Journal of Privacy and Confidentiality
Vol. 9 (2) 2019

Submitted Mar 8, 2019
Published Oct 2019

A PRACTICAL METHOD TO REDUCE PRIVACY LOSS WHEN DISCLOSING STATISTICS BASED ON SMALL SAMPLES

RAJ CHETTY AND JOHN N. FRIEDMAN

Harvard University and NBER

Brown University and NBER
e-mail address: john_friedman@brown.edu

Abstract. We develop a simple method to reduce privacy loss when disclosing statistics such as OLS regression estimates based on samples with small numbers of observations. We focus on the case where the dataset can be broken into many groups ("cells") and one is interested in releasing statistics for one or more of these cells. Building on ideas from the differential privacy literature, we add noise to the statistic of interest in proportion to the statistic's *maximum observed sensitivity*, defined as the maximum change in the statistic from adding or removing a single observation across all the cells in the data. Intuitively, our approach permits the release of statistics in arbitrarily small samples by adding sufficient noise to the estimates to protect privacy. Although our method does not offer a formal privacy guarantee, it generally outperforms widely used methods of disclosure limitation such as count-based cell suppression both in terms of privacy loss and statistical bias. We illustrate how the method can be implemented by discussing how it was used to release estimates of social mobility by Census tract in the Opportunity Atlas. We also provide a step-by-step guide and illustrative Stata code to implement our approach.

1. Introduction

Social scientists increasingly use confidential data held by government agencies or private firms to publish statistics based on small samples, from descriptive statistics on income distributions and health expenditures in small areas (e.g., Cooper et al. 2018, Chetty et al. 2018) to estimates of the causal effects of specific schools and hospitals (e.g., Angrist, Pathak

Key words and phrases: Differential Privacy, Administrative Data.

We thank John Abowd, Cynthia Dwork, Simson Garfinkel, Nathaniel Hendren, James Honaker, Adam Smith, Salil Vadhan, anonymous referees, and the Harvard University Privacy Tools Project for valuable feedback, as well as Federico Gonzalez, Jamie Gracie, Martin Koenen, Kate Musen, Daniel Reuter, Jesse Silbert, Nicolaj Thor, and our other pre-doctoral fellows at Opportunity Insights for outstanding research assistance. We are grateful for funding from the Chan-Zuckerberg Initiative, the Bill and Melinda Gates Foundation, the Overdeck Foundation, and Harvard University. A shorter, earlier version of this manuscript was published in the *American Economic Review Papers and Proceedings* as Chetty and Friedman (2019*b*). That version is Copyright, American Economic Association; sections that overlap are reproduced with permission. Illustrative code for the algorithm (along with a replication archive) is available on GitHub (see also Chetty and Friedman, 2019*a*).

JPrivacy Confidentiality

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.716

and Walters 2013, Hull 2018). Such statistics allow researchers and policymakers to answer important questions. But releasing such statistics also raises concerns about privacy loss – the disclosure of information about a specific individual – which can undermine public trust and is typically prohibited by law in government agencies and user agreements in the private sector.

In this paper, we develop a simple method to reduce privacy loss when disclosing statistics such as OLS regression estimates based on small samples. We add noise to the statistic of interest that is inversely proportional to the number of observations in the sample, choosing the amount of noise that is added based on how much the statistic of interest changes when one includes or excludes a single observation. Intuitively, our approach permits the release of statistics in arbitrarily small samples by adding sufficient noise to the estimates to protect privacy. We discuss an application of our noise-infusion method to releasing Census tract-level estimates of social mobility and present a step-by-step guide and code for implementing the method in other settings.

Currently, the most widely applied approaches to limiting such disclosure risks in social science are cell suppression (omitting data for small cells) and data swapping (switching individual values across cells). These techniques are simple to understand and are practical in the sense that they are almost universally applicable to any statistic of interest. Unfortunately, they remain prone to divulging information about specific individuals (e.g., Abowd and Schmutte 2019). For example, even when one suppresses cells with a count of fewer than say 100 individuals, one could in principle recover a single individual's income by releasing a mean over 150 individuals and a mean over 151 individuals and differencing the two statistics. Such concerns are not merely theoretical: one can reconstruct individual data with surprising accuracy from tables released by the U.S. Census Bureau that employed traditional disclosure avoidance methods (Garfinkel, Abowd and Martindale, 2018).

The recent literature on differential privacy, initiated in seminal work by Dwork (2006) and Dwork et al. (2006*a*),[1] provides a path to solving this problem by developing metrics for the privacy loss associated with the release of a statistic that can be held below a desired risk tolerance threshold. This literature has developed straightforward methods to protect privacy for simple statistics such as means and counts by adding noise to the estimates (e.g., Dwork et al. 2006*b*, McSherry and Talwar 2007, Dwork 2006, Kasiviswanathan et al. 2011). However, methods to protect privacy when disclosing other parameters – such as regression coefficients or quasi-experimental estimators – are considerably more complex, often relying on either asymptotic results in large samples (e.g., Blum et al. 2005, Smith 2011, Chaudhuri, Monteleoni and Sarwate 2011, Kifer, Smith and Thakurta 2012) or the use of robust statistics such as median regression (e.g., Nissim, Raskhodnikova and Smith 2007, Dwork and Lei 2009), limiting their application in social science.

Here, we build on ideas from the differential privacy literature to develop a method of reducing the privacy loss from disclosing arbitrarily complex statistics in small samples. Our approach combines some of the advantages of the differential privacy approach while retaining the practical benefits of traditional approaches such as cell suppression. In particular, the differential privacy literature generally focuses on developing mechanisms that are "provably private" in the sense of offering well-defined (probabilistic) guarantees about the risk of disclosing information about a single individual. We pursue a less ambitious goal. Rather than attempting to develop a provably private algorithm, we propose a method that

[1] See also Dwork et al. (2017).

outperforms existing methods of disclosure limitation such as cell suppression both in terms of privacy loss and statistical bias.

For concreteness, we focus on the problem of releasing estimates from univariate ordinary least squares (OLS) regressions estimated in small samples (e.g., small geographic units). We consider the case where the dataset can be broken into many groups ("cells") and one is interested in releasing statistics for one or more of these cells. For example, we may be interested in disclosing the predicted values from a regression of children's income percentile ranks in adulthood on their parents' income ranks in each Census tract in the U.S. Following the differential privacy literature, we add noise to each regression estimate that is proportional to the *sensitivity* of the estimate, defined as the impact of changing a single observation on the statistic. Intuitively, if a statistic is very sensitive to a single observation, one needs to add more noise to keep the likelihood of disclosing a single person's data below a given risk tolerance threshold.

The key technical challenge is determining the sensitivity of the regression estimates. The most common approach in the differential privacy literature is to measure the *global* sensitivity of the statistic by computing the maximum amount a regression estimate could change when a single observation is added or removed for any possible realization of the data. The advantage of this approach is that the actual data are not used to compute sensitivity, permitting formal guarantees about the degree of privacy loss. The problem is that in practice, the global sensitivity of regression estimates is infinite: one can always formulate a dataset (intuitively, with sufficiently little variance in the independent variable) such that the addition of a single observation will change the estimate by an arbitrarily large amount. As a result, respecting global sensitivity effectively calls for adding an infinite amount of noise and hence does not provide a path forward to disclose standard OLS regression estimates.

At the other extreme, one can compute the *local* sensitivity of a regression statistic as the maximum amount a regression estimate changes when a single observation is added or removed from the actual data in a given sample. While this is a finite value, the problem with this approach is that releasing the local sensitivity of statistics may *itself* release confidential information. Intuitively, local sensitivity is itself a statistic computed in a small sample and thus reveals some information about the underlying data.[2]

Our approach to computing sensitivity is a hybrid that lies between local and global sensitivity. We calculate local sensitivity in each cell (e.g., each Census tract) and then define the *maximum observed sensitivity* (MOS) of the statistic as the maximum of the local sensitivities across all cells (e.g. across all tracts in a given state), adjusting for differences in the number of observations across cells.[3] Drawing on results from the differential privacy literature, we show that by adding noise proportional to the MOS, one can guarantee that the privacy loss from releasing the cell-specific statistics (e.g., regression estimates) themselves falls below any desired exogenously specified risk tolerance threshold ε. The only uncontrolled privacy risk comes from the release of the MOS parameter (a single number), which is disclosed without noise and hence reveals information about the underlying data

[2] For example, outliers may greatly affect local sensitivity and hence the disclosure of local sensitivity can reveal information about the presence of outliers. See Section 3 and Figure 1 below for an illustration and further discussion of these issues.

[3] When one is interested in releasing an estimate for a single cell (e.g., a quasi-experimental estimate based on policy changes in a single school), one can construct "placebo" estimates by pretending that similar changes occurred in other cells (other schools) and then following the same approach to compute the MOS. See Step 1c of the implementation guide in the Appendix for further details.

that has unknown privacy risk. Importantly, however, we can compute the MOS in a sufficiently large sample that the disclosure risk from releasing it is likely to be negligible. For example, the Census Bureau's Disclosure Review Board has adopted the interim policy of not requiring additional noise infusion for statistics based on populations at least as large as the smallest state, based on the rationale that the number of individuals in such groups is large enough that it is unlikely one could identify a single person using typical statistics.[4]

We illustrate how the method can be implemented by discussing how we used it to produce the Opportunity Atlas, which provides public estimates of children's long-term outcomes by the tract in which they grew up (Chetty et al. 2018). We reduced the sensitivity of the statistics we released through procedures such as bounding variables and winsorization. We then chose the privacy threshold ε by following Abowd and Schmutte (2019) and weighing the privacy losses of a higher ε against the social benefits, which we defined as providing more accurate information to a family seeking to move to a higher-opportunity neighborhood. Ultimately, the noise we added to the estimates to protect privacy was smaller than the sampling error inherent in the estimates themselves and hence did not affect the precision of the statistics significantly. The tract-level estimates released using this approach have been viewed by half-a-million users, are currently being used to inform moving-to-opportunity housing voucher policies by housing authorities, and have been used as inputs by other researchers in downstream analyses (e.g., Morris, Gregory and Hartley 2018). The Opportunity Atlas thus provides a large-scale, real-world demonstration that our approach can be used to construct statistics from confidential data that provide useful information for social science and policy applications while limiting privacy risk.

Our approach outperforms the most popular disclosure limitation protocol that social scientists currently use (suppression of cells based on small counts) both in terms of reducing privacy loss *and* statistical bias.[5] In terms of privacy loss, it is straightforward to show that cell suppression has infinite (uncontrolled) privacy risk. As discussed above, even if one suppresses cells with counts below some threshold, one can recover information about a single individual by releasing statistics (e.g., means) from adjacent datasets that differ by a single observation.[6] In contrast, our noise-infusion approach would yield only probabilistic information about the additional observation, with a probability that is controlled by the choice of the risk tolerance threshold ε. Our approach reduces the dimensionality of the statistics that create uncontrolled privacy risks to a single number (the MOS parameter) that can be estimated in large samples, thereby significantly reducing the scope for privacy loss.

We demonstrate the benefits of our noise infusion approach in terms of statistical bias using an example from the Opportunity Atlas. Using noise-infused tract-level data, Chetty et al. (2018) show that black women who grow up in Census tracts with more single parents have significantly higher teenage birth rates. If one were to instead conduct their analysis

[4] Of course, this logic cannot be uniformly applied to all statistics; for instance, if one were to release the maximum income observed in a given state, one might be able to identify the person whose income is being reported. Nevertheless, for typical statistics such as means or medians of bounded variables, there is a common intuition – though no formal proof – that the privacy risks in large samples are generally small enough to be ignored.

[5] We focus on comparisons to count-based suppression mechanisms, but similar points apply to data swapping as well (Alexander, Davern and Stevenson 2010, Abowd and Schmutte 2015).

[6] Of course it is theoretically possible to prevent such releases by tracking every release and the exact sample used to generate it, but in practice this would be very difficult given the broad uses of many administrative data sets.

suppressing cells where where very few (less than 5) teenage births occur – a common approach to limit disclosure risk for rare outcomes – this strong relationship would vanish and the correlation would be zero. This is because the suppression rule leads to non-random missing data by excluding cells with low teenage birth rates (as pointed out more generally by Abowd and Schmutte (2015)). In short, count suppression would have led Chetty et al. (2018) to miss the relationship between teenage birth rates and single parent shares, illustrating how our algorithm outperforms existing approaches not just in principle but in practical applications of current interest to social scientists.

The rest of this paper is organized as follows. The next section sets up the problem and defines the key concepts. Section 3 describes our noise infusion method, both in general terms and in the application to the Opportunity Atlas. Section 4 shows how our noise infusion method outperforms cell suppression methods. Section 5 concludes. A step-by-step guide to implementing the method (along with illustrative Stata code) is provided in Appendix A.

2. The Problem

Our goal is to disclose a statistic θ that is a scalar estimated using a small number of observations in a confidential dataset while minimizing the risk of privacy loss. Although our approach can be applied to any statistic, we focus for concreteness on the problem of releasing predicted values from univariate regressions that are estimated in subgroups of the data, indexed by g:

$$y_{ig} = \alpha_g + \beta_g x_{ig} + \nu_{ig}.$$

For example, Chetty et al. (2018) regress children's income ranks in adulthood (y) on their parents' income ranks (x) by Census tract g. They then seek to release the predicted values from these regressions at the 25th percentile of the parent income distribution $\theta_g = \alpha_g + 0.25 \times \beta_g$ in their Opportunity Atlas. Because each Census tract contains relatively few observations, releasing $\{\theta_g\}$ raises concerns about preserving the privacy of the underlying individual data.

Noise Infusion. One intuitive way to reduce the risk of privacy loss is to add noise to the estimates $\{\theta_g\}$. An attractive feature of this approach is that the privacy loss from publishing noise-infused statistics can be quantified and thereby controlled below desired levels (Dwork et al. 2006*a* and Wasserman and Zhou 2010). To see this, let $\tilde{\theta}_g = \theta_g + \omega_g$ denote the noise-infused statistic, where ω_g is an independently and identically distributed draw from distribution $F(\omega)$, so that the conditional distribution of $\tilde{\theta}_g$ given θ_g is $F\left(\tilde{\theta}_g - \theta_g\right)$. Let $D_g = \{x_{ig}, y_{ig}\}$ denote the empirically observed data in cell g and $\mathbb{D}_g$ denote the set of theoretically possible datasets for tract g. The privacy loss from disclosing $\tilde{\theta}_g$ can be measured using the log likelihood ratio

$$\log \frac{f(\tilde{\theta}_g - \theta_g(D_g^1))}{f(\tilde{\theta}_g - \theta_g(D_g^2))}, \tag{2.1}$$

where $D_g^1, D_g^2 \in \mathbb{D}_g$ are two adjacent datasets (i.e., differ by only one observation) and $f()$ denotes the density of $F(\omega)$. Intuitively, this ratio measures the likelihood that the published statistic $\tilde{\theta}_g$ stems from underlying dataset D_g^1, relative to D_g^2; from a Bayesian perspective, the larger this ratio (in absolute value), the more one could update one's priors between D_g^1 and D_g^2 given the release of statistic $\tilde{\theta}_g$.

When no noise is infused (i.e., $Var(\omega_g) = 0$), this likelihood ratio will be infinite (except in the knife-edge case where two datasets produce exactly the same value of θ_g), as one could perfectly distinguish between any two datasets D_g^1 and D_g^2 that do not happen to produce exactly the same value of θ_g. As the noise variance increases, the likelihood ratio falls, and it becomes more difficult to determine whether the published statistic results from one dataset or another.

Differential Privacy. Modern privacy mechanisms limit privacy loss by placing an upper bound on the likelihood ratio in (2.1), effectively providing a "worst case" guarantee on the degree of privacy loss. In the terminology introduced by Dwork (2006) and Dwork et al. (2006*a*), a privacy algorithm is "ε-differentially private" if

$$\log \frac{f(\tilde{\theta}_g - \theta_g(D_g^1))}{f(\tilde{\theta}_g - \theta_g(D_g^2))} < \varepsilon \quad \forall\, D_g^1, D_g^2 \in \mathbb{D}_{\mathrm{g}}\,, \forall\, \tilde{\theta}_g \in \mathbb{R}. \tag{2.2}$$

The parameter ε can be interpreted as the maximum risk one is willing to tolerate when releasing the statistic of interest. If one uses a mean-zero Laplace distribution for noise ω_g (with density $l(\omega; 0, b) = \frac{1}{2b} \exp\left[-\frac{|\omega|}{b}\right]$), where ω_g is independent of the statistic of interest, the log-likelihood ratio is

$$\log \frac{f(\tilde{\theta}_g - \theta_g(D_g^1))}{f(\tilde{\theta}_g - \theta_g(D_g^2))} = \frac{\left|\tilde{\theta}_g - \theta_g(D_g^2)\right| - \left|\tilde{\theta}_g - \theta_g(D_g^1)\right|}{b}.$$

It follows that one can achieve the desired bound from (2.2) by setting the Laplace scale parameter $b = \frac{\Delta\theta_g}{\varepsilon}$, where

$$\Delta\theta_g = \max_{D_g^1, D_g^2 \in \mathbb{D}_{\mathrm{g}}} \left|\theta_g\left(D_g^1\right) - \theta_g\left(D_g^2\right)\right|$$

is the "sensitivity" of the statistic θ_g (Dwork et al. 2006*a*). Sensitivity measures the maximum amount that the statistic can change between any two adjacent datasets. When sensitivity is higher – that is when changing a single observation changes θ_g more – one must add more noise to prevent people from distinguishing one dataset from another. To see the intuition, consider releasing the mean wealth for a small group of households. If a very wealthy individual is potentially in that small group, the inclusion or exclusion of her data could change the reported mean substantially. One must therefore infuse a large amount of noise to protect her privacy when releasing statistics on mean wealth. In contrast, if one seeks to release the mean education in a group, there is less scope for outliers and hence the inclusion or exclusion of any one individual is unlikely to have a significant impact (i.e., sensitivity is low). In this case, privacy loss can be limited by adding a modest amount of noise.

If sensitivity $\Delta\theta_g$ were publicly known, one could obtain differentially private statistics that satisfy any exogenously specified privacy loss threshold ε simply by adding noise

$$\omega_g \sim L\left(0, \frac{\Delta\theta_g}{\varepsilon}\right). \tag{2.3}$$

to the statistics one seeks to release.[7] In practice, $\Delta\theta_g$ is not known ex-ante, as it depends on the particular values in the dataset D_g; hence, the key remaining question is how it should be calculated.

Global Sensitivity. The standard approach to measuring $\Delta\theta_g$ in the differential privacy literature is to calculate *global* sensitivity, the maximum amount a statistic can change under any theoretically possible configuration of the data. For instance, consider releasing the mean of N observations that are bounded between 0 and 1. The most that this statistic can change by changing a single observation in the data is by replacing a value of 0 with a value of 1 (or the reverse), thereby changing the mean by $\frac{1}{N}$. Hence, global sensitivity is $\frac{1}{N}$ in this case. Since this computation of global sensitivity does not rely on the actual data, it can be released publicly along with the statistic $\tilde{\theta}_g$ without any further privacy loss, yielding a fully differentially private disclosure mechanism. Researchers have applied this global-sensitivity approach to release simple statistics such as counts and means (Dwork et al. 2006*a*); indeed, the privacy protection plan for tabular data publications from the 2020 Decennial Census uses such methods (U.S. Census Bureau 2018).

Unfortunately, global sensitivity is typically infinite for OLS regression estimates and many other statistics of interest to social scientists. To see this in our setting, consider the limiting case where $Var(x_{ig})$ approaches 0 (e.g., all parents in a given cell have virtually the same income). In this case, the slope of the regression line (and therefore the predicted value θ_g) grows arbitrarily large. Adding a single value (x, y) to the dataset that is sufficiently far from the estimated regression line could therefore have an arbitrarily large effect on the statistic of interest, as illustrated in Appendix Figure 1. Thus, global sensitivity is infinite, implying that adding any finite amount of noise will not meet the differential privacy guarantee in (2.2).

In summary, standard methods of computing global sensitivity in the differential privacy literature do not provide a straightforward way to disclose many statistics of interest to social scientists. We propose an alternative approach to computing sensitivity in the next section.

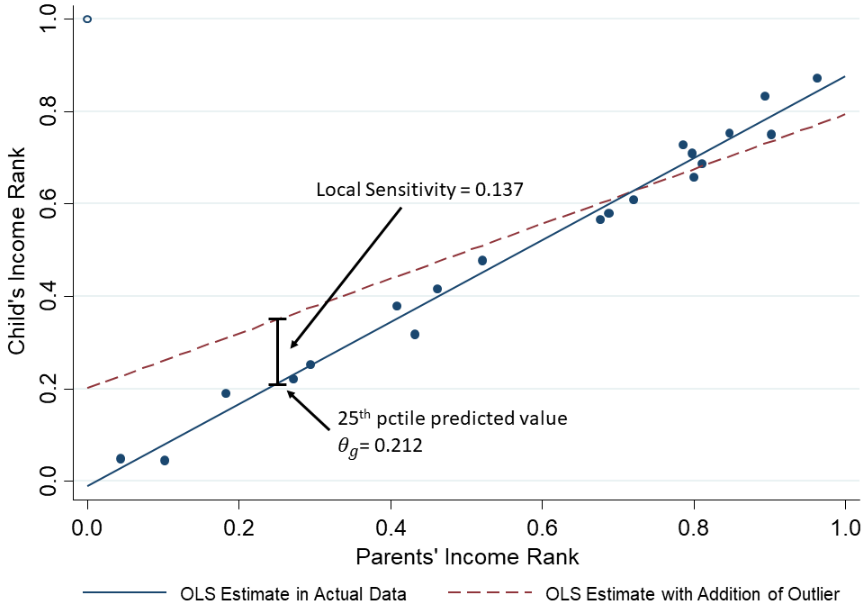
3. Maximum Observed Sensitivity Algorithm

The problem with global sensitivity is that empirically unrealistic but theoretically feasible data configurations drive sensitivity to infinity. In this section, we propose an algorithm that instead focuses on values of sensitivity that are empirically relevant. Our approach is analogous to an Empirical Bayes estimator, in that we use the data itself to construct a prior on possible levels of sensitivity rather than using an uninformed prior that permits all theoretically possible values (as in the calculation of global sensitivity).

Local Sensitivity. The starting point for our algorithm is measuring *local sensitivity* (Nissim, Raskhodnikova and Smith 2007) defined as the largest amount that adding or removing a single point can affect the statistic θ_g given the data that is actually observed in cell g. Figure 1 illustrates the computation of local sensitivity by considering a hypothetical Census tract with twenty observations of parent and child income percentiles. Based on these observations, the predicted value of children's income (y) at the 25th percentile of the parental income distribution ($x = 0.25$) is $\theta_g = 0.212$.

[7] As in much of the differential privacy literature, we take the privacy loss threshold ε as given. One way to choose ε is to weigh the tradeoffs between the social value of a more accurate statistic and the costs of potential privacy loss (Abowd and Schmutte 2019).

Figure 1: Calculation of local sensitivity



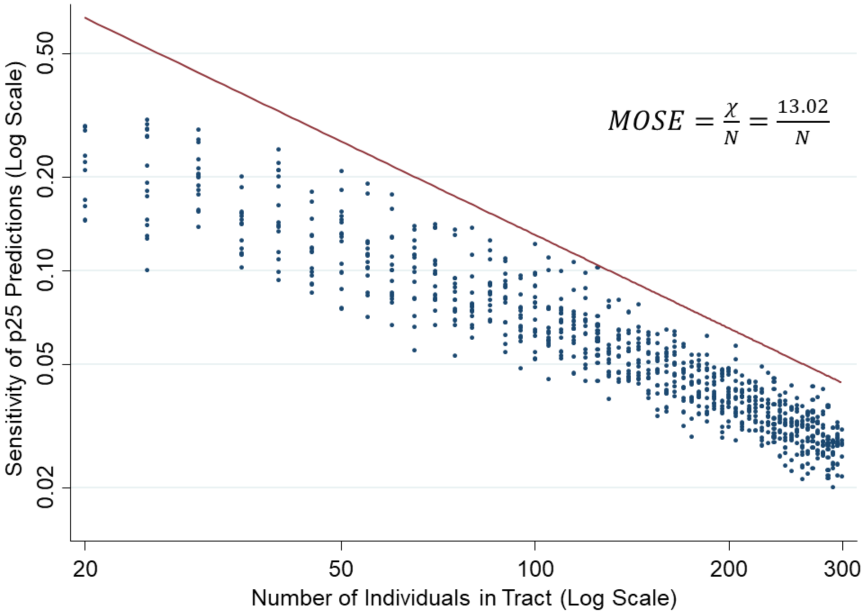
Notes: This figure shows how we calculate local sensitivity in a hypothetical cell (Census tract) with 20 individuals. The figure presents a scatter plot of children's income ranks in adulthood vs. their parents' income rank. The parameter of interest (θ_g) is the predicted value of child income rank at the 25^{th} percentile of the parent income distribution, as calculated from a univariate regression of child income rank on parent income rank in these data (shown by the solid best-fit line). In these data, the predicted value is $\theta_g = 0.212$. Local sensitivity is defined by the maximum absolute change in the predicted value by adding a point to or removing a point from the data. In this example, that occurs when adding the point (0,1), shown by the hollow dot in the upper left corner of the figure. With the addition of that point, the estimated regression line shifts to the dashed line, increasing θ_g by 0.137 - the local sensitivity of θ_g .

To compute local sensitivity, we recalculate this predicted value, adding new points one by one to see how much they affect the estimate of θ_g. In the example in Figure 1, adding a point at $(0, 1)$ – that is, an outlier where a child from a very low income family has a very high income in adulthood – has the biggest impact on the predicted value.[8] If that point is added, the original regression line flattens to become the dashed line, and the predicted value at the 25th percentile rises to $\theta_g = 0.349$. The local sensitivity in this example is therefore $LS_{\theta,g} = 0.349 - 0.212 = 0.137$.

Adding noise proportional to this level of sensitivity would, per equation (2.2), guarantee the desired upper bound on privacy loss from the public release of the statistic $\tilde{\theta}_g$. However, in order for users of this statistic to know the variance of the noise $Var(\omega)$ that was added – which is necessary for valid downstream inference – one must also release the value of local

[8] Formally, measuring local sensitivity requires consideration of three sets of cases – adding a point, removing a point, or changing a point – and finding the case that produces the largest change in θ_g. In the example in Figure 1 and in most practical applications with well-behaved distributions, adding a point in the corner of the dataspace typically produces the largest change in θ_g and thereby pins down sensitivity. Hence, the computation of local sensitivity can generally be simplified using a grid search in which one adds points to the corners of the dataspace.

Figure 2: Maximum Observed Sensitivity Envelope



Notes: This figure demonstrates our calculation of the Maximum Observed Sensitivity Envelope (MOSE) for a hypothetical dataset consisting of several cells (Census tracts) analogous to that in Figure 1. To construct this figure, we calculate the local sensitivity within each cell as described in Figure 1, and then plot the local sensitivity vs. the number of individuals in the cell. We use log scales on both axes. The MOSE, depicted by the solid line, is the function $MOSE(N_g) = \frac{\chi}{N_g}$, where $\chi = max_g\,[N_g \times LS_{\theta,g}] = 13.02$ in this example.

sensitivity $LS_{\theta,g}$, which discloses additional information and thereby can create a privacy risk.[9] Intuitively, $LS_{\theta,g}$ is itself a statistic that is estimated from the data D_g, just like $\tilde{\theta}_g$, and so it may reveal something about the underlying individual data. For instance, if sensitivity is very large, that may reveal that the data in cell g are tightly clustered around the regression line (as in the example in Figure 1). Hence, measuring sensitivity locally in each cell does not directly provide a feasible path to disclosing the statistics of interest while controlling privacy risk.

Maximum Observed Sensitivity. To reduce the information loss associated with disclosing local sensitivity in each cell, we measure sensitivity based on the largest local sensitivity across *all* cells. If all cells have the same number of observations N_g, we simply define sensitivity as $\Delta\theta_g = max_g[LS_{\theta,g}]$. In most empirical applications, however, cells differ in size. Since smaller cells typically have higher sensitivity, defining $\Delta\theta_g = max_g[LS_{\theta,g}]$ yields too conservative a bound on sensitivity. Figure 2 illustrates this point by presenting a scatter plot of local sensitivity $LS_{\theta,g}$, calculated as in Figure 1, vs. N_g across cells (using log scales). If we were to simply define $\Delta\theta_g = max_g[LS_{\theta,g}]$, sensitivity would be pinned down entirely by the smallest cells and would far exceed the actual local sensitivity of the estimates in larger cells.

[9] In principle, a privacy risk exists even when $Var(\omega)$ is not directly released, because one may be able to deduce information about $Var(\omega)$ from the observed $\tilde{\theta}_g$.

To achieve a tighter bound, we define an upper envelope to the set of points in Figure 2, which we term the *maximum observed sensitivity envelope*, as

$$MOSE(N_g) = \frac{\chi}{N_g},$$

where $\chi = max_g\left[N_g \times LS_{\theta,g}\right]$ is a scalar pinned down by the local sensitivity in one cell. The MOSE, illustrated by the solid line in Figure 2, is linear because both axes in the figure use log scales. Importantly, the MOSE weakly exceeds local sensitivity $LS_{\theta,g}$ in *all* cells by construction, as shown in the Figure 2, but falls as N_g rises.[10] Hence, by adding noise proportional to sensitivity $\Delta\theta_g = \frac{\chi}{N_g}$ in cell g, we can achieve the privacy guarantee in (2.2) when releasing $\left\{\tilde{\theta}_g\right\}$.

Our MOS method is still not differentially private because the scaling parameter χ is released publicly without noise, which discloses information that may not satisfy the guarantee in (2.2). However, the only potential uncontrolled privacy risk arises from the release of the single number χ; the privacy loss from releasing the cell-specific statistics $\left\{\tilde{\theta}_g\right\}$ themselves is guaranteed to be below ε. Moreover, we can take steps to reduce (though not formally bound) the privacy risk from releasing χ by computing it in a sufficiently large sample (e.g., across all tracts in a state). For example, the Census Bureau's Disclosure Review Board has adopted the interim policy of not requiring additional noise infusion for statistics based on populations at least as large as the smallest state, because the number of individuals in such groups is large enough that it is unlikely one could identify a single person using typical statistics.

Our method can be summarized as follows.

Maximum Observed Sensitivity (MOS) Disclosure Algorithm

To publish a statistic θ_g estimated using confidential data given a privacy risk threshold ε, release $\tilde{\theta}_g = \theta_g + \omega_g$, where the noise

$$\omega_g \sim L(0, \frac{\chi}{\varepsilon N_g}) \quad \text{or} \quad \omega_g \sim N\left(0, \sigma = \sqrt{2}\frac{\chi}{\varepsilon N_g}\right)$$

follows a LaPlace or Gaussian distribution, $\chi = max_g\left[N_g \times LS_{\theta,g}\right]$ is the MOS parameter, and $LS_{\theta,g}$ is local sensitivity, the maximum amount the statistic changes by adding or removing one observation in cell g.

See Appendix A for a step-by-step guide to implementing this method (accompanied by illustrative Stata code). Cell-specific counts can be released as $\tilde{N}_g = N_g + \nu_g$, where $\upsilon_g \sim L(0, \frac{1}{\varepsilon})$. The release of the counts in addition to the point estimates effectively doubles the privacy loss, making the algorithm 2ε-differentially private (aside from the release of the MOS parameter χ). Standard errors in each cell can be released using analogous methods;

[10] One could potentially achieve even tighter bounds using other functional forms rather than the $\frac{1}{N_g}$ scaling we use to define the upper envelope. In practice, the $\frac{1}{N_g}$ functional form yields a tight envelope (as illustrated in Figure 2) because the sensitivity of many common statistics (e.g., means, variances, and covariances) decays at rate $\frac{1}{N_g}$. However, sensitivities for other statistics decay at other rates (e.g., the decay rate for standard deviations is $\frac{1}{\sqrt{N_g}}$); users may wish to choose an appropriately scaled envelope to optimize the method for their setting.

see Appendix A for details. Releasing standard errors in addition to the point estimates and counts further increases the privacy risk threshold to 3ε.

Application: Opportunity Atlas. To further facilitate implementation, we discuss how we applied this method to release the Opportunity Atlas, which provides publicly available estimates of children's outcomes in adulthood by parental income, race, gender, and the tract in which they grew up (see Chetty et al. (2018) for details). This application illustrates how estimators can be optimized to minimize privacy loss (and hence the amount of noise that must be added to protect privacy) and maximize their utility in practical applications.

First, we worked only with bounded variables and used statistical transformations that limit the influence of outliers. For instance, rather than attempting to report estimates of mean income measured in dollars, we converted both children's and parents' incomes into percentile ranks.

Second, we winsorized both parent and child income ranks within each tract at the 5th and 95th percentiles by replacing all observations lying outside those quantiles in the distribution with the values of the cutoffs. In small tracts, we always replaced at least one high and low point with the next most extreme values. We found that winsorization substantially reduced the MOSE (by reducing the influence of outliers on each $\Delta\theta_g$), and thereby allowed us to release more accurate estimates at a given level of privacy loss.[11]

Third, we entirely omitted very small cells with fewer than 20 children to comply with other regulations governing the use of the data and because the estimates from these cells were too noisy to be useful. More generally, excluding very small cells can be useful to stabilize the estimates and reduce the risk of extremely high values of sensitivity $\Delta\theta_g$ that may in turn end up affecting the maximum observed sensitivity calculation. Note however that such censoring, if based on the true value of N_g, can introduce additional privacy risk by implicitly disclosing additional information. If regulations permit censoring instead on the released statistic $\tilde{N}_g$, there would be no additional privacy loss. More generally, any such pre-processing of the data based on non-public information potentially introduces additional privacy loss, and so such adjustments should ideally be made on the basis of publicly available information.

Fourth, we estimated the scaling parameter χ separately by state-gender-race groups, a level of aggregation that our data provider (the Census Bureau) determined had negligible privacy risks in our application.[12] We chose the privacy parameter ε by weighing the privacy losses against the potential social benefits of the statistics, as in Abowd and Schmutte (2019). Motivated by the real-world application of these data to help households with housing vouchers find higher-opportunity neighborhoods in which to live (Seattle Housing Authority 2017), we measured the social benefits of accuracy as the potential error rates faced by a housing authority wishing to identify the best and worst tracts in a given county for a given outcome. Specifically, we calculated the probability that tracts which appear in the top or bottom tail of the distribution of public (noise-infused) estimates in a given county are

[11] One must account for winsorization (and any other features of the estimation process) in the calculation of local sensitivity, in order to estimate the sensitivity of the composed function including both the winsorization and the estimation. That is, one must add an additional point to the pre-winsorized data and then winsorize before running the regression of interest.

[12] Estimating the envelope at the state-by-subgroup level reduces the scale on which the MOSE is based. If this approach is implemented at too granular a level, release of the MOSE could raise privacy concerns. Researchers should consult with their data providers or other context-specific experts to determine the appropriate level at which to estimate the MOSE.

actually in the true top or bottom tail in the confidential data for different values of ε. After plotting these error rates vs. ε and consulting with the Census Bureau, we set $\varepsilon = 8$ as a value that preserved sufficient accuracy for this application while injecting adequate noise to provide meaningful privacy protection.

Finally, we used a Normal distribution for the noise ω_g instead of a Laplace distribution because we expected the statistics we released to be used as an input in many downstream analyses (e.g., Morris, Gregory and Hartley 2018). Normally distributed noise is convenient for downstream statistical inference, such as the construction of confidence intervals or Bayesian shrinkage estimators.[13]

4. Comparison to Current Methods of Disclosure Limitation

In this section, we compare the properties of our noise infusion approach to existing methods of disclosure limitation. In particular, we contrast our method with count-based cell suppression – the leading technique used to limit disclosure risk – on three dimensions: privacy loss, statistical bias, and statistical precision.

Privacy Loss. Like most noise-infusion approaches, our method is likely to reduce the risk of privacy loss substantially relative to count-based cell suppression. This is because even if one suppresses cells with counts below some threshold, one can recover information about a single individual by releasing statistics (e.g., sample means) from adjacent datasets that differ by a single observation. Hence, statistics released after cell suppression still effectively have infinite (uncontrolled) privacy risk ε. In contrast, our maximum observed sensitivity approach reduces the dimensionality of the statistics that create uncontrolled privacy risks to one number (χ). Moreover, that number can typically be estimated in a sufficiently large sample that its release could reasonably be viewed as posing negligible privacy risk.[14]

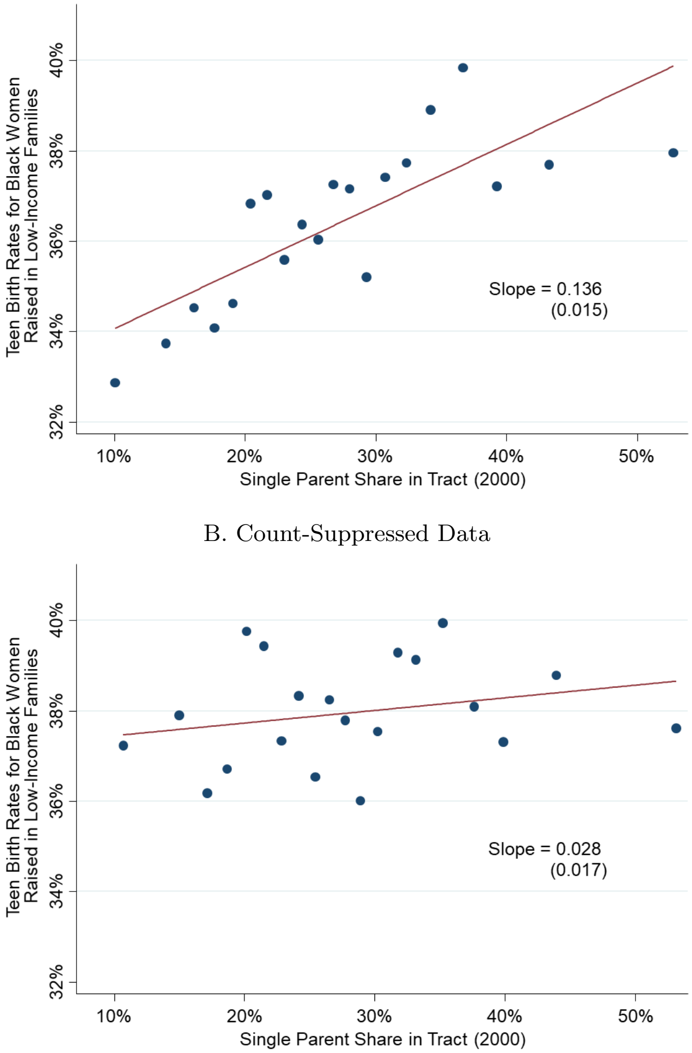
Statistical Bias. Our method also offers significant advantages in downstream statistical inference. Because we infuse random noise using parameters that are publicly known, one can obtain unbiased estimates of any parameter of interest using standard techniques. In contrast, count-based suppression can create bias in ways that cannot be easily identified or corrected ex-post.

To illustrate this point, we examine how results reported by Chetty et al. (2018) in their analysis of the Opportunity Atlas tract-level data would have changed had they used cell suppression. In particular, the authors show that black women who grow up in Census tracts with more single parents have significantly higher teenage birth rates, even among tracts with low poverty rates. Figure 3a shows a version of this finding by presenting a binned scatter plot of teenage birth rates for black women with parents at the 25th percentile vs. the share of single-parent families in the tracts in which they grew up, restricting the sample to low-poverty Census tracts (below 7%). There is a clear positive relationship between the two variables: an OLS regression implies that a 1 percentage point increase in single parent shares is associated with a 0.136 percentage point increase in teenage birth rates for

[13] Although the Census Bureau has permitted experimental approaches like the one used here and the one used in OnTheMap to recommend privacy loss levels, the Data Stewardship Executive Policy Committee, which oversees the Disclosure Review Board, controls privacy loss levels for production applications.

[14] We do not formalize this idea here, but the prior literature has identified some conditions under which low-dimensional summaries of large data sets can be proven to carry little risk (e.g. Bhaskar et al. 2011, Bassily et al. 2013).

Figure 3: Association between Teenage Birth Rates and Single Parent Shares



A. Noise-Infused Data

B. Count-Suppressed Data

Notes: This figure presents binned scatter plots of the relationship between teenage birth rates for black women and single parent shares across low-poverty Census tracts. Teenage birth rates are obtained from the publicly available Opportunity Atlas data and are defined as the fraction of black women who have a teenage birth among those born in the 1978-1983 birth cohorts and raised in families at the 25^{th} percentile of the household income distribution in a given Census tract. Data on the fraction of single headed households is obtained from the 2000 Decennial Census. We restrict the sample to Census tracts with a poverty rate of less than 7% based on the 2000 Decennial Census and winsorize tracts in the bottom or top 1% of the distribution of teenage birth rates to reduce the influence of outliers. To construct the binned scatter plots, we first bin tracts into 20 groups based on their single-parent share, weighting each tract by the number of black children under the age of 18 living in households with below median income. Each dot then plots the mean teenage birth rate (y-axis) vs. the mean single-parent share (x-axis) in each of the twenty bins. We estimate the best-fit line using an OLS regression on the tract-level data, again weighting by the number of black children under the age of 18 living in households with below median income. Panel A shows this relationship directly using the noise-infused, publicly available Opportunity Atlas data on teenage birth. Panel B replicates Panel A after omitting tracts where relatively few women have teenage births. Specifically, we impute the number of teenage births in a tract as the product of the predicted teenage birth rate for black women with parents at the 25^{th} percentile of the income distribution, the total count of black women in the sample, and the fraction of black women with parents with below median income. We then suppress cells if the implied count lies in the interval $[0.5, 4.5)$.

black women growing up in low-income families in low-poverty areas. The OLS regression coefficient provides an unbiased estimate of this statistic despite the addition of noise to the tract-level estimates because the noise simply enters the error term and is orthogonal to the independent variable by construction.[15]

We now examine how this result would have changed with cell suppression. When studying binary outcomes such as teenage birth, a common practice in the cell suppression approach is to omit data in tracts where very few (e.g., fewer than 5) teenage births occur (Washington State Department of Health 2018). (Current practice typically does not suppress a count of 0.) We mimic this rule in the Opportunity Atlas data by omitting tracts where black women raised in low-income families have between 1 and 4 teenage births (inclusive).[16]

Figure 3b replicates Figure 3a in the sample where tracts with 1-4 teenage births are suppressed. The strong positive correlation in Figure 3a disappears, with a slope that is now not statistically distinguishable from 0. The reason is that count-based suppression induces measurement error that is correlated with single parent shares through two sources. First, suppressing cells with few teenage births mechanically omits tracts with low teenage birth rates (Appendix Figure 2a), which are concentrated in areas with few single parents. Second, black women who grow up in areas with a smaller black population tend to have fewer teenage births (Appendix Figure 2b); tracts with a small black population in turn are more likely to be suppressed and also tend to be areas with few single parents. Identifying and correcting for these biases would be very difficult if one only had access to the post-suppression data. In short, one would likely have missed the association between teenage birth rates and single parent shares in low-poverty areas had Chetty et al. (2018) released data that followed standard cell-suppression techniques – illustrating that our noise infusion approach has significant advantages in terms of statistical bias not only in theory but in practice.

Statistical Precision. The key drawback of adding noise – which is typically the primary concern of most researchers – is that the estimates are less precise than those that would be obtained using cell suppression techniques (for the cells that are not suppressed). We again assess the practical importance of this concern in the context of the Opportunity Atlas. We find that the noise that was added to protect privacy does not meaningfully decrease precision because it is much smaller than the noise already present in the estimates due to sampling variation.

Table 1 demonstrates this point by decomposing the total (count-weighted) variance in the publicly-available tract-level statistics into the components reflecting sampling noise variance (based on the standard errors of the estimates), privacy noise variance (based on the known parameters of the noise distribution), and the remaining "signal" variance (which reflects the variance of the underlying "truth" under the assumptions used to estimate the standard errors). The first row shows this breakdown for teenage birth rates for black women raised in low-income families, the outcome analyzed in Figure 3. Just 0.8% of the

[15] Raw estimates of other statistics, such as the correlation between teenage birth rates and single parent shares, will be biased because of the addition of noise. But those biases can be easily corrected using standard techniques to correct for measurement error, e.g., by rescaling the correlation by the (known) amount of variance in teenage birth rates that is due to noise.

[16] For simplicity, we conduct this analysis in the publicly available Opportunity Atlas data rather than the confidential data. To do so, we impute the number of teenage births to black women in low-income families as the product of the predicted teenage birth rate for black women with parents at the 25th percentile of the income distribution, the total count of black women in the sample, and the fraction of black women with parents with below-median income. We then suppress cells if the implied count lies in the interval $[0.5, 4.5)$.

Table 1: Variance Decomposition for Tract-Level Estimates:
Selected Outcomes and Demographic Groups

	Signal Variance (1)	Sampling Noise Variance (2)	Privacy Noise Variance (3)
Panel A. Teenage Birth Rate, for Daughters of Parents at the 25th Percentile			
Black Females	71.00 %	28.18 %	0.82 %
White Females	70.58 %	28.31 %	1.11 %
Panel B. Share Incarcerated, for Sons of Parents at the 25th Percentile			
Black Males	56.39 %	40.21 %	2.32 %
White Males	33.64 %	38.85 %	27.51 %
Panel C. Household Income Rank, for Children of Parents at the 25th Percentile			
All Children	90.96 %	8.97 %	0.08 %
Black Children	75.19 %	23.59 %	1.22 %
White Children	78.90 %	20.82 %	0.27 %
Hispanic Children	69.62 %	29.08 %	1.30 %
Asian Children	69.99 %	28.94 %	1.06 %
American Indian & Alaska Native Children	81.90 %	17.28 %	0.82 %
White Males	64.71 %	34.60 %	0.69 %
White Females	70.36 %	28.93 %	0.71 %
Black Males	66.58 %	31.29 %	2.13 %
Black Females	73.16 %	25.17 %	1.67 %

Notes: This table reports a variance decomposition of Census-tract-level statistics from the Opportunity Atlas (Chetty et al., 2018), which are predicted outcomes for children based on the tract in which they grow up. We focus on predicted outcomes of children with parents at the 25th percentile of the parental income distribution. See Chetty et al. (2018) for definitions of these variables. We restrict the sample to tracts in which the outcome variable of interest is calculated using more than fifty observations. We then decompose the total tract-level variance into the fraction that comes from signal variance (reflecting the variance of the latent parameters of interest under our modelling assumptions), noise variance due to sampling variation, and noise variance from the noise we infused to protect privacy. Each row in the table presents this decomposition for a particular outcome variable and demographic group. The three percentages add to 100% across each row (as the three variance components are independent). To calculate the decomposition, we first calculate the total variance in the outcome across tracts, weighting by the number of children in the relevant demographic group with parent income below the national median. We then estimate the noise variance due to sampling error as the mean of the squared standard errors (again using the same weights). We calculate the privacy noise variance in each tract based on the sensitivity and privacy risk parameters used in our application of the MOS algorithm and the tract-specific count, and again take a weighted mean across tracts. Finally, we compute the signal variance as the total variance minus the sum of the two noise variances. Panel A presents this variance decomposition for the teenage birth rate for the demographic subgroup specified in the first column, following the U.S. Census Bureau's definitions of race and Hispanic identity. The first row (of Panel A) corresponds to data plotted in Figure 3. Panels B and C replicate Panel A using the share incarcerated and child household income rank, respectively.

total variance across tracts and only 2.8% of the total noise variance comes from the added privacy noise. Phrased differently, the reliability of the estimates (the ratio of signal variance to total variance) falls very slightly, from 71.8% to 71.0%, due to the addition of noise to protect privacy. The other rows of Table 1 provide a similar breakdown for other outcomes and subgroups. For most outcomes, the privacy noise variance is even smaller than for teenage birth rates. For a few variables, such as the incarceration rate for white men, the privacy noise variance share is significantly higher, but it is still always smaller than the noise due to sampling error.

Of course, noise infusion would have larger effects on reliability in any given application with a lower value of ε, and even with the same value of ε, it could have larger effects in other applications. Nevertheless, the Opportunity Atlas demonstrates that one can achieve substantial gains in terms of bias and privacy protection while incurring only small losses in statistical precision using our method, especially by optimizing the estimators one uses as discussed at the end of Section 3.

5. Conclusion

Building on ideas from the differential privacy literature, this paper has developed a practical noise-infusion method for reducing the privacy loss from disclosing statistics based on confidential data. The method outperforms existing, widely-used methods of disclosure limitation *both* in terms of privacy loss and statistical bias. Importantly, it can be easily applied to virtually any statistic of interest to social scientists. For example, consider difference-in-differences or regression discontinuity estimators. Even if there is only one quasi-experiment (e.g., a single policy change in a given area), one can construct "placebo" estimates by pretending that a similar change occurred in other cells of the data and computing the maximum observed sensitivity of the estimator across all cells.

In future work, it would be useful to develop metrics for privacy loss for algorithms in which a single statistic (e.g., sensitivity) is disclosed based on a large sample (e.g., at the state or national level). Here, we argued on an intuitive basis that the release of such statistics has small privacy costs, but formalizing this idea – perhaps by placing restrictions on distributions or the set of estimators – could provide a way to offer formal privacy guarantees. More broadly, developing differential privacy techniques that can be applied to many estimators – as we have done here – without requiring users to develop new algorithms for each application may help increase the use of such methods in social science.

References

Abowd, John M., and Ian M. Schmutte. 2015. “Economic Analysis and Statistical Disclosure Limitation.” *Brookings Papers on Economic Activity*, 221–267. http://www.jstor.org/stable/43684103.

Abowd, John M., and Ian M. Schmutte. 2019. “An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices.” *American Economic Review*, 109(1): 171–202. https://doi.org/10.1257/aer.20170627.

Alexander, J. Trent, Michael Davern, and Betsey Stevenson. 2010. “Inaccurate age and sex data in the Census PUMS files: Evidence and Implications.” National Bureau of Economic Research Working Paper 15703, https://doi.org/10.3386/w15703.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters. 2013. “Explaining Charter School Effectiveness.” *American Economic Journal: Applied Economics*, 5(4): 1–27. https://doi.org/10.1257/app.5.4.1.

Bassily, Raef, Adam Groce, Jonathan Katz, and Adam Smith. 2013. “Coupled-Worlds Privacy: Exploiting Adversarial Uncertainty in Statistical Data Privacy.” https://doi.org/10.1109/FOCS.2013.54.

Bhaskar, Raghav, Abhishek Bhowmick, Vipul Goyal, Srivatsan Laxman, and Abhradeep Thakurta. 2011. “Noiseless Database Privacy.” https://doi.org/10.1007/978-3-642-25385-0_12.

Blum, Avrim, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. 2005. “Practical Privacy: The Sulq Framework.” *24th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems, Baltimore (PODS 2005)*. https://doi.org/10.1145/1065167.1065184.

Chaudhuri, Kamalika, Claire Monteleoni, and Anand D. Sarwate. 2011. “Differentially Private Empirical Risk Minimization.” *Journal of Machine Learning Research*, 12: 1069–1109. http://www.jmlr.org/papers/v12/chaudhuri11a.html.

Chetty, Raj, and John N. Friedman. 2019*a*. “Code for ”A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples.” Zenodo Computer Code jpc.716.1, https://doi.org/10.5281/zenodo.3476957.

Chetty, Raj, and John N. Friedman. 2019*b*. “A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples.” *AEA Papers and Proceedings*, 109: 414–20. https://doi.org/10.1257/pandp.20191109.

Chetty, Raj, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. 2018. “The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.” National Bureau of Economic Research Working Paper 25147, https://doi.org/10.3386/w25147.

Cooper, Zack, Stuart V Craig, Martin Gaynor, and John Van Reenen. 2018. “The Price Ain’t Right? Hospital Prices and Health Spending on the Privately Insured.” *The Quarterly Journal of Economics*, 134(1): 51–107. https://doi.org/10.1093/qje/qjy020.

Dwork, Cynthia. 2006. “Differential privacy.” *Automata, Languages, and Programming. ICALP 2006. Lecture Notes in Computer Science*, , ed. M. Bugliesi, B. Preneel, V. Sassone and I. Wegene Vol. 4052, Chapter 1, 1–12. Springer. https://doi.org/10.1007/11787006_1.

Dwork, Cynthia, and Jing Lei. 2009. “Differential Privacy and Robust Statistics.” . Proceedings of the 41th Annual ACM Symposium on Theory of Computing (STOC) ed. Association for Computing Machinery, Inc. https://www.microsoft.com/en-us/research/publication/differential-privacy-and-robust-statistics/.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006*a*. "Calibrating Noise to Sensitivity in Private Data Analysis." 265–284. Berlin, Heidelberg:Springer Berlin Heidelberg. https://doi.org/10.1007/11681878_14.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2017. "Calibrating Noise to Sensitivity in Private Data Analysis." *Journal of Privacy and Confidentiality*, 7(3): 17–51. https://doi.org/10.29012/jpc.v7i3.405.

Dwork, Cynthia, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006*b*. "Our Data, Ourselves: Privacy Via Distributed Noise Generation." 486–503. Springer. https://doi.org/10.1007/11761679_29.

Garfinkel, Simson, John M Abowd, and Christian Martindale. 2018. "Understanding database reconstruction attacks on public data." *Queue*, 16. https://doi.org/10.1145/3291276.3295691.

Hull, Peter. 2018. "Estimating Hospital Quality with Quasi-experimental Data." https://doi.org/10.2139/ssrn.3118358.

Kasiviswanathan, Shiva Prasad, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. 2011. "What Can We Learn Privately?" *SIAM Journal of Computing*, 40(3). https://doi.org/10.1137/090756090.

Kifer, Daniel, Adam Smith, and Abhradeep Thakurta. 2012. "Private Convex Empirical Risk Minimization and High-dimensional Regression." Vol. 23 of *Proceedings of Machine Learning Research*, 25.1–25.40. Edinburgh, Scotland:PMLR. http://proceedings.mlr.press/v23/kifer12.html.

McSherry, Frank, and Kunal Talwar. 2007. "Mechanism Design via Differential Privacy." *48th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. https://doi.org/10.1109/FOCS.2007.66.

Morris, Davis A., Jesse Gregory, and Daniel A. Hartley. 2018. "The Long-Run Effects of Low-Income Housing on Neighborhood Composition." https://www.ssc.wisc.edu/~jmgregory/DGTcurrent.pdf.

Nissim, Kobbi, Sofya Raskhodnikova, and Adam Davison Smith. 2007. "Smooth sensitivity and sampling in private data analysis." 75–84. https://doi.org/10.1145/1250790.1250803.

Seattle Housing Authority. 2017. "Creating Moves to Opportunity: Seattle & King County." https://www.seattlehousing.org/sites/default/files/CMTO_Fact_Sheet.pdf.

Smith, Adam. 2011. "Privacy-preserving Statistical Estimation with Optimal Convergence Rates." *STOC '11*, 813–822. New York, NY, USA:ACM. https://doi.org/10.1145/1993636.1993743.

U.S. Census Bureau. 2018. "Statistical Safeguards." *Data Protection and Privacy Program, U.S. Census Bureau*. https://www.census.gov/about/policies/privacy/statistical_safeguards.html.

Washington State Department of Health. 2018. "Department of Health Agency Standards for Reporting Data with Small Numbers." 1–24. https://www.doh.wa.gov/portals/1/documents/1500/smallnumbers.pdf.

Wasserman, Larry, and Shuheng Zhou. 2010. "A statistical framework for differential privacy." *Journal of the American Statistical Association*, 105(489): 375–389. https://doi.org/10.1198/jasa.2009.tm08651.

Appendix A. Implementation Guide

This guide provides step-by-step instructions for implementing our algorithm to release statistics constructed from a confidential database. It also provides some suggestions to simplify computation and minimize the amount of noise that has to be infused to achieve a given level of privacy protection. Illustrative Stata code that implements the five steps below in the context of a regression estimate is available here.

Step 0. **Estimate the statistic** θ you are interested in releasing – e.g., a coefficient from a regression or a parameter from another statistical model – in the confidential data.

a. All variables must be bounded for the algorithm below to work (i.e., yield finite estimates of sensitivity). If you are working with unbounded variables, bottom- and top-code them before proceeding (but do so in a way that does not depend on the particular realized values in any given sample).

b. Consider alternative estimators that reduce the influence of outliers and will thereby reduce the amount of noise you need to add to meet a given privacy threshold. For example, in the context of regression, winsorizing variables at the 5th and 95th percentiles can reduce the influence of outliers without significantly affecting estimates of the parameters of interest. Estimators such as median regression may also be less sensitive to outliers (Dwork and Lei 2009).

c. It may be useful to implement Steps 1-4 below with alternative estimators to calculate the amount of noise that must be added to the estimates using a given estimator. Then choose the estimator (e.g., the winsorization threshold) that minimizes noise while yielding suitable estimates for your application.

Implement the following five steps to add noise to the estimates and release them publicly:

Step 1. **Calculate local sensitivity** $LS_{\theta,g}$ for the scalar statistic θ_g in each cell g of your data, defined as the largest absolute change in θ from adding or removing a single observation d:

$$LS_{\theta,g} = max_{d \in D_g} |\theta_{\pm d} - \theta|,$$

where $\theta_{\pm d}$ is the estimate obtained when adding or removing observation d from the dataset D_g in cell g.

a. Local sensitivity can be calculated using grid search and other standard optimization techniques; for well-behaved relationships, adding points to the corners of the dataspace D_g will typically be sufficient to calculate local sensitivity. For example, in a univariate regression where both the dependent and independent variables are bounded between 0 and 1, adding the points $(0, 1)$, $(1, 0)$, $(0, 0)$, and $(1, 1)$ is typically adequate to calculate local sensitivity, although users should use finer grid searches or other numerical optimization methods to confirm the accuracy of any particular approach.

b. In high-dimensional dataspaces (e.g., multivariable regression with many regressors), removing points is more computationally tractable than adding new points, since the number of potential points that can be added grows exponentially with the number of variables. In such cases, it may be convenient to consider only removals, perhaps after establishing that estimates of local sensitivity are similar whether or not one allows for the addition of points in a subset of cells. Examining only removals is most likely to be adequate for privacy when one uses estimators that are not sensitive to outliers (e.g., through winsorization).

c. If you are interested in reporting a statistic from a single cell (e.g., a treatment effect estimate you have constructed for a specific subgroup or geographic unit), find other similar units in your dataset and treat them as distinct cells. Then replicate your estimator in each of those cells, assigning "placebo" treatment variables that have the same structure as the actual treatment if necessary, to obtain estimates of local sensitivity across several analogous cells.

Step 2. **Compute the maximum observed sensitivity envelope scaling parameter** χ:

$$\chi = max_g \left[N_g \times LS_{\theta,g}\right],$$

where N_g is the number of observations (e.g., individuals) used to estimate the statistic θ in cell g.

a. Compute χ by taking the maximum across cells at a sufficiently high level of aggregation that your data provider considers the privacy risks from releasing the exact value of χ to be negligible (e.g., the state or national level).

Step 3. **Determine the privacy parameter** ε for your release using one of the following methods:

a. Follow established guidelines on ε from your data provider.

b. Choose ε by plotting the social gain from greater accuracy vs. ε and choosing a value of ε that you and the data provider agree optimizes this tradeoff (Abowd and Schmutte 2019). If there is no clear loss function or decision problem, two practical definitions of the social gain from accuracy are the mean squared error (MSE) or the classification error in the noise-infused statistic relative to the truth. The MSE can be computed by calculating the error $(\widetilde{\theta}_g - \theta_g)^2$ based on the estimate constructed in step 4 below and averaging over several draws of the noise distribution. Classification error is the probability that the true value of θ_g falls below a certain threshold conditional on the noise-infused value $\widetilde{\theta}_g$ falling above that threshold. For example, one might calculate the probability that θ_g falls outside the top 10% of the distribution of $\{\theta_g\}$ conditional on observing $\widetilde{\theta}_g$ in the top 10% of the distribution of $\left\{\widetilde{\theta}_g\right\}$.

Step 4. **Add random noise** proportional to maximum observed sensitivity χ and the privacy parameter ε to each statistic:

$$\widetilde{\theta}_g = \theta_g + \sqrt{2}\frac{\chi}{\varepsilon N_g}\omega,$$

where ω is a random variable with mean 0 and standard deviation 1.

a. The distribution of ω can be chosen depending upon the application and the requirements of the data provider. If the statistics will be used for downstream analysis, it is convenient to add $N(0,1)$ noise so that the total noise variance remains normally distributed. If not, using a LaPlace distribution $L(0, \frac{1}{\sqrt{2}})$ conforms more precisely to the desired privacy loss limit at all points in the distribution, with the differences largest in the tails of the distribution.

b. To report standard errors, first estimate the standard error (SE) of the estimate including the noise added to protect privacy:

$$SE(\widetilde{\theta}_g) \quad = \quad \sqrt{SE\left(\theta_g\right)^2 + 2\left(\frac{\chi}{\varepsilon N_g}\right)^2}.$$

Then apply the same procedure as above to construct public, privacy-protected estimates $\widetilde{SE}(\widetilde{\theta}_g)$ of the standard errors themselves, treating $SE(\widetilde{\theta}_g)$ like the statistics θ_g above.

c. Construct noise-infused estimates of the count of observations in each cell as follows, using the same definition of ω as in Step 4a:

$$\widetilde{N}_g = N_g + \sqrt{2}\frac{\omega}{\varepsilon}.$$

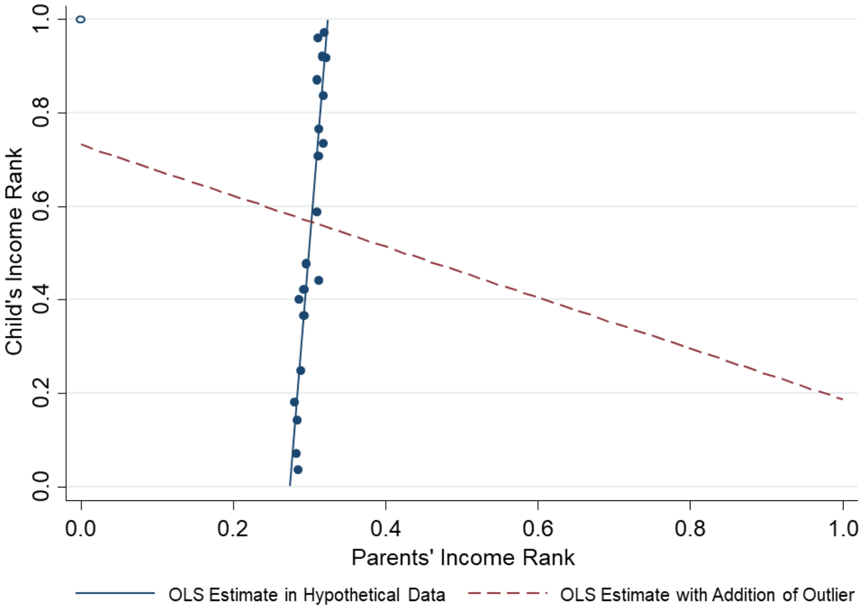
d. To quantify the amount of noise added, compute the standard deviation of the noise distribution in your cell of interest, $\sqrt{2}\frac{\chi}{\varepsilon N_g}$, or the share of the variance in your cell-specific estimates that is due to noise, $2E[(\frac{\chi}{\varepsilon N_g})^2]/Var(\widetilde{\theta}_g)$, where the expectation is taken over the cells in the dataset.

Step 5. **Release the noise-infused statistics** $\left\{\widetilde{\theta}_g\right\}$ and $\{\widetilde{SE}(\widetilde{\theta}_g)\}$, counts $\left\{\widetilde{N}_g\right\}$, and parameters that control the amount of noise added (ε and χ) for the groups of interest publicly.

Questions? Email info@opportunityinsights.org

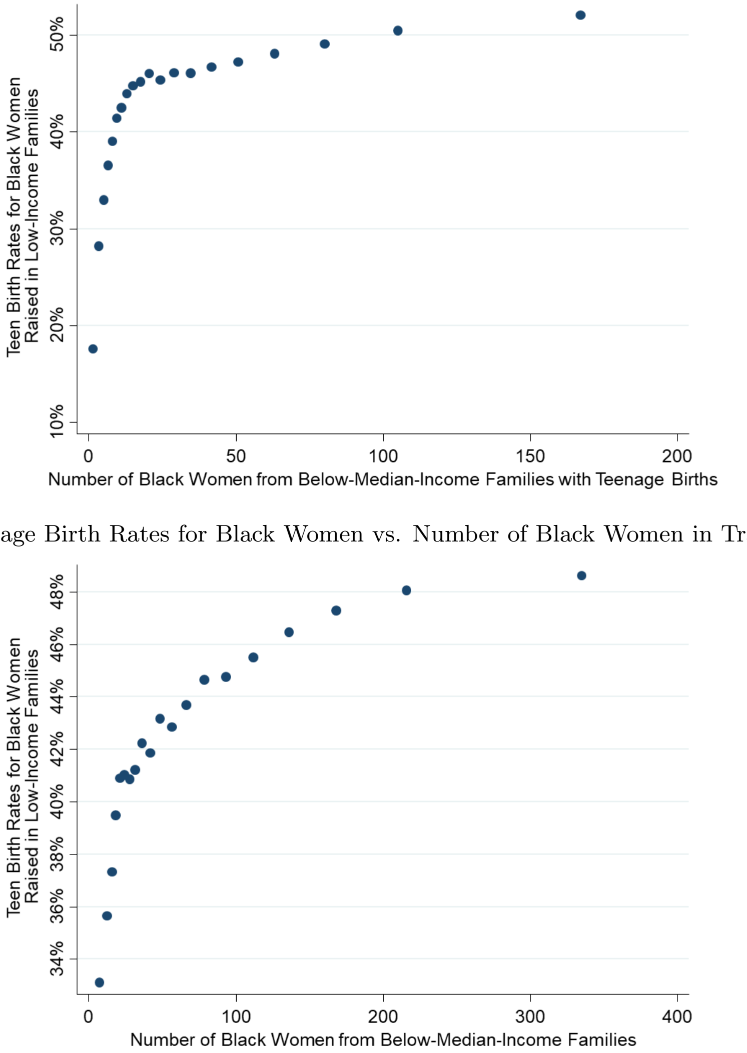
2. Additional Figures

Appendix Figure 1: Calculation of Global Sensitivity



Notes: This figure shows how we calculate global sensitivity in a hypothetical cell (Census tract) with 20 individuals. As in Figure 1, the figure presents a scatter plot of children's income ranks in adulthood vs. their parents' income rank. However, rather than using empirically observed values, we choose a set of data points that make the sensitivity of the regression estimates to the addition of a single point very large, which occurs as the variance of the dots on the x-axis becomes small. In this case, adding a single point at (0,1) changes the estimated slope of the regression line (and the corresponding predicted values) substantially, showing why the global sensitivity of OLS regression estimates can be arbitrarily large.

Appendix Figure 2: Teenage Birth Rates vs. Counts, by Census Tract



A. Teenage Birth Rates for Black Women vs. Number of Black Women with Teenage Births in Tract

B. Teenage Birth Rates for Black Women vs. Number of Black Women in Tract

Notes: This figure presents binned scatter plots of tract-level teenage birth rates for black women raised in families at the 25^{th} percentile vs. two measures of tract-specific counts. Panel A plots teenage birth rates for black women vs. the number of black women from low-income families who have teenage births in the tract, computed as the product of the predicted teenage birth rate for black women with parents at the 25^{th} percentile of the income distribution, the total count of black women in the sample, and the fraction of black women with parents with below-median income. (i.e., the numerator of the teenage birth rate). Panel B plots teenage birth rates for black women vs. the total number of black women from below-median-income families in the tract (i.e., the denominator of the teenage birth rate). The binned scatter plots are constructed by binning tracts into twenty equal-sized bins based on the x-axis variable and plotting the means of the x and y variables within each bin.