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. Discussion 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 the authors can't cite sources that document each of these components, they should be required to include the information in the 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, but we don't 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.

8.3 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, *etc*. Very specific information is released about items that have been edited, imputed or otherwise altered to address data quality concerns. Virtually nothing – *specific* – about the SDL parameters is released. 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 at they begin to release data products like Google Trends for use by the research community.

9 Conclusion

While 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 define ignorable SDL methods in a model-dependent manner that is exactly parallel to the way ignorability is defined for missing data models. 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 – re-identify 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, at least for the models we considered, the consequences of the SDL for the data analysis were limited. 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: (1) 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 (2) less trepidation by our 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.

Bibliography

Abowd, J. M. and Woodcock, S. (2001). Disclosure limitation in longitudinal linked data, *in* P. Doyle, J. Lane, L. Zayatz and J. Theeuwes (eds), *Confidentiality, Disclosure, and Data Access: Theory and Practial Applications for Statistical Agencies*, North Holland, pp. 215–277.

Alexander, J. T., Davern, M. and Stevenson, B. (2010). Inaccurate age and sex data in the Census PUMS files: Evidence and implications, *Public Opinion Quarterly* **74**(3): 551–569.

Anderson, M. and Seltzer, W. (2007). Challenges to the confidentiality of US federal statistics, 1910-1965, *Journal of Official Statistics* **23**(1): 1–34.

Anderson, M. and Seltzer, W. (2009). Federal statisitcal confidentiality and business data: Twentieth century challenges and continuing issues, *Journal of Privacy and Confidentiality* **1**(1): 7–52.

Benedetto, G. and Stinson, M. (2015). Disclosure review board memo: Second request for release of SIPP synthetic beta version 6.0, *Technical report*, U. S. Census Bureau, Survey Improvement Research Branch, Social, Ecomomic, and Housing Statistics Division.

Bertrand, M., Pan, J. and Kamenica, E. (2015). Gender identity and relative income within households, *Quarterly Journal of Economics* **130**(2): 571–614.

Bollinger, C. R. and Hirsch, B. T. (2006). Match bias from earnings imputation in the Current Population Survey: The case of imperfect matching, *Journal of Labor Economics* **24**(3): 483–520.

Burkhauser, R. V., Feng, S., Jenkins, S. P. and Larrimore, J. (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**(2): 371–388.

Card, D., Lee, D., Pei, Z. and Weber, A. (2012). Nonlinear policy rules and the identification and estimation of causal effects in a generalized regression kink design, *Working Paper 18564*, National Bureau of Economic Research.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control, *Statistik Tidskrift* **15**: 429–444.

Duncan, G. and Lambert, D. (1986). Disclosure-limited data dissemination, *Journal of the American Statistical Association* **81**(393): 10–18.

Duncan, G. T., Elliot, M. and Salazar-González, J.-J. (2011). *Statistical Confidentiality Principles and Practice*, Statistics for Social and Behavioral Sciences, Springer New York.

Duncan, G. T. and Fienberg, S. E. (1999). Obtaining information while preserving privacy: A markov perturbation method for tabular data, *Statistical Data Protection (SDP '98)*, Eurostat, pp. 351–362.

Duncan, G. T., Jabine, T. B. and de Wolf, V. A. (eds) (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics; Panel on Confidentiality and Data Access [of the] Committee on National Statis- tics, Commission on Behavioral and Social Sciences and Education, National Research Council and the Social Science Research Council*, The National Academies Press, Washington, DC.

Dwork, C. (2006). Differential privacy, *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, pp. 1–12.

Dwork, C. (2014). Differential privacy: A cryptographic approach to private data analysis, *in* J. Lane, V. Stodden, S. Bender and H. Nissenbaum (eds), *Privacy, Big Data, and the Public Good*, Cambridge University Press, chapter 14, p. 296.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis, *Proceedings of the Third conference on Theory of Cryptography*, TCC'06, Springer-Verlag, Berlin, Heidelberg, pp. 265–284.

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

Evans, T., Zayatz, L. and Slanta, J. (1998). Using noise for disclosure limitation for establishment tabular data, *Journal of Official Statistics* **14**(4): 537–551.

Evfimievski, A., Gehrke, J. and Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining, *ACM SIGMOD Principles of Database Systems (PODS)*, pp. 211–222.

Federal Reserve Board of Governors (2013). Codebook for 2013 Survey of Consumer Finances, *Technical report*, Federal Reserve Board of Governors.

Fellegi, I. P. (1972). On the question of statistical confidentiality, *Journal of the American Statistical Association* **67**(337): pp. 7–18.

Goldwasser, S. and Micali, S. (1982). Probabilistic encryption & how to play mental poker keeping secret all partial information, *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, ACM, pp. 365–377.

Hardt, M., Ligett, K. and McSherry, F. (2012). A simple and practical algorithm for differentially private data release, *Advances in Neural Information Processing (NIPS)* **abs/1012.4763**.

Harris-Kojetin, B. A., Alvey, W. L., Carlson, L., Cohen, S. B., Cohen, S. H., Cox, L. H., Fay, R. E., Fecso, R., Fixler, D., Gates, G., Graubard, B., Iwig, W., Kennickell, A., Kirkendall, N. J., Schechter, S., Schmitt, R. R., Seastrom, M., Sirken, M. G., Spruill, N. L., Tucker, C., Tupek, A. R., Williamson, G. D. and Groves, R. (2005). Report on statistical disclosure limitation methodology, *Research Report Statistical Policy Working Paper 22*, Federal Committee on Statistical Methodology.

Heffetz, O. and Ligett, K. (2014). Privacy and data-based research, *Journal of Economic Perspectives* **28**(2): 75–98.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data, *The Annals of Statistics* **19**(4): 2244–2253.

Hirsch, B. T. and Schumacher, E. J. (2004). Match bias in wage gap estimates due to earnings imputation, *Journal of Labor Economics* **22**(3): 689–722.

Holan, S. H., Toth, D., Ferreira, M. A. and Karr, A. F. (2010). Bayesian multiscale multiple imputation with implications for data confidentiality, *Journal of the American Statistical Association* **105**(490).

Holland, P. W. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **81**(396): 945–960.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice, *Journal of Econometrics* **142**(2): 615–635.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference for Statistics, Social and Biomedical Sciences*, Cambridge University Press New York.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality, *The American Statistician* **60**(3): 224–232.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 survey of consumer finances, *Technical report*, Washington DC: National Academy Press.

Kennickell, A. and Lane, J. (2006). Measuring the impact of data protection techniques on data utility: Evidence from the Survey of Consumer Finances, *in* J. Domingo-Ferrer and L. Franconi (eds), *Privacy in Statistical Databases*, Vol. 4302 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 291–303.

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

Larrimore, J., Burkhauser, R. V., Feng, S. and Zayatz, L. (2008). Consistent cell means for topcoded incomes in the public use March CPS (1976–2007), *Journal of Economic and Social Measurement* **33**(2): 89–128.

Lauger, A., Wisniewski, B. and McKenna, L. (2014). Disclosure avoidance techniques at the U.S. Census Bureau: Current practices and research, *Technical Report 2014-02*, Center for Disclosure Avoidance Research U.S. Census Bureau.

Lee, D. S. and Card, D. (2008). Regression discontinuity inference with specification error, *Journal of Econometrics* **142**(2): 655–674.

Little, R. J. A. (1993). Statistical analysis of masked data, *Journal of Official Statistics* **9**(2): 407–426.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L. (2008). Privacy: Theory meets practice on the map, *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pp. 277 –286.

McSherry, F. (2009). Privacy integrated queries: An extensible platform for privacy-preserving data analysis, *SIGMOD '09*.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets, *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, IEEE, pp. 111–125.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization, *UCLA Law Review* **57**: 1701.

Piketty, T. and Saez, E. (2003). Income inequality in the United States, 1913–1998, *The Quarterly Journal of Economics* **118**(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**(1): 1–16.

Reiter, J. P. (2004). Smultaneous use of multiple imputation for missing data and disclosure limitation, *Survey Methodology* **30**: 235–242.

Reiter, J. P. (2005). Estimating risks of identification disclosure in microdata, *Journal of the American Statistical Association* **100**(472): 1103–1112.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**(5): 688–701.

Rubin, D. B. (1976). Inference and missing data, *Biometrika* **63**(3): 581–592.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation, *Journal of Official Statistics* **9**(2): 461–468.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata, *Journal of Official Statistics* **14**(4): 361–372.

Skinner, C. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models, *Journal of the American Statistical Association* **103**(483): 989–1001.

Sweeney, L. (2000). Uniqueness of simple demographics in the U.S. population, *Technical report*, Technical report, Carnegie Mellon University.

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

U.S. Census Bureau (2013b). Synthetic longitudinal business database: Version 2.0 [computer file], Washington DC; Cornell University, Synthetic Data Server [distributor], Ithaca, NY.

U.S. Census Bureau (2015). Lehd origin-destination employment statistics (lodes), Washington DC; U.S. Census Bureau [distributor].

Warner, S. L. (1965). A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**(309): 63–69.

Yakowitz, J. (2011). Tragedy of the data commons, *Harvard Journal of Law and Technology* **25**: 1.

Table 1: Estimated Variance of QWI Establishment Noise Factor (δ)

	County-Sector		County		State	
	ln(Emp.)	ln(Payroll)	ln(Emp.)	ln(Payroll)	ln(Emp.)	ln(Payroll)
	(1)	(2)	(3)	(4)	(5)	(6)
$\ln(Num.Estab.)$	$-.281$	$-.211$	$-.209$	$-.155$	-0.144	0.205
	$(.0016)$	$(.0017)$	$(.0061)$	$(.0062)$	$(.0679)$	$(.0987)$
Constant	-1.537	-1.610	-2.027	-1.962	-4.679	-6.747
	$(.0153)$	$(.0070)$	$(.0408)$	$(.0422)$	$(.7885)$	(1.159)
Num. Obs.	$228,770$	$236,925$	$18,000$	$18,057$	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 2006–2011. 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 $\widehat{V[\delta_k]} = \exp(-2 \times Constant)$. Columns (1) and (2) report estimates from data disaggregated by county and NAICS major sector. Columns (3) and (4) report county-level estimates. Columns (5) and (6) report state-level estimates.

Table 2: Estimated Variance of QWI Establishment Noise Factor (δ): Mixed Models

	County-Sector		County		State	
	ln(Emp.)	ln(Payroll)	ln(Emp.)	ln(Payroll)	ln(Emp.)	ln(Payroll)
	(1)	(2)	(3)	(4)	(5)	(6)
$\ln(Num.Estab.)$	$-.321$	$-.219$	$-.224$	$-.153$	-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)
Num. Obs.	$228,770$	$236,925$	$18,000$	$18,057$	282	282
Var. Fuzz $(V[\delta])$	0.060	0.043	0.019	0.019	0.0001	0.000

SOURCE: QCEW and QWI data for Q1 for years 2006–2011. 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. The table layout is identical to Table 1.

Technical Appendix for "Economic Analysis and Statistical Disclosure Limitation"

John M. Abowd
Department of Economics
Labor Dynamics Institute
Cornell University
john.abowd@cornell.edu

Ian M. Schmutte
Department of Economics
Terry College of Business
University of Georgia
schmutte@uga.edu

August 7, 2015

A Ignorable and Nonignorable SDL

We formalize the role of SDL in economic analysis using the concept of ignorability. Our approach is a direct extension of the ignorability of missing data developed by Rubin (1976). Little (1993) anticipated much of our analysis, including the use of hierarchical models that introduced SDL via generalized randomized response. We first define the economic process model that the econometrician is trying to learn. We then define the inclusion process that determines which parts of the economic process are actually observed. This gives rise to the well-known concept of *ignorable missing data* or, equivalently, *ignorable inclusion*. Finally, we formally define the SDL model and define *ignorable statistical disclosure limitation*.

A.1 The Economic Process Model

We consider a population of N entities that is described by a *complete-data* matrix Y, $N \times K$, a *process-parameter* vector θ_p, $P \times 1$, and two probability distributions: the *data model* $p_Y(Y|\theta_p)$ and the *process-parameter prior distribution* $p_{\theta_p}(\theta_p)$.

The econometrician seeks to conduct estimation and inference concerning finite-population estimands, functions of Y only, and super-population estimands, functions of the parameters θ_p. We distinguish between these two estimand types because the statistical agencies that collect and disseminate the data we are discussing in this paper consider themselves to be engaged in producing finite-population estimands whereas the economists who analyze these data are primarily conducting super-population estimation and inference.[9]

A.2 The Data Inclusion Model and Ignorable Inclusion

Next, we define the tools necessary to understand the properties of published (released) data from conventional surveys, censuses, and administrative record systems. The population *inclusion matrix*, R, $N \times K$, indicates that an entity i has data for the associated

[9] Many SDL methods, as well as methods from the newer data-privacy literature in computer science, explicitly consider the properties of these methods for finite-population estimands whereas econometricians tend to focus on parametric (or semi-parametric) modeling focused on θ_p. The concept of ignorability was invented to allow a clean characterization of how the data collection process affects both types of modeling. We are not trying to be overly philosophical, just to provide a direct link between the way the data collectors think about the methods they use and the way data analysts trained in economics and econometrics use those data.

variable, $r_{ij} = 1$, or not, $r_{ij} = 0$. If you think that this is needlessly complex, remember that we have not said that N is known nor how the statistician came to observe any element of Y. That is the role of the *inclusion model*: the distribution of R given Y is $p_{R|Y}\left(R\left|Y,\theta_D\right.\right)$. θ_D, is the *design* parameter vector, so named because it characterizes how Y is observed, or the design of the survey or experiment. The *design-parameter prior distribution* is $p_{\theta_D|\theta_p}\left(\theta_D\left|\theta_p\right.\right)$ allows for potential dependence of the design on the process parameters. The complete-data likelihood function[10] is then

$$\mathcal{L}_\theta\left(\theta_p,\theta_D\left|Y,R\right.\right) = p_Y\left(Y\left|\theta_p\right.\right)p_{R|Y}\left(R\left|Y,\theta_D\right.\right) = p_{YR}\left(Y,R\left|\theta_p,\theta_D\right.\right). \tag{A.1}$$

The term "complete data" means that this likelihood function applies to estimation and inference on the process and design parameters given a realization of Y, R from the superpopulation.

The *observed data* matrix, in the absence of SDL, is $Y^{(obs)}$, $N \times P$, contains a data item in $y_{ij}^{(obs)}$, if and only if $r_{ij} = 1$. The complement to the observed data matrix, in the absence of SDL is $Y^{(mis)}$, which contains the unobserved data items corresponding to $r_{ij} = 0$. The observed data likelihood function, in the absence of SDL is

$$\begin{aligned}\mathcal{L}_\theta^{(obs)}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right) &= p_{Y^{(obs)}R}\left(Y^{(obs)},R\left|\theta_p,\theta_D\right.\right) & \text{(A.2)}\\ &= \int p_{YR}\left(Y,R\left|\theta_p,\theta_D\right.\right)dY^{(mis)}. & \text{(A.3)}\end{aligned}$$

The term "observed data" derives from the application of these modeling concepts to sampling, experimental design, and unintentionally missing data (missing survey records or responses, unreported administrative records, etc.). In the standard analysis of ignorability (e.g., Gelman et al. 2013), the published data would be $Y^{(obs)}$. The notation may seem awkward for the application to SDL, but it seems better to us to use this conventional notation. Wherever the term $Y^{(obs)}$ occurs, think: the actual confidential data collected by the statistical agency.

Inference and estimation, in the absence of SDL, are based on the joint posterior distribution of (θ_p, θ_D), given the observed data, which we assemble from the pieces defined above as

$$\begin{aligned}p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right) &\propto p_{\theta_D|\theta_p}\left(\theta_D\left|\theta_p\right.\right)p_{\theta_p}\left(\theta_p\right)p_{Y^{(obs)}R}\left(Y^{(obs)},R\left|\theta_p,\theta_D\right.\right)\\ &= p_{\theta_D|\theta_p}\left(\theta_D\left|\theta_p\right.\right)p_{\theta_p}\left(\theta_p\right)\mathcal{L}_\theta^{(obs)}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right). \quad \text{(A.4)}\end{aligned}$$

In general, we focus interest on the posterior distribution of θ_p which, in the absence of

[10] The Rubin formulation includes the notion of fully observed covariates–variables that are never missing in the population and never have to be collected. In a known, finite population, these consist of variables on the frames used for sampling. Since these variables are also subjected to SDL when the data are published, we include them in the population data matrix Y.

SDL, is

$$p_{\theta_P|Y^{(obs)}R}\left(\theta_p\left|Y^{(obs)},R\right.\right) = \int p_{\theta|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right)d\theta_D \tag{A.5}$$
$$\propto \int\int p_Y\left(Y\left|\theta_p\right.\right)p_{R|Y}\left(R\left|Y,\theta_D\right.\right)p_{\theta_D|\theta_p}\left(\theta_D\left|\theta_p\right.\right)p_{\theta_p}\left(\theta_p\right)dY^{(mis)}d\theta_D$$

The data inclusion model is *ignorable* if

$$p_{\theta_P|Y^{(obs)}R}\left(\theta_p\left|Y^{(obs)},R\right.\right) \equiv p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right). \tag{A.6}$$

For reasons that will be clear shortly, we call this *ignorable inclusion* (or *ignorable sampling*, or *ignorable missing data*, if the context of the inclusion model is clear).

Our definition of ignorability is general enough to cover observational data, survey designs, experiments, and unintentional missing data models. It says that inference and estimation about the super-population parameters is ignorable if it does not depend on the unobserved data, $Y^{(mis)}$. It is not general enough to cover SDL because $Y^{(obs)}$ undergoes an additional transformation before being published.

A.3 The SDL Model and Ignorable SDL

We characterize the SDL probabilistically using the same tools as we have used for the data model, the inclusion model, and their parameters. The *published data* Z, $N \times K$, are generated by the *SDL model* $p_{Z|Y,R}\left(Z\left|Y,R,\theta_S\right.\right)$ with *SDL-parameter* vector θ_S. The *SDL-parameter prior distribution* is $p_{\theta_S|\theta_D\theta_p}\left(\theta_S\left|\theta_D,\theta_p\right.\right)$. The likelihood function for the published data is

$$\pounds_\theta^{(pub)}\left(\theta_p,\theta_D,\theta_S\left|Z,R\right.\right) = \int p_{Z|YR}\left(Z\left|Y,R,\theta_S\right.\right)p_{YR}\left(Y,R\left|\theta_p,\theta_D\right.\right)dY \tag{A.7}$$
$$= \int p_{Z|YR}\left(Z\left|Y,R,\theta_S\right.\right)p_{R|Y}\left(R\left|Y,\theta_D\right.\right)p_Y\left(Y\left|\theta_p\right.\right)dY$$

Once again, estimation and inference are based on the posterior distribution of the process parameters, which is derived from the joint posterior distribution of the model, inclusion, and publication parameters given the published data and the inclusion matrix

$$p_{\theta|ZR}\left(\theta_p,\theta_D,\theta_S\left|Z,R\right.\right) \propto \int p_{Z|YR}\left(Z\left|Y,R,\theta_S\right.\right)p_{YR}\left(Y,R\left|\theta_p,\theta_D\right.\right)p_\theta\left(\theta\right)dY$$
$$= p_\theta\left(\theta\right)\pounds_\theta^{(pub)}\left(\theta_p,\theta_D,\theta_S\left|Z,R\right.\right),$$

where $p_\theta\left(\theta\right) = p_{\theta_S|\theta_D\theta_p}\left(\theta_S\left|\theta_D,\theta_p\right.\right)p_{\theta_D|\theta_p}\left(\theta_D\left|\theta_p\right.\right)p_{\theta_p}\left(\theta_p\right)$. So that the posterior distribution of the process parameters is

$$p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) = \int\int p_{\theta|ZR}\left(\theta_p,\theta_D,\theta_S\left|Z,R\right.\right)d\theta_D d\theta_S. \tag{A.8}$$

The relation between equations (A.5) and (A.8) is

$$p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right)=\int p_{\theta_P|Y^{(obs)}R}\left(\theta_p\left|Y^{(obs)},R\right.\right)p_{Y^{(obs)}|ZR}\left(Y^{(obs)}\left|Z,R\right.\right)dY^{(obs)}. \tag{A.9}$$

That is, the posterior distribution of the process parameters θ_p given the published data and inclusion matrix is the expectation of the posterior distribution of the process parameters given the observed data (the actual confidential data used by the agency) and inclusion matrix with the expectation taken over the posterior predictive distribution of the observed data given the published data and inclusion matrix. This formulation assumes that the agency also publishes R, which is not innocuous but we will usually be analyzing models in which we assume ignorable inclusion.

We define *ignorable statistical disclosure limitation* as

$$p_{\theta_P|Y^{(obs)}R}\left(\theta_p\left|Y^{(obs)}=Z,R\right.\right)\equiv p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) \tag{A.10}$$

for all $Y^{(obs)}$, Z, and R.

The definition is subtle, so we repeat it in words. The SDL is ignorable if and only if analyzing the posterior distribution of the process parameters given the published data is equivalent to analyzing the posterior distribution of process parameters given the observed data and assuming that the published data are identical to the (confidential) observed data.

If the model possesses both ignorable inclusion and ignorable SDL then

$$p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}=Z\right.\right)\equiv p_{\theta_P|Z}\left(\theta_p\left|Z\right.\right) \tag{A.11}$$

for all $Y^{(obs)}$ and Z. Equation (A.11) summarizes both the sampling (or inclusion) and SDL assumptions that are embodied in any economic analysis that treats the published data as if they had been produced by an ignorable inclusion process without SDL; that is, without explicitly modeling the sample design and SDL.

A.4 Implementing SDL-aware Data Analysis

Since equation (A.9) is an identity, it is, in principle, possible to do any data analysis using methods that account for the SDL. In practice, we must confront whether or not the SDL process is known, and if it is known, whether the components required to compute $p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right)$ can be assembled. We will define an SDL method as *fully discoverable* if $p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right)$ can be computed. If the SDL process is not fully discoverable, then we will consider some diagnostic methods that can be used to approximate $p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right)$ or to detect failures of equation (A.10).

At the heart of the implementation is the computation of $p_{Y^{(obs)}|ZR}\left(Y^{(obs)}\left|Z,R\right.\right)$, which is the posterior predictive distribution of the data that would have been published in the absence of SDL, given the published data and the inclusion matrix. In the absence of any ignorability assumptions the computations can be done using Markov Chain Monte

Carlo sampling from the conditional distributions

$$p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right)$$

$$p_{\theta_S|ZR\theta_p\theta_D}\left(\theta_S\left|Z,R,\theta_p,\theta_D\right.\right)$$

$$p_{Y^{(obs)}|ZR\theta_p\theta_D\theta_S}\left(Y^{(obs)}\left|Z,R,\theta_p,\theta_D,\theta_S\right.\right)$$

starting from arbitrary initial values of $Y^{(obs)}$, and $(\theta_p,\theta_D,\theta_S)$.

In many ways, implementing SDL-aware data analysis is similar to implementing ignorable and nonignorable missing data models. Since there are many excellent discussions of missing data issues and in order to focus our contribution more clearly, we consider next implementing SDL-aware analysis when the inclusion model is provably ignorable. A leading case is the inclusion model in which data are missing at random in the sense of Rubin (1987); then, inclusion model can be ignored because

$$p_{R|Y}\left(R\left|Y,\theta_D\right.\right)=p_{R|Y}\left(R\left|Y^{(obs)},\theta_D\right.\right)$$

and

$$p_\theta\left(\theta\right)=p_{\theta_S|\theta_p\theta_D}\left(\theta_S\left|\theta_p,\theta_D\right.\right)p_{\theta_D}\left(\theta_D\right)p_{\theta_p}\left(\theta_p\right)$$

To further simplify, simple random sampling implies that the inclusion model does not depend upon any unknown parameters nor on the population data; hence $p_{R|Y}\left(R\left|Y,\theta_D\right.\right)=p_R\left(R\right)$, which allows R and θ_D to be eliminated altogether from the analysis of the published data.

It is enlightening to study the SDL-aware data analysis equations under the assumption that the inclusion model is ignorable and known. Then,

$$\begin{aligned}p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) &= p_{\theta_P|Z}\left(\theta_p\left|Z\right.\right)\\ &= \int p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right)p_{Y^{(obs)}|Z}\left(Y^{(obs)}\left|Z\right.\right)dY^{(obs)} \qquad \text{(A.12)}\end{aligned}$$

$$p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right)=p_{\theta_p|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right) \qquad \text{(A.13)}$$

$$p_{\theta_S|ZR\theta_p\theta_D}\left(\theta_S\left|Z,R,\theta_p,\theta_D\right.\right)=p_{\theta_S|Z\theta_p}\left(\theta_S\left|Z,\theta_p\right.\right) \qquad \text{(A.14)}$$

and

$$p_{Y^{(obs)}|ZR\theta_p\theta_D\theta_S}\left(Y^{(obs)}\left|Z,R,\theta_p,\theta_D,\theta_S\right.\right)=p_{Y^{(obs)}|Z\theta_p\theta_S}\left(Y^{(obs)}\left|Z,\theta_p,\theta_S\right.\right). \qquad \text{(A.15)}$$

Estimation and inference using the SDL-aware system described by equations (A.12)-(A.15) can be applied to many common SDL methods, including those introduced in the data-privacy literature in CS.

Although we largely limit our attention in this paper to SDL-aware analyses that assume that the inclusion model is known and ignorable, we do not mean to endorse these assumptions universally. In particular, we have chosen many examples where the inclusion model's properties are well understood or provably ignorable.

A.5 Using Conditional Probability Models to Discover Nonignorable SDL

Consider the data model in which y_i contains K variables with y_{i1} binary and the remaining variables either continuous or discrete. Although the formal data model remains $p_{y|\theta_P}\left(Y^{(obs)}\,|\theta_P\right)$, interest focuses on estimation and inference for the conditional probabilities

$$\Pr\left[y_{i1} = 1\,|y_{i2}, \beta\right]$$

where β is the process parameter vector of interest (linear probability model coefficients, logit coefficients, probit coefficients, etc.). The remaining process parameters are nuisance parameters in an analysis with access to $Y^{(obs)}$. We consider here the use of conditional probability models as diagnostic tools for discovering nonignorable SDL.

If the analyst is completely ignorant of the process generating y_i the SDL is not discoverable unless its details are published by the agency or it is generated by a formal privacy model with public parameters. The intuitive notion that information in related data can be used to discover SDL properties lies at the heart of the Alexander et al. (2010) analysis of the 2000 Census and ACS Public-Use Microdata Samples (PUMS). In those examples y_{i1} is the individual's sex and y_{i2} is the individual's birth date (or age). They (implicitly) use an informative prior on β based on the population summary files for the 2000 Census (which are based on all records, not just the PUMS records) and the published tabulations for the ACS (which are based on the full ACS sample, not just the records in the PUMS) to estimate the effects of SDL on analyses using the PUMS files. In their case, the informative prior distribution was sufficient to estimate β accurately because β was actually a finite-population estimand (the proportion of the age cohort that is in each sex for the U.S. population at a point in time). In addition, because they used a finite-population estimand where the variability of β in the prior distribution was negligible, they could assess the probability that the differences were due to chance from the posterior variability in the PUMS files alone. In general, this won't be the case, but the intuition underlying their method is more broadly applicable for discovering nonignorable SDL.

Conditional probability models analyzed using SDL-aware procedures with informative priors can render the SDL discoverable in both our formal sense and the intuitive sense used by Alexander et al. To develop this point formally, we can no longer assume that the inclusion process, in this case the sampling model, is ignorable because this process contributes to the posterior distribution of the process parameters and to an informative prior distribution on those parameters, but not necessarily in the same manner. In addition, we will need to be precise in making assumptions about the SDL. SDL processes used in related data publications may share parameters, random noise, and conditioning variables. We will have to be formal about conditioning on or integrating out these SDL components in the informative prior as well as in posterior of interest. The payoff is that we can get probability models that provide a formal basis for what Alexander et al. did and are more generally applicable. In our empirical examples, we are careful to select published data files where the dependencies in the SDL have been documented by the suppliers.

B Details of Estimating Population Proportions with Noise Infusion

Suppose the confidential data, y_i, contain K variables with y_{i1} binary and the remaining variables either continuous or discrete. We are interested in estimation and inference for the conditional probabilities $\Pr\left[y_{i1} = 1\,|y_{i2}, \beta\right]$, where β is the parameter of interest. The problem arises from using $\Pr\left[z_{i1} = 1\,|z_{i2}, \beta, \theta_S\right]$ where the z_i variables are the published versions of y_i and θ_S are the parameters of the SDL.

To facilitate the exposition, consider just one outcome z_{i1}, which can be either zero or one. For example, the observed z_{i1} could be an indicator that the respondent is male and the conditioning set, z_{i2} could be age 65. With probability ρ, the published data come from the same conditioning set as in the confidential data; that is, $z_{i2} = y_{i2}$. For example, if the stratification is on age, then with probability ρ, the observed outcome comes from the true age category; that is, $z_{i1} = y_{i1}$ for $y_{i2} = 1\left[\text{true age } = 65\right]$. With the complementary probability, the observed outcome is a binary random variable with expected value $\mu \neq \beta$, for example, the average value of proportion male over all age categories at risk to be changed by the SDL model.

Under these conditions and using $\mathrm{E}\left[z_{i1} = 1\,|z_{i2}, \beta, \rho, \mu\right]$ the consistent estimator for the process parameter of interest, β, is

$$\hat{\beta} = \frac{\bar{z}_1 - (1-\rho)\,\mu}{\rho} \tag{B.16}$$

where $\bar{z}_1$ is the estimated sample proportion of ones (i.e., males). 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 none of the published age data has been infused with noise. If all of observations have been subjected to this noise infusion, then $\hat{\beta}$ is undefined, and the expected value of $\bar{z}_1$ is just μ. In the starkest possible terms, the estimator in equation (B.16) is hopelessly underidentified in the absence of information about ρ and μ.

If ρ and μ are not known, they may still be discoverable if the analyst has access to estimates of conditional probabilities like β from an alternative source. Here is an example based on the analysis in ADS. Comparing the sex proportions estimated from the Census 2000 PUMS to the published Census 2000 data, and treating the published Census 2000 estimates as the true values, we have

$$\mathrm{E}\left[\bar{z}_{j1} - \bar{y}_{j1}\,|\bar{y}_{j1}\right] = \rho_j \bar{y}_{j1} + \left(1-\rho_j\right)\mu_j - \bar{y}_{j1} \tag{B.17}$$

for $j =$ ages $65, 66, 67, \ldots 89$.

The SDL process is still underidentified if we consider only a single outcome like sex, 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 ages 65 and greater and their comparable Census 2000 PUMS estimates have exactly the same functional form as equation (B.17) with exactly the same SDL parameters. Since these proportions condition on the same age variable, all of the other outcomes that also have an official Census 2000 published proportion can be used to estimate ρ_j and μ_j. The identifying assumptions are: (1) all proportions are all conditioned on the same noisy age variable, and (2) the noisy age variable can be reasonably modeled as randomized-response noise.

B.1 History of the Census Bureau's Correction to Census 2000 and ACS PUMS Files

The original announcement that the PUMS files would not be corrected can be found in Census 2000 Public Use Microdata Sample Data Note 12 (October 2010) and the reversal in Data Note 13 (October 2010) http://www.census.gov/prod/cen2000/doc/pums.pdf. The original announcement that the ACS PUMS files would not be corrected can be found in Errata 47 (February 18, 2010) and 50 (December 18, 2009). The reversal is in Erratum 65 (January 25, 2012). See also User note 3. Cited documents http://www.census.gov/acs/www/data_documentation/errata/#Err47, http://www.census.gov/acs/www/data_documentation/errata/#Err50, http://www.census.gov/acs/www/data_documentation/errata/index.php#Err65, and http://www.census.gov/acs/www/data_documentation/user_notes/index.php#n03(cited March 19, 2015).

C Details of Estimating Regression Models with SDL

C.1 Bias due to SDL in the Dependent Variable

For the case in which SDL is applied to the dependent variable, our derivation of the bias formula is a direct extension of the analysis in Sections 1.1, 1.2, and 1.3 in the Appendix to Bollinger and Hirsch (2006). Our only modification is to the equation characterizing the distribution of imputed data. In the Bollinger and Hirsch Appendix, equation (1) states

$$f_I\left(y_i, z_i|x_i\right) = f_O\left(y_i|x_i\right) f_M(z_i|x_i),$$

where $f_I\left(y_i, z_i|x_i\right)$ is the joint distribution of y and z in the imputed data, $f_O\left(y_i|x_i\right)$ is the distribution of the dependent variable, y, given the matching variables, x, among the observed data, and $f_M(z_i|x_i)$ is the distribution of the regressors, z, conditional on x, among the missing data.

In our application, y is sometimes missing, not because it was not reported, but because it is suppressed. That means the imputed values can be drawn from the distribution of the suppressed data. Formally, this just amounts to changing the above equation to

$$f_I\left(y_i, z_i|x_i\right) = f_M\left(y_i|x_i\right) f_M(z_i|x_i).$$

This change does not affect the remaining derivations in sections 1.1–1.3 of the Bollinger and Hirsch Appendix; the bias formula remains the same. This is just a change of interpretation. Specifically, whereas Bollinger and Hirsch must make an assumption on the missing data process (namely, that the data are conditionally missing at random), we require no such assumption on the suppression process.

The SDL is ignorable for estimation and inference of β if the solution to the least squares projection of z_{1i} on z_{2i} yields estimates consistent for the parameters of the true regression model: $\mathrm{E}\left[y_{i1}\left|y_{i2}\right.\right] = \alpha + y_{i2}\beta$. The solution to the least squares projection is $\left(\hat{a}, \hat{b}\right) = \arg\min_{a,b} \mathrm{E}\left[\left(z_{i1} - a - z_{2i}b\right)^2\right]$.

We start with the case of a single right-hand side variable, where the intuition is simpler. The regressor $z_{2i} = y_{2i}$ and the conditioning variables x_i are scalar. Allowing SDL to be conditional on the suppression indicator, we follow the derivations in the unpublished Appendix to Bollinger and Hirsch (2006) to obtain the following result:

$$\operatorname{plim} \hat{b} = \beta - (1-\rho)\,\mu\beta = (1 - (1-\rho)\,\mu)\beta. \tag{C.18}$$

The bias term on the right hand side depends on two factors: the share of suppressed observations, $(1-\rho)$, and the error from using x_i to impute the suppressed value instead of z_{2i}, measured by μ. The term μ may be derived as follows. First, compute the residual from predicting the regressor with the conditioning variables: $e_i = z_{2i} - E(z_{2i}|x_i, \gamma_i = 0)$. Now μ is the slope parameter from the regression $e_i = \ell + \mu z_{2i}$. That is, μ measures the signal from the regressor z_{2i} left in e_i after conditioning on x_i and $\gamma_i = 0$.

The same result holds for the more general case in which z_{2i} and x_i are vectors. Now

$$\operatorname{plim} \hat{b} = \beta - (1-\rho)\,M\beta = (I - (1-\rho)\,M)\beta, \tag{C.19}$$

Formally, M is derived analogously to μ. First, measure the vector of residuals from the system $e_i = z_{2i} - E(z_{2i}|x_i, \gamma_i = 0)$. Then, M is the parameter matrix from estimating of $e_i = L + Mz_{2i}$.

The case of general z_i is similar, but the derivations are complicated and provide little intuition for the applications under consideration here. They are available upon request.

C.2 Bias Due to SDL in a Single Regressor

The case in which SDL is applied to a single regressor turns out to be identical to the case of SDL applied to the dependent variable. That this is so may be intuitive when the regression model includes only one regressor. It is less transparent in the case of multiple regressors, so we present the relevant derivations here. For ease of presentation, we use notation similar to Bollinger and Hirsch.

The data vector is $(y_i, z_i, t_i, x_i, R_i)$. y_i is the dependent variable, t_i is the scalar variable to which SDL is applied, z_i is a vector of regressors that are not distorted, x_i is a vector of conditioning variables used to impute replacements for t when it is suppressed, and R_i is a variable equal to 1 if t_i is suppressed, and 0 otherwise. Define the population distributions $f_O\left(y_i, z_i, t_i, x_i|R_i = 0\right)$ for observations with no suppression, and $f_M\left(y_i, z_i, t_i, x_i|R_i = 1\right)$ for observations where t was suppressed and imputed. Also, let $p = \Pr\left[R_i = 1\right]$.

We make the following assumptions on the data generating process:

- Only t_i is suppressed;
- the matching variables depend only on z_i and t_i, $x_i = h\left(z_i, t_i\right);$
- the researcher has the correct model, and so x cannot provide any additional information:
$$E\left[y_i|z_i, t_i, x_i\right] = E\left[y_i|z_i, t_i\right] = \alpha + z_i^T\beta + \gamma t_i$$
- when $R_i = 1$, the published value t_i is sampled from the distribution $f_M\left(t|x_i\right)$.

The conditional distribution of the suppressed data is

$$f_I\left(y_i, z_i, t_i|x_i\right) = f_M(y_i, z_i|x_i)f_M\left(t_i|x_i\right).$$

It follows the distribution of the published data is

$$f_S\left(y_i, z_i, t_i|x_i\right) = \left(1-p\right)f_O\left(y_i, z_i, t_i|x_i\right) + pf_M(y_i, z_i|x_i)f_M\left(t_i|x_i\right).$$

After some algebraic transformations, and taking expectations with respect to z_i,t_i, and x_i, we get the key moment equation characterizing the conditional expectation of y_i given $\left(z_i, t_i, x_i\right)$ in the published data:

$$\begin{aligned} E_S\left[y_i|z_i, t_i, x_i\right] &= \left(1-p\right)E_O\left[y_i|z_i, t_i, x_i\right]\frac{f_O\left(z_i, t_i, x_i\right)}{f_S\left(z_i, t_i, x_i\right)} \\ &\quad + pE_M\left[y_i|z_i, x_i\right]\frac{f_M(z_i, x_i)f_M\left(t_i|x_i\right)}{f_S\left(z_i, t_i, x_i\right)}. \end{aligned}$$

Using the definition of $E_O\left[y_i|z_i, t_i, x_i\right]$ and adding and subtracting $p\gamma t\frac{f_M(z_i,x_i)f_M(t_i|x_i)}{f_S(z_i,t_i,x_i)}$,

$$\begin{aligned} E_S\left[y_i|z_i, t_i, x_i\right] &= \alpha + z_i^T\beta + \gamma t_i \\ &\quad - p\gamma\left[t - E_M\left(t_i|x_i\right)\right]\frac{f_M(z_i, x_i)f_M\left(t_i|x_i\right)}{f_S\left(z_i, t_i, x_i\right)}. \end{aligned}$$

We now show that the least-squares solution will not be consistent for the parameters of interest and derive the bias correction. In the published data, the least-squares solution to the regression of y_i on z_i and t_i is

$$\arg\min_{a,b,c} E_S\left[\left(E_s\left[y_i|z_i, t_i, x_i\right] - \left(a + z_i^Tb + ct_i\right)\right)^2\right];$$

that is

$$\arg\min_{a,b,c}\int\left(E_s\left[y_i|z_i, t_i, x_i\right] - \left(a + z_i^Tb + ct\right)\right)^2 f_s\left(z_i, t_i, x_i\right)dz_i dt_i dx_i.$$

The first-order conditions for a minimum are given by:

$$\begin{gathered} \alpha + E_S\left(z_i^T\right)\beta + \gamma E_s\left(t_i\right) - \gamma pE_I\left(t_i - E_M\left(t_i|x_i\right)\right) \\ -\left(a + E_S\left(z_i^T\right)b + cE_s\left(t_i\right)\right) = 0 \end{gathered}$$

from differentiating with respect to a,

$$\alpha E_S\left(z_i\right)+E_S\left(z_i z_i^T\right)\beta+\gamma E_s\left(z_i t_i\right)-\gamma p E_I\left(z_i\left(t_i-E_M\left(t_i|x_i\right)\right)\right) \\ -\left(a E_S\left(z_i\right)+E_S\left(z_i z_i^T\right)b+c E_s\left(z_i t_i\right)\right)=0$$

from differentiating with respect to b, and

$$\alpha E_S\left(t_i\right)+E_S\left(t_i z_i^T\right)\beta+\gamma E_s\left(t_i^2\right)-\gamma p E_I\left(t_i\left(t_i-E_M\left(t_i|x_i\right)\right)\right) \\ -\left(a E_S\left(t_i\right)+E_S\left(t_i z_i^T\right)b+c E_s\left(t_i^2\right)\right)=0$$

from differentiating with respect to c.

In the case where t_i is the only regressor (z is identically zero), it is easy to show

$$c=\gamma\left\{1-p\left[E_I\left[t_i\left(t_i-E_M\left(t_i|x_i\right)\right)\right]-E_I\left[t-E_M\left(t_i|x_i\right)\right]E_s\left(t_i\right)\right]\right\}.$$

By inspection, this is identical to the formula for the case in which SDL is applied to the dependent variable.

D Details of the RD Model

D.1 Generalized Randomized Response SDL

In our analysis of the effect of SDL on regression discontinuity designs, we consider the case in which the following model of SDL was applied to the running variable. The published data are

$$\begin{array}{rcl} \omega_i & = & w_i^* \\ z_{i3} & & \text{sampled from } p_{Z_3|Y_3}\left(z_{i3}\left|y_{i3},\theta_S\right.\right) \\ z_{i4} & = & 1\left[z_{i3}\geq\tau\right] \end{array}$$

with $p_{Z_3|Y_3}\left(z_{i3}\left|y_{i3},\theta_S\right.\right)$ given by the following mixture model, which is a generalization of randomized response. The randomization variable is $\gamma_i\sim\mathrm{Bin}\left(\rho,1\right)$. When $\gamma_i=1$, $z_{i3}=y_{i3}$; otherwise $z_{i3}=y_{i3}+\varepsilon_i$ with $\varepsilon_i\sim\mathrm{N}\left(0,\delta^2\right)$, (*i.e.*, additive noise infusion).

These assumptions imply

$$z_{i3}=\gamma_i y_{i3}+\left(1-\gamma_i\right)\left(y_{i3}+\varepsilon_i\right),$$

$$z_{i4}=\begin{cases}1\left[y_{i3}\geq\tau\right] \text{ if } \gamma_i=1\\ 1\left[y_{i3}+\varepsilon_i\geq\tau\right] \text{ if } \gamma_i=0\end{cases}$$

and

$$p_{Z_3Z_4|Y_3}\left(z_{i3},z_{i4}\left|y_{i3},\theta_S\right.\right)=\rho p_{Y_3Y_4}\left(Z_3,Z_4\left|\theta_p\right.\right)+\left(1-\rho\right)p_{Y_3Y_4}^*\left(Z_3,Z_4\left|\theta_p,\delta^2\right.\right),$$

where $p_{Y_3Y_4}^*\left(Z_3,Z_4\left|\theta_p,\delta^2\right.\right)$ is the distribution function from the convolution of $p_{Y_3Y_4}\left(Y_3,Y_4\left|\theta_p\right.\right)$ and $\mathrm{N}\left(0,\delta^2\right)$.

D.2 SDL Aware Analysis of the RD Model

Using the posterior predictive distribution for y_{i3} given z_{i3} and assuming that the SDL parameters are fixed at the known values ρ_0 and δ_0, we have

$$\mathrm{E}\left[y_{i3}\left|z_{i3},\rho_0,\delta_0\right.\right]=\mathrm{E}\left[z_{i3}-\left(1-\gamma_i\right)\varepsilon_i\left|z_{i3},\rho_0,\delta_0\right.\right]=z_{i3}$$

and

$$\begin{aligned}\mathrm{E}\left[y_{i4}\left|z_{i3},\rho_0,\delta_0\right.\right] &= \mathrm{E}\left[1\left[y_{i3}\geq\tau\right]\left|z_{i3},\rho_0,\delta_0\right.\right] \qquad \text{(D.20)}\\ &= \rho_0\,1\left[z_{i3}\geq\tau\right]+\left(1-\rho_0\right)\Phi\left(\frac{z_{i3}-\tau}{\delta_0}\right)\end{aligned}$$

where $\Phi\left(\right)$ is the standard normal cumulative distribution function. The SDL-aware analysis has converted the original sharp RD into a fuzzy RD. To complete the analysis we should use the posterior distribution of θ_{RD} given the published data Z and the SDL parameters, assumed known or with an informative prior given agency-provided data.

In the RD literature, functional form assumptions about $f_1\left(y_{i3}\right)$, $f_2\left(y_{i3}\right)$, and $\pounds_\theta^{(obs)}\left(\theta_p\left|Y^{(obs)}\right.\right)$ are minimized. Respecting this analysis style, without implying that it is the best way to analyze a finite sample of size n from a superpopulation with size N, we analyze a few posterior moments, making the assumption that those exist.

We want to estimate

$$\begin{aligned}\mathrm{E}\left[\theta_{RD}\left|Z,\rho_0,\delta_0\right.\right] &= \mathrm{E}\left[\lim_{y_{i3}\downarrow\tau}\mathrm{E}\left[y_{i2}\left|y_{i3}=\tau\right.\right]\left|Z,\rho_0,\delta_0\right.\right] \qquad \text{(D.21)}\\ &\quad -\mathrm{E}\left[\lim_{y_{i3}\uparrow\tau}\mathrm{E}\left[y_{i1}\left|y_{i3}=\tau\right.\right]\left|Z,\rho_0,\delta_0\right.\right] \qquad \text{(D.22)}\\ &= \mathrm{E}\left[\lim_{y_{i3}\downarrow\tau}f_2\left(y_{i3}\right)\left|Z,\rho_0,\delta_0\right.\right]-\mathrm{E}\left[\lim_{y_{i3}\uparrow\tau}f_1\left(y_{i3}\right)\left|Z,\rho_0,\delta_0\right.\right]\\ &= \rho_0\left\{\begin{array}{c}\mathrm{E}\left[\lim_{z_{i3}\downarrow\tau}f_2\left(z_{i3}\right)\left|Z,\gamma_i=1,\delta_0\right.\right]\\ -\mathrm{E}\left[\lim_{z_{i3}\uparrow\tau}f_1\left(z_{i3}\right)\left|Z,\gamma_i=1,\delta_0\right.\right]\end{array}\right\}\\ &+ \left(1-\rho_0\right)\left\{\begin{array}{c}\mathrm{E}\left[\lim_{z_{i3}\downarrow\tau}f_2\left(z_{i3}-\varepsilon_i\right)\left|Z,\gamma_i=0,\delta_0\right.\right]\\ -\mathrm{E}\left[\lim_{z_{i3}\uparrow\tau}f_1\left(z_{i3}-\varepsilon_i\right)\left|Z,\gamma_i=0,\delta_0\right.\right]\end{array}\right\}\\ &= \rho_0\left(\lim_{z_{i3}\downarrow\tau}f_2\left(\tau\right)-\lim_{z_{i3}\uparrow\tau}f_1\left(\tau\right)\right)\end{aligned}$$

and

$$\begin{aligned}\rho_0 &= \lim_{z_{i3}\downarrow\tau}\left[\rho_0\,1\left[z_{i3}\geq\tau\right]+\left(1-\rho_0\right)\Phi\left(\frac{z_{i3}-\tau}{\delta_0}\right)\right]\\ &\quad -\lim_{z_{i3}\uparrow\tau}\left[\rho_0\,1\left[z_{i3}\geq\tau\right]+\left(1-\rho_0\right)\Phi\left(\frac{z_{i3}-\tau}{\delta_0}\right)\right]\end{aligned}$$

The regime where $\gamma_i=1$ is a conventional RD. The existence of the regime $\gamma_i=0$ converts the problem to a fuzzy RD where $\mathrm{E}\left[y_{i4}\left|z_{i3},\rho_0,\delta_0\right.\right]=g\left(z_{i3}\right)$ plays the role of the "compliance status" function. The term

$$(1-\rho_0)\left\{\mathrm{E}\left[\lim_{z_{i3}\downarrow\tau} f_2\left(z_{i3}-\varepsilon_i\right)|Z,\gamma_i=0,\delta_0\right]-\mathrm{E}\left[\lim_{z_{i3}\uparrow\tau} f_1\left(z_{i3}-\varepsilon_i\right)|Z,\gamma_i=0,\delta_0\right]\right\} \tag{D.23}$$

is zero because $\varepsilon_i \sim \mathrm{N}\left(0,\delta^2\right)$ implies that in the regime $\gamma_i=0$, there is no point mass at $\varepsilon_i=0$; hence there is no jump at τ–the continuous function $f_1\left(z_{i3}\right)$ transitions smoothly to $f_2\left(z_{i3}\right)$ over the support of ε_i. The SDL noise needn't be normal, but it must be drawn from a continuous distribution.

D.2.1 Implications of SDL in the Running Variable for other RD Models

If generalized random response SDL is applied to the running variable, then the SDL is ignorable for parameter estimation when the true RD design is fuzzy. The FRD compliance function, augmented with the contribution from SDL, becomes

$$\begin{aligned} h(z_i) &= \mathrm{E}\left[t_i\,|z_i,\rho_0,\delta_0\right] && \text{(D.24)}\\ &= \rho_0 p_{T|R}\left(t_i=1|z_i\right)+(1-\rho_0)\int p_{T|R}(t_i=1|r_i)p_{R|Z}(r_i|z_i)dr. && \text{(D.25)}\end{aligned}$$

It immediately follows

$$\lim_{z_i\downarrow\tau} h\left(z_i\right)-\lim_{z_i\uparrow\tau} h\left(z_i\right)=\rho_0\left[\lim_{z_i\downarrow\tau} p_{T|R}\left(t_i=1|z_i\right)-\lim_{z_{i1}\uparrow\tau} p_{T|R}\left(t_i=1|z_i\right)\right].$$

The second summand in the expression for $h(z_i)$ is zero. When the running variable is distorted with normally distributed noise, there is no point mass anywhere, and hence no discontinuity in the probability of treatment at τ. The claim that the SDL is ignorable for consistent estimation of the treatment effect in the fuzzy RD design follows. Imbens and Lemieux (2008) show that the 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.

E Details for Tabular Methods

E.1 Swapped Household Data

This part of our discussion of tabular data that applies only to tabulations based on household data. The tables produced from the decennial censuses and the American Community Survey are based on swapped input data. The effects of swapping can be assessed using the methods we discussed in Technical Appendix Section B. The condition for discoverable consequences of swapping, without the cooperation of the data provider, requires getting at least two tabulations that cover the same subpopulation, have known sampling variation, and use independent SDL models. The best general diagnostic we can derive is to perform simulations under the assumption that the contribution to the posterior variance of a parameter of interest due to swapping is less than the contribution due to edit and imputation. If an agency were to state in its published documentation that this hypothesis was correct, then it might be worth unleashing the full posterior simulation technology.

E.2 Custom Tabulations

An agency's officially tabulated estimates are those listed in the defined data products for the agency's publications. If the tabulation isn't listed in the defined data products, then an official estimate of that item is called a custom tabulation. All custom tabulations (also called special tabulations) are done sequentially, then released to the general public. The suppression rules applied to the official tabulations carry over to the first custom tabulation, and then to all successive custom tabulations. The effects are order dependent and cumulative. If an item was explicitly suppressed from any previous official or custom tabulation, then it will be suppressed for all future tabulations as well. This statement applies to both primary and complementary suppressions. Some agencies will not produce custom tabulations as a matter of policy. It is also worth pointing out that not all suppressions are due to SDL. There are also minimum data quality standards that can result in a suppression. These do not always cause additional complementary suppressions, but they do always cumulate. The data quality cannot be improved by calling it a custom tabulation.

E.3 Directly Tabulating Published Microdata

After data collection has ended, the raw survey, census or administrative-record data are edited, imputed for missing data, weight-corrected, and subjected to SDL. Only then are the publication tables generated. The statistical agencies consider any released public-use microdata samples to also be publication tables. In general, if a researcher computes an estimate of a moment or quantile from the public-use microdata, then compares that estimate to its published equivalent in the tabular summaries, those two estimates will not agree exactly. Assuming that the correct selection criteria and weights were used, there remain three reasons why these calculations don't match. The first possible cause is differences in the computational formulas used. The second possible cause is sampling variability.

The third possible cause is SDL. The Census Bureau, for example, applies additional swapping and noise infusion to the publication-ready ACS records before selecting the PUMS. Public-use files produced by the Statistics of Income Division of the IRS are also subjected to extensive SDL beyond what is used for the tabular summaries.

By far, the most important SDL explanation for a discrepancy is that the equivalent tabular estimate is suppressed. A researcher can calculate some estimate of interest from the public-use microdata file, but the agency didn't release an official tabulation of the same item for several reasons. Some researchers may not consider suppressions in official tabulations to be a discrepancy with respect to estimates produced from the public-use microdata files, but there is an important sense in which they are. The microdata files are produced so that researchers can perform analyses that are not possible using the aggregated tabulations. If there were no microdata files, as we will see is usually the case for the aggregated business data discussed in Section 5.2.2, then any research design would require a strategy for handling the missing data caused by the suppression.

Even when public-use microdata are available, suppression is still a problem. If a substantial proportion of the estimates a researcher computes from the microdata files correspond to suppressed official tabulations, it is a warning sign that the inputs to the researcher's statistical model may be of poor quality. Household tabular estimates are suppressed most often when the number of households in the cell is below the publication threshold. The statistical agency considers that item to be poorly estimated in the underlying confidential data. Furthermore, these are exactly the cells most likely to contain edit, imputation, and SDL-induced noise. A good research strategy is to consider pooling those estimates, for example by using a shrinkage estimator that averages a specific moment with a pooled estimate of the same moment.

E.4 Tabular Regression Models with Nonignorable SDL

The noise infusion in QWI may be nonignorable. Univariate regression of a variable, say from another dataset, onto a QWI aggregate, provides a simple illustration. Suppose the part of the process model of interest is:

$$\mathrm{E}\left[Y_{(k)t}\left|W_{(k)t}\right.\right] = \alpha + \beta W_{(k)t} \tag{E.26}$$

where $W_{(k)t}$ is the quarterly payroll in county k and $Y_{(k)t}$ is any outcome of interest collected from a different data source. $Y_{(k)t}$ can also be subject to SDL, but we will assume that it is statistically independent of the SDL applied to the QWI data. The published aggregate data are $[Y_{(k)t}, W^*_{(k)t}]$. The undistorted values, $W_{(k)t}$ are confidential.

The probability limit of the OLS estimator for β based using the published data is

$$p\lim \hat{\beta}_{OLS} = \frac{\mathrm{Cov}\left[Y_{(k)t}, W_{(k)t}\right]}{\mathrm{Var}\left[\delta_j\right]\mathrm{E}\left[W^2_{(k)t}H^W_{(k)t}\right] + \mathrm{Var}\left[W_{(k)t}\right]} \tag{E.27}$$

The term $\mathrm{E}\left[W^2_{(k)t}H^W_{(k)t}\right]$ is the expected Herfindahl index for payroll within aggregate k, as derived in the Data Appendix E.5. The noise infusion is clearly nonignorable in this setting. Algebraic manipulation reveals the bias to be

$$p\lim \frac{\hat{\beta}_{OLS}}{\beta} = \frac{\mathrm{Var}\left[W_{(k)t}\right]}{\mathrm{Var}\left[\delta_j\right]\mathrm{E}\left[W^2_{(k)t}H^W_{(k)t}\right] + \mathrm{Var}\left[W_{(k)t}\right]}. \tag{E.28}$$

The bias factor lies between 0 and 1.

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

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, since

$$p\lim \hat{\beta}_{IV} = \frac{\mathrm{Cov}\left[\alpha + \beta W_{(k)t} + \varepsilon_{(k)t}, Z_{(k)t}\right]}{\mathrm{Cov}\left[W^{*}_{(k)t}, Z_{(k)t}\right]} = \frac{\beta\,\mathrm{Cov}\left[W_{(k)t}, Z_{(k)t}\right]}{\mathrm{Cov}\left[W_{(k)t}, Z_{(k)t}\right]} = \beta. \tag{E.29}$$

E.5 Details of Estimating the Variance Contribution of SDL for the QWI

It is possible to recover the variance of the noise factor $\mathrm{Var}\left[\delta_j\right]$, which is needed to correct directly for bias in the univariate and multivariate regression examples using the QWI. The noise in a magnitude estimate from a particular cell and the confidential magnitude value are independent by construction. By design, there is no bias:

$$\mathrm{E}\left[W^{*}_{(k)t} - W_{(k)t} \left| W_{(k)t}\right.\right] = \mathrm{E}\left[\sum_{j\in\Omega_{(k)t}} W_{jt}\left(\delta_j - 1\right)\left| W_{(k)t}\right.\right] = 0, \tag{E.30}$$

where the last equality results from the independence of W_{jt} and δ_j for all t. This is a common feature of noise-infusion SDL. The designers eliminated the bias in published tabulations. However, this was accomplished by inflating the variance of the published aggregate. The exact formula for the variance in the difference between noisy and noise-free estimates of is

$$\mathrm{V}\left[W^{*}_{(k)t} - W_{(k)t} \left| W_{(k)t}\right.\right] = \mathrm{V}\left[\delta\right] \sum_{j\in\Omega_{(k)t}} W^2_{jt}. \tag{E.31}$$

Our leverage in this analysis comes from the fact that QWI and QCEW use identical frames (QCEW establishments). Hence, we can use $W^{(QCEW)}_{(k)t}$ as the noise-free estimate of $W_{(k)t}$, as long as it has not been suppressed too often.

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. The formulas for recovering both systems SDL parameters are in the Data Appendix Section H.

Data Appendix for "Economic Analysis and Statistical Disclosure Limitation"

John M. Abowd	Ian M. Schmutte
Department of Economics	Department of Economics
Labor Dynamics Institute	Terry College of Business
Cornell University	University of Georgia
john.abowd@cornell.edu	schmutte@uga.edu

August 7, 2015

F Variables from the QWI, QCEW, and CBP

We work with several variables from the QWI data:

- $B_{jt} \equiv$ employment at establishment j at the beginning of quarter t (record-linkage definition; first calendar day of the quarter)
- $W_{jt} \equiv$ total quarterly payroll for all statutory employees during the quarter (state unemployment insurance system definition)

From the QCEW, we use the following variables, which are analogous to the employment and payroll variables reported in the QWI.

- $E_{jt}^{(QCEW,1)} \equiv$ month 1 employment (on the payroll for the pay period covering the 12^{th} day of the first month of the quarter)
- $E_{jt}^{(QCEW,3)} \equiv$ month 3 employment (on the payroll for the pay period covering the 12^{th} day of the third month of the quarter)
- $W_{jt}^{(QCEW)} \equiv$ total quarterly payroll for all statutory employees during the quarter (state unemployment insurance system definition)

From the CBP, we use the following variables, which are analogous to the employment and payroll variables in the QWI.

- $L_{jt} \equiv$ employment (on the payroll for the pay period covering the March 12^{th})
- $P_{jt} \equiv$ total first-quarter payroll for all statutory employees (Federal Insurance Contributions Act (FICA) definition)

G Statistical Disclosure Limitation Methods

G.1 QWI

The QWI SDL system is based on multiplicative input noise infusion applied to all variables used to compute tabular magnitude estimates. These include employment and payroll, of course, but also hires, separations, job creations, job destructions, and similar statistics for stocks and flows based on stable employment definitions.

A random fuzz factor, δ_j, is drawn for each establishment, j, from a double-ramp distribution with the following probability distribution function:

$$p\left(\delta_j\right) = \begin{cases} 0,\ \delta < 1-b \\ \left(1+b+\delta-2\right)/\left(b-a\right)^2,\ \delta \in [1-b, 1-a] \\ 0,\ \delta \in (1-a, 1+a) \\ \left(1+b-\delta\right)/\left(b-a\right)^2,\ \delta \in [1+a, 1+b] \\ 0,\ \delta > 1+b \end{cases} \tag{G.32}$$

and associated density

$$F\left(\delta_j\right) = \begin{cases} 0,\ \delta < 1-b \\ \left(\delta+b-1\right)^2/\left[2\left(b-a\right)^2\right],\ \delta \in [1-b, 1-a] \\ 0.5,\ \delta \in (1-a, 1+a) \\ 0.5+\left[\left(b-a\right)^2-\left(1+b-\delta\right)^2\right]/\left[2\left(b-a\right)^2\right],\ \delta \in [1+a, 1+b] \\ 1,\ \delta > 1+b \end{cases} \tag{G.33}$$

The values $0 < a < b < 1$ are parameters chosen such that each establishment's value of the statistic is distorted by a minimum of $100a$ percent and a maximum of $100b$ percent.[11] Thus δ_j has the following properties:

$$\mathrm{E}\left[\delta_j\right] = 1$$

and

$$\mathrm{V}\left[\delta_j\right] = a^2 + \frac{1}{6}\left(b-a\right)^2 + \frac{2}{3}a\left(b-a\right).$$

The probability distribution of δ_j is plotted in Figure H.1a on page 67 and the cumulative distribution is plotted in Figure H.1b on page 67 for the values $a = 0.05, b = 0.3$, which were chosen for illustrative purposes only. Note, the distribution of δ_j is independent of all other variables. The SDL system is implemented so that an establishment is assigned a value of δ_j at the time it first enters the database. The establishment retains the assigned value until it disappears from the data permanently.

G.2 SDL for the QCEW

The QCEW data use a primary/complementary suppression system for SDL. The only public information about this system appears in Statistical Policy Working Paper 22:

"For example, the Quarterly Census of Employment and Wages (QCEW), a census of monthly employment and quarterly wage information from Unemployment Insurance filings, uses a threshold rule and the p percent rule for calendar year (CY) 2002 data and beyond. Prior to CY 2002, QCEW used a threshold rule and a concentration rule of (n, k). In a few cases, a two-step rule is used–an (n, k) rule for a single establishment is followed

[11] The exact percentage distortions are Census Bureau confidential.

by an (n, k) rule for two establishments." (Harris-Kojetin et al. (2005), page 47)

The BLS quantifies the amount of suppression with the following statement:

"The finest level of geographic detail is the county-industry level, as aggregates of establishments classified to varying degrees of industry detail. While the input data are coded with meaningful address locations, the data are generally unavailable at greater detail. The QCEW program is constrained by the need to protect the confidentiality of data provided by employers, and richer geographic detail would threaten that confidentiality. Even the county by industry data cited above is at the margin of being disclosable-approximately 60 percent of the most detailed level data are suppressed for confidentiality reasons." (http://www.bls.gov/cew/cewfaq.htm)

The only public detail of the complementary suppression algorithm is that it does not include table margins (unless the margin itself fails the primary suppression rule):

"However, published totals of higher-level aggregations, when disclosed, include the suppressed lower-level data." (http://www.bls.gov/cew/cewfaq.htm)

The public QCEW data are not rounded.

G.3 SDL for County Business Patterns

CBP uses noise infusion that is similar in the cross-section to the method used by QWI. There are also primary suppressions when the number of establishments in a cell is deemed too small to allow publication and when the value of the cell was distorted by more than five percent. The official specification of the noise infusion system is quoted here from the CBP documentation.

"County Business Patterns continues to apply the Noise Infusion method of data protection that began in 2007. Noise infusion is a method of disclosure avoidance in which values for each establishment are perturbed prior to table creation by applying a random noise multiplier to the magnitude data (i.e., characteristics such as first-quarter payroll, annual payroll, and number of employees) for each company. Disclosure protection is accomplished in a manner that results in a relatively small change in the vast majority of cell values. Each published cell value has an associated noise flag, indicating the relative amount of distortion in the cell value resulting from the perturbation of the data for the contributors to the cell. The flag for 'low noise' (G) indicates the cell value was changed by less than 2 percent with the application of noise, and the flag for 'moderate noise' (H) indicates the value was changed by 2 percent or more but less than 5 percent. Cells that have been changed by 5 percent or more are suppressed from the published tables. Additionally, other cells in the table may be suppressed for additional protection from disclosure or because the quality of the data does not meet publication standards. Though some of these suppressed cells may be derived by subtraction, the results are not official and may differ substantially from the true estimate. The number of establishments in a particular tabulation cell is not considered a disclosure; therefore, this information may be released without the addition of protective noise." (http://www.census.gov/econ/cbp/methodology.htm, citing Evans et al. (1998)).

Tabular cell magnitudes for L_{jt} and P_{jt} in CBP are computed using a multiplicative fuzz factor from a ramp distribution as in equations (G.32) and (G.33) with confidential parameters. The factor $\delta_{jt}^{(CBP)}$ is drawn fresh for each establishment every year. The same fuzz factor is applied to all values from an establishment. Census uses a "balancing" algorithm to reduce the amount of noise in a particular cell of CBP. In this context, balancing means that the conditional distribution of $\delta_{jt}^{(CBP)}$ is not independent of the values of L_{jt} or P_{jt} (depending upon which variable has been used to balance, which is not disclosed). CBP also uses non-standard establishment level rounding to ensure the protection of noise infusion for cells with a small number of small establishments. Employment size class distributions are tabulated from unfuzzed employment data.

The published CBP data on payroll are rounded to the nearest thousand dollars, which is also an SDL. Published employment is not further rounded from the record-level edits.

H Discovery of SDL parameters in QWI data

It is the differences in data construction that give rise to our strategy for revealing features of the SDL applied in each source. Therefore, it is necessary to discuss in some detail how each data source constructs and reports aggregate summaries from the underlying microdata.

QWI variable construction: Aggregates are formed over a classification $k = 1, \ldots, K$ that partitions the universe of establishments Ω_t into K mutually exclusive and exhaustive subsets $\Omega_{(k)t}$. These partitions usually have detailed geographic and industrial dimensions. For all three data sources, geography is coded using FIPS county codes. Industrial classifications are by NAICS sectors, sub-sectors, and industry groups.

The tabular magnitudes are computed by aggregating the values over the establishments in the partition k. In the QWI, total private employment in a state is benchmarked to the month-1 employment from QCEW using an establishment weight, ω_{jt}. In the absence of SDL, the beginning-of-quarter employment in k would be estimated by

$$B_{(k)t} = \sum_{j \in \Omega_{(k)t}} \omega_{jt} B_{jt}.$$

The QCEW does not use weights. The comparable employment magnitudes for months 1 and 3 are

$$E_{(k)t}^{(QCEW,1)} = \sum_{j \in \Omega_{(k)t}} E_{jt}^{(QCEW,1)}$$

and similarly for $E_{(k)t}^{(QCEW,3)}$.

SDL through multiplicative noise infusion: Published aggregates from the QWI are computed using the multiplicative noise factors δ_j. Beginning-of-quarter employment is computed as

$$B^*_{(k)t} = \sum_{j \in \Omega_{(k)t}} \delta_j \omega_{jt} B_{jt},$$

where we have adopted the convention of tagging the post-SDL value with an asterisk.

Similarly, the unprotected and protected values of total payroll in QWI are computed as

$$W_{(k)t} = \sum_{j \in \Omega_{(k)t}} \omega_{jt} W_{jt}$$

and

$$W^*_{(k)t} = \sum_{j \in \Omega_{(k)t}} \delta_j \omega_{jt} W_{jt}.$$

Notice that the same weight and the same fuzz-factor are used to aggregate total payroll and beginning-of-quarter employment (and, in fact, for all of the QWI).

QCEW variable construction: The total payroll variable in the QCEW is computed as

$$W^{(QCEW)}_{(k)t} = \sum_{j \in \Omega_{(k)t}} W^{(QCEW)}_{jt}.$$

To implement the SDL system for the QCEW, order statistics for the employment and payroll variables from the establishments in $\Omega_{(k)t}$ are used to compute the p-percent primary suppression rule. The partition size, $\left|\Omega_{(k)t}\right|$, is used to compute cell size thresholds, when they are used for suppression. As noted above, the formulas for the complementary suppressions have not been published.

Comparability of QWI and QCEW data: If the only partition is geography at the state level, so k indexes states, then the QWI benchmarking ensures that

$$B_{(k)t} = E^{(QCEW,1)}_{(k)t}. \tag{H.34}$$

We note for clarity that equation (H.34) holds only for beginning-of-quarter employment at the state level for all private employers, and not for any other variable or aggregation. In addition, the benchmarking is performed using the confidential inputs before SDL; hence, it does not hold exactly for the published values.

CBP data are not weighted. Hence, the published values are computed as

$$L^*_{(k)t} = \sum_{j \in \Omega_{(k)t}} \delta^{(CBP)}_{jt} L_{jt}$$

and

$$P^*_{(k)t} = \sum_{j \in \Omega_{(k)t}} \delta^{(CBP)}_{jt} P_{jt}.$$

$\left|\Omega_{(k)t}\right|$ is used to compute cell sizes for primary suppressions. The criteria for suppression based on "data quality" have not been published. There are no complementary suppressions. When computing the bins for the employment size distribution of establishments, L_{jt} is used without fuzzing.[12]

[12] Some details of the CBP protections are taken from an non-confidential presentation by Richard Moore, 2010, available from the authors.

H.1 Estimating the Variance Contribution of SDL for the QWI

In QWI, the variance in the difference between noisy (published) and noise-free estimates of start-of-quarter employment, $B_{(k)t}$, conditional on the noise-free estimates, is

$$\begin{aligned} \mathrm{V}\left[B^*_{(k)t} - B_{(k)t} \left| B_{(k)t}\right.\right] &= \mathrm{E}\left[\sum_{j\in\Omega_{(k)t}} \omega_{jt}\left(B_{jt}\left(\delta_j - 1\right)\right)^2 \left| B_{(k)t}\right.\right] \\ &= \mathrm{V}\left[\delta\right] \sum_{j\in\Omega_{(k)t}} \omega_{jt} B^2_{jt} \end{aligned} \tag{H.35}$$

with a similar formula for $W_{(k)t}$.

If information about the properties of the size distribution of employment are available for the classification represented by Ω, then equation (H.35) can be re-expressed as

$$\mathrm{V}\left[B^*_{(k)t} - B_{(k)t} \left| B_{(k)t}\right.\right] = \mathrm{V}\left[\delta\right] B^2_{(k)t} H^{(B)}_{(k)t}$$

where $H^{(B)}_{(k)t}$ is the Herfindahl index of employment shares of establishments in category k. It is straightforward to derive an equivalent equation for the variance of the difference between the published and noise-free payroll totals, conditional on the noise-free level, which depends on $H^{(W)}_{(k)t}$, the Herfindahl index of payroll shares of establishments in category k. Dividing both sides of Equation (H.35) by the square of the noise-free estimate and taking positive square roots yields

$$\sqrt{\frac{\mathrm{V}\left[B^*_{(k)t} - B_{(k)t} \left| B_{(k)t}\right.\right]}{B^2_{(k)t}}} \equiv \mathrm{CV}\left[B^*_{(k)t} - B_{(k)t} \left| B_{(k)t}\right.\right] \tag{H.36}$$

$$= \sqrt{\mathrm{V}\left[\delta\right] \frac{\sum_{j\in\Omega_{(k)t}} \omega_{jt} B^2_{jt}}{B^2_{(k)t}}} \tag{H.37}$$

$$= \sqrt{\mathrm{V}\left[\delta\right] H^{(B)}_{(k)t}}$$

with a similar formula for $W^*_{(k)t}$.

H.1.1 Empirical Specification

Taking logarithms of Equation (H.36) yields the estimating equation

$$\ln \mathrm{CV}\left[B^*_{(k)t} - B_{(k)t} \left| B_{(k)t}\right.\right] = \frac{1}{2} \ln \mathrm{V}\left[\delta\right] + \ln \sqrt{H^{(B)}_{(k)t}}. \tag{H.38}$$

The dependent variable is defined in terms of the noise-free estimates, which are confidential in the QWI system. Fortunately, we have access to noise-free variables from the published QCEW and CBP data. We assume

$$\mathrm{CV}\left[B^*_{(k)t} - B_{(k)t}\,\middle|B_{(k)t}\right] = \mathrm{CV}\left[B^*_{(k)t} - E^{(QCEW,1)}_{(k)t}\,\middle|E^{(QCEW,1)}_{(k)t}\right]$$

and

$$\mathrm{CV}\left[W^*_{(k)t} - W_{(k)t}\,\middle|W_{(k)t}\right] = \mathrm{CV}\left[W^*_{(k)t} - W^{(QCEW)}_{(k)t}\,\middle|W^{(QCEW)}_{(k)t}\right].$$

Both assumptions are justified by the fact that the noise factors in the QWI are completely independent of the underlying data. Furthermore, all three sources use identical frames and identical input data sources to measure the same variables in the same manner. The concepts used to measure $B_{(k)t}$ in QWI were constructed to approximate as closely as possible first month employment in the QCEW. Furthermore, the QWI establishment weights force the private state-level aggregate $B_{(k)t}$ to match exactly its QCEW counterpart.

The key explanatory variable in Equation H.38 is based on the Herfindahl index over employment shares, $H^{(B)}_{(k)t}$. This can be computed directly when size class information about the distribution within $B_{(k)t}$ and $W_{(k)t}$ is available, as is the case in the CBP data. Alternatively, we model this term as a power law (Cobb-Douglas) function of the number of establishments used to form the cell k

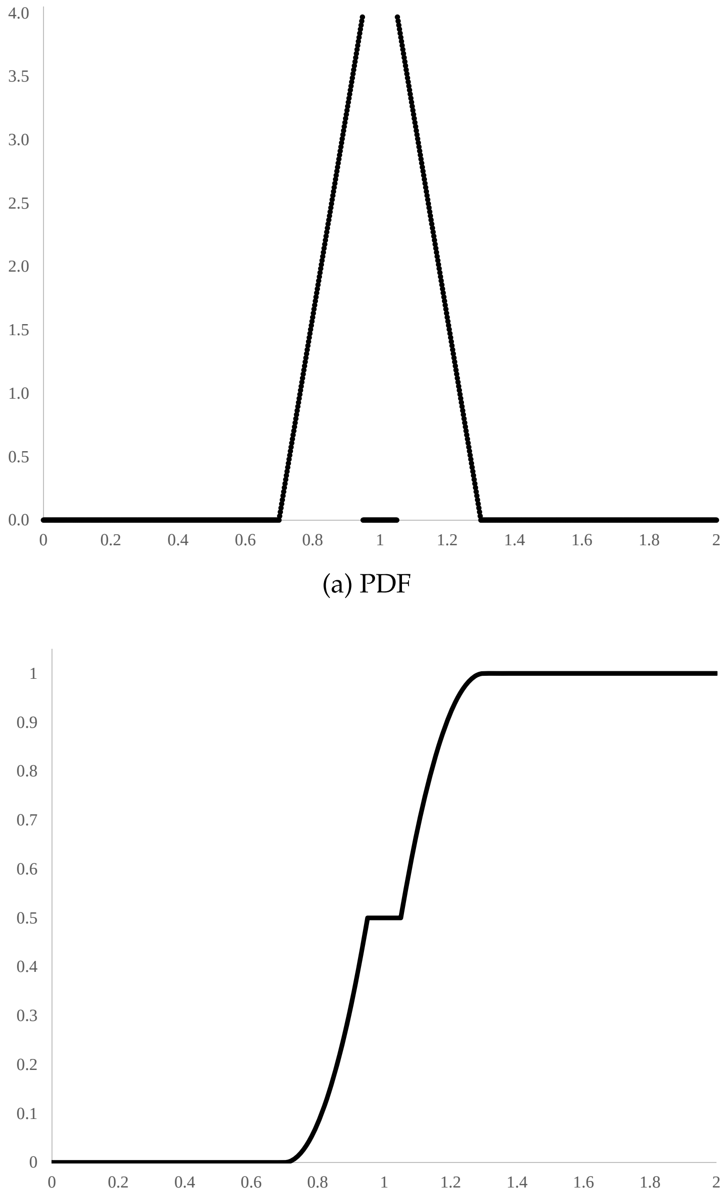
$$H^{(B)}_{(k)t} = \frac{\sum_{j\in\Omega_{(k)t}} \omega_{jt} B^2_{jt}}{B^2_{(k)t}} = \alpha_{(k)} N^{\beta_{(k)}}_{(k)t}, \tag{H.39}$$

where the scaling coefficient $\alpha_{(k)}$ is a potential confounder for the estimation of $\mathrm{V}[\delta]$.

When no size-class information is available, substitution of (H.39) gives the estimating equation:

$$\ln \mathrm{CV}\left[B^*_{(k)t} - E^{(QCEW,1)}_{(k)t}\,\middle|E^{(QCEW,1)}_{(k)t}\right] = \frac{1}{2}\ln \mathrm{V}[\delta] + \ln \alpha_{(k)} + \frac{\beta_{(k)}}{2}\ln N_{(k)t}. \tag{H.40}$$

Equation (H.40) is a smooth function of the logarithm of the number of establishments used to form the table cell with estimates $B^*_{(k)t}$ and $W^*_{(k)t}$. Data on the number of establishments in each cell is reported in the QCEW. The derivation of the estimating equation for the coefficient of variation in payroll is identical.



(a) PDF

(b) CDF

Figure H.1: Distribution of Establishment Noise for the Quarterly Workforce Indicators