A Practical Method to Reduce Privacy Loss when Disclosing Statistics Based on Small Cells*

Raj Chetty, Harvard University and NBER
John N. Friedman, Brown University and NBER

January 2019

Abstract

We develop a simple method to reduce privacy loss when disclosing statistics such as OLS regression estimates based on samples with a small number 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 statistics of interest in proportion to the statistic’s maximum observed sensitivity, defined as the maximum change in the statistic from removing a single observation across all the cells in the data. Although not provably private, our method 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.

*We thank John Abowd, Simson Garfinkel, James Honaker, Adam Smith, Salil Vadhan, and the Harvard University Privacy Tools Project for valuable feedback, as well as Jamie Gracie, Martin Koenen, Daniel Reuter, Jesse Silbert, Joseph Winkelmann, 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.
I Introduction

Social scientists increasingly use confidential data held by government agencies or private firms to publish statistics based on cells with a small number of individuals, from descriptive statistics on local income distributions and health expenditures (e.g., Cooper et al. 2015, Chetty et al. 2018) to estimates of the causal effects of specific schools and hospitals (e.g., Angrist et al. 2013, Hull 2018). Such statistics allow researchers and policymakers to develop tailored solutions to 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.

The most widely applied approaches to limiting disclosure risk in social science are cell suppression (omitting data for small cells) or data swapping (randomly switching individual values across cells). These techniques have an intuitive appeal 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 2015). 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.

The recent literature on differential privacy, initiated in seminal work by Dwork 2006, 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., McSherry and Talwar 2007, Dwork 2008, Kasiviswanathan et al. 2011). However, methods to protect privacy when disclosing more complex estimators – such as regression or quasi-experimental estimators – rely on either asymptotic results in large samples (e.g., Smith 2011 and Kifer et al. 2012) or the use of robust statistics such as median regression (e.g., Dwork and Lei 2009), limiting their application in social science.

In this paper, we develop an easily implementable method of reducing the privacy loss from disclosing arbitrarily complex statistics in small cells. Our approach aims to combine 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 approach, we propose a method that outperforms the most widely used methods of disclosure limitation such as cell suppression both in terms of privacy loss and statistical bias. We then illustrate how this method can be used to estimate social mobility in small areas – a topic of central interest in the recent literature on neighborhood effects in economics and sociology.

For concreteness, we focus on the problem of releasing estimates from univariate ordinary least squares (OLS) regressions estimated in small cells (e.g., small geographic units). In the neighborhood effects example, this could represent the predicted values from a regression of children’s income ranks in adulthood on their parents’ income ranks in a given Census tract. Following a common approach in 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 removed for any possible configuration 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 removal 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 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 removed from the actual data in a given cell. While this is a finite value, the problem with this approach is that releasing the local sensitivity of statistics in a given cell may itself release confidential information even if the privacy loss from the regression estimates themselves is small. Intuitively, the local sensitivity measure is itself a statistic computed in a small cell and thus reveals some information about the underlying data.
Our approach to computing sensitivity is a hybrid that lies between local and global sensitivity. We calculate local sensitivity in each of the cells of interest (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 in the sample (e.g., across all tracts in a given state), adjusting for differences in the number of observations across cells.\footnote{Our maximum observed sensitivity approach is analogous to using an Empirical Bayes estimator, where the empirical distribution of the sensitivities across all cells provides a guide as to the potential sensitivity in the cell of interest. In contrast, the global sensitivity approach is analogous to using a minimax estimator with an unknown prior on the distribution of sensitivities.} Critically, we compute the MOS at a level of aggregation of the data that is high enough that disclosure risks are considered minimal ex-ante. For example, the Census Bureau’s Disclosure Review Board currently does not consider most statistics aggregated to state (or higher) level to pose disclosure risks because the number of individuals living in a state is large enough that it is unlikely one could identify a single person based purely on state-level statistics.\footnote{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—even though no formal proof—that the privacy risks in state-level data are generally small enough to be ignored, at least at present.}

We illustrate how the method can be implemented by discussing how we used it to produce the [Opportunity Atlas](https://www2.census.gov/programs-surveys/opportunity/), a tool that provides public estimates of children’s long-term outcomes by the tract in which they grew up. We reduce the sensitivity of the statistics we release through procedures such as bounding variables and winsorization. We then choose the privacy threshold $\varepsilon$ by following Abowd and Schmutte [2019](https://www2.census.gov/programs-surveys/opportunity/) by weighing the privacy losses of a higher $\varepsilon$ against the social benefits, which we define as providing more accurate information to a family seeking to move to a higher-opportunity neighborhood. Ultimately, the noise infused in the estimates to protect privacy was smaller than the sampling error inherent in the estimates themselves and hence do not reduce statistical accuracy significantly in most cases. The Opportunity Atlas estimates released using this approach have been viewed by half-a-million users, are currently being used to inform moving-to-opportunity housing policies by housing authorities, and have been used as inputs by other researchers in downstream analyses (e.g., Morris et al. [2018](https://www2.census.gov/programs-surveys/opportunity/)). The Opportunity Atlas thus provides a large-scale, real-world demonstration that our approach to reducing privacy loss yields estimates that remain quite useful for practical social science and policy applications.

Our approach outperforms the most popular disclosure limitation protocols that social scientists currently use (suppression of cells with small counts) both in terms of privacy loss and statistical
In terms of privacy loss, it is straightforward to show that cell suppression is a technique that has infinite sensitivity. As discussed above, even if one suppresses cells with counts below some threshold, one can recover information about a single individual by releasing statistics from two regressions on a constant (i.e., sample means) on adjacent datasets that differ by a single observation. 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 $\varepsilon$. Although our approach is not provably private in the sense of guaranteeing that there is no statistic that would release information at a rate higher than $\varepsilon$ – in particular because our computation of the MOS does not satisfy such a guarantee – this example shows that our approach is at least weakly preferable to conventional cell suppression methods in terms of privacy loss. Moreover, if one takes the view that the MOS does not itself disclose sensitive information, our approach allows users to control the rate of privacy loss in a well-defined, quantifiable manner, unlike traditional cell suppression techniques.

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 teenage birth rates are significantly higher in Census tracts with more single parents (in the previous generation). If one were to instead conduct their analysis using data where tracts with very few (less than 3) teenage births occur – a common approach to limit disclosure risk of rare outcomes – the correlation is actually slightly negative. This is because the suppression rule leads to non-random missing data (in particular, excluding cells with particularly low teenage birth rates) and because the tracts with the lowest teenage birth rates tend to be those with the smallest number of black women, i.e. precisely the tracts that are omitted under count-based suppression. 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 paper is organized as follows. The next section sets up the problem. Section III describes our noise infusion method, both in general terms and in the application to the Opportunity Atlas. Section IV shows how our noise infusion method outperforms traditional count-based cell suppression methods. Section V concludes. A step-by-step guide to implementing the method (along with illustrative Stata code) is provided in the Appendix.

---

3We present a comparison to widely applied count-based suppression mechanisms for concreteness here, but similar points would apply to data swapping as well (Alexander et al. 2010, Abowd and Schmutte 2015).
II The Problem

Our goal is to disclose a statistic $\theta$ that is a scalar (XXRaj just took “scalar” here) 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 parent’s 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 $\theta_g = \alpha_g + 0.25 \times \beta_g$ as measured of upward mobility 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 [2006]). To see this, let $\tilde{\theta}_g = \theta_g + \omega_g$ denote the noise-infused statistic, where $\omega_g$ is an independently and identically distributed draw from distribution $F(\omega)$, so that the conditional distribution of $\tilde{\theta}_g$ given $\theta_g$ is $F(\tilde{\theta}_g - \theta_g)$. Let $D_g = \{x_{ig}, y_{ig}\}$ denote the empirically observed data in cell $g$ and $\mathbb{D}_g$ denote the set of potential datasets. The privacy loss from disclosing $\tilde{\theta}_g$ can be measured using the log likelihood ratio

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

(1)

where $D^1_g, D^2_g \in \mathbb{D}_g$ 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 $D^1_g$, relative to $D^2_g$; from a Bayesian perspective, the larger is this ratio, the more one could update one’s priors between $D^1_g$ and $D^2_g$ given the release of statistic $\tilde{\theta}_g$.

When no noise is infused (i.e., $\text{Var}(\omega_g) = 0$), this likelihood ratio will almost surely be infinite, as one could perfectly distinguish between any two datasets $D^1_g$ and $D^2_g$ that do not happen to produce exactly the same value of $\theta_g$. 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 (1), effectively providing a “worst case” guarantee on the degree of privacy loss. In the terminology introduced by Dwork et al. [2006], a privacy algorithm is “$\varepsilon$-differentially private” if
\[
\log \frac{f(\tilde{\theta}_g - \theta_g(D^1_g))}{f(\tilde{\theta}_g - \theta_g(D^2_g))} < \varepsilon \quad \forall D^1_g, D^2_g \in \mathbb{D}_g, \forall \tilde{\theta}_g \in \mathbb{R}.
\] (2)
The parameter $\varepsilon$ can be interpreted as the maximum risk one is willing to tolerate when releasing the statistic of interest. Dwork et al. [2006] establish that one can achieve this bound by adding Laplacian noise (XXRaj wrote out Laplace density) with density $l(x; \mu, b) = \frac{1}{2b} \exp \left[ -\frac{|x-\mu|}{b} \right]$ to the estimates:
\[
\omega_g \sim L \left( 0, \frac{\Delta \theta_g}{\varepsilon} \right),
\] (3)
where $\Delta \theta_g$ is the “sensitivity” of the statistic $\theta_g$, defined as
\[
\Delta \theta_g = \max_{D^1_g, D^2_g \in \mathbb{D}_g} \left| \theta_g(D^1_g) - \theta_g(D^2_g) \right|.
\]
Sensitivity measures the maximum amount that the statistic can change between any two adjacent datasets. When sensitivity is very high – that is when changing a single observation changes $\theta_g$ a lot – one must add a lot of 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.

The amount of noise that must be added to achieve the bound in (2) depends upon two parameters: the privacy loss threshold $\varepsilon$ and sensitivity $\Delta \theta_g$. The privacy loss threshold is a parameter that is typically specified exogenously based on a judgment about the tradeoffs between the social utility of a more accurate statistic and the costs of potential privacy loss, as discussed in Abowd and Schmutte [2019]. The remaining question is how one calculates sensitivity $\Delta \theta_g$; if $\Delta \theta_g$ were publicly known, then one could obtain differentially private statistics simply by adding noise as in (3).

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 does not reveal any information about the actual dataset \( D_g \), and can therefore be released publicly along with the statistic \( \hat{\theta}_g \) without any further privacy loss, yielding a fully differentially private disclosure mechanism. Researchers have applied this global-sensitivity approach to provide to release statistics such as counts and means (Dwork et al. 2006); 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 regression estimates and many other statistics of interest to social scientists. To see this in our setting, consider the limiting case where \( \text{Var}(x_{ig}) \) 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 \( \hat{\theta}_g \)) 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. Therefore, global sensitivity is infinite, implying that adding any finite amount of noise will not meet the differential privacy guarantee in (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. How can we proceed? We propose an alternative approach to computing sensitivity in the next section.

### III The 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 conceptually similar

---

4The literature has developed other approaches that allow the publication of regression estimates that are differentially private with finite levels of infused noise. For instance, one can release a noise-infused estimate of the covariance of the dependent and independent variable, and then separately release the a noise-infused estimate of the variance of the independent variable. However, the resulting estimate of the regression slope, calculated as the ratio of these two statistics, is biased to the noise-infused estimate of the variance in the denominator. We restrict ourselves in this paper to publication of statistics that are unbiased. Similarly, a number of results exist for asymptotically limiting cases in large dataset, or when restricted to robust statistics such as median regression, but not in the finite sample OLS regression case we consider here.
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, defined as the largest amount that adding or removing a single point can affect the statistic $\theta_g$ 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 $\theta_g = 0.212$, denoted by the square point.

To compute local sensitivity, we recalculate this predicted value, adding new points one by one to see how much they affect the estimate of $\theta_g$. 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. 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), 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 sensitivity $LS_{\theta,g}$, which discloses additional information and thereby can create a privacy risk. 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

---

5Formally, 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 $\theta_g$. 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 $\theta_g$ 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.
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.

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 $\chi = \max_g [N_g \times LS_{\theta,g}]$ 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. 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) when releasing $\{\tilde{\theta}_g\}$.

Our MOS method is still not differentially private because the scaling parameter $\chi$ is released publicly without noise, which discloses information that may not satisfy the guarantee in (2). However, the only potential uncontrolled privacy risk arises from the release of the single number $\chi$; the privacy loss from releasing the cell-specific statistics $\{\tilde{\theta}_g\}$ themselves is guaranteed to be below $\varepsilon$. Importantly, we can take steps to reduce (though not formally bound) the privacy risk from releasing $\chi$ by computing it a sufficiently large sample (e.g., across all tracts in a state). For example, the Census Bureau’s Disclosure Review Board currently does not consider most statistics aggregated to the state or higher level to pose disclosure risks because the number of individuals living in a state is large enough that it is unlikely one could identify a single person using typical state-level statistics. To summarize:

---

6One 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 decays at a rate of approximately $\frac{1}{N_g}$. For example, as discussed in Section II, the sensitivity of the mean of a bounded variable is proportional to $\frac{1}{N_g}$.
### Maximum Observed Sensitivity (MOS) Method of Releasing Statistics

The MOS method of reducing privacy loss is to release \( \{\hat{\theta}_g, \tilde{N}_g, \chi, \varepsilon\} \) where \( \hat{\theta}_g = \theta_g + \omega_g, \)
\( \tilde{N}_g = N_g + \nu_g, \) where \( \omega_g \sim L(0, \frac{1}{\chi\varepsilon N_g}), \chi = \max_g [N_g \times L S_{\theta,g}], \nu_g \sim L(0, \frac{1}{\varepsilon}), \) and \( L S_{\theta,g} \) is the maximum amount the statistic changes by changing one observation in cell \( g. \)

See the Online Appendix for a step-by-step guide to implementing this method (accompanied by illustrative Stata code).

**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 use 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 replace at least one high and low point with the next most extreme values. We found that winsorization substantially reduced privacy loss while having a minimal influence on our estimates, by directly reducing the influence of outliers and sensitivity \( \Delta \theta_g. \) (XXRaj added this sentence to clarify) Note that one must include the Winsorization in the calculation of local sensitivity, that is one must add the additional point to the pre-Windsorized data and then Winsorize (as one had, but including the adding point) before running the regression.

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.

---

7 In practice, it may be convenient to use a Normal distribution with standard deviation (matching the SD of the LaPlace) instead of the LaPlace distribution to simplify downstream estimation.

8 Standard errors in each cell can be released using analogous methods; see the Appendix for details.
Fourth, we estimated the scaling parameter $\chi$ 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. We chose the privacy parameter $\varepsilon$ 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 actually in the true top or bottom tail in the confidential data, for different values of $\varepsilon$. After plotting these error rates vs. $\varepsilon$ 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 $\omega_g$ 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 et al. 2018). Normally distributed noise is convenient for downstream statistical inference, e.g., the construction of confidence intervals or Bayesian shrinkage estimators.

IV 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 $\varepsilon$. In contrast, our maximum observed sensitivity approach reduces the dimensionality of the statistics that create uncontrolled privacy risks to one number ($\chi$). Moreover, that number can typically be estimated in a sufficiently large sample that its release could reasonably be viewed as posing negligible privacy risk.
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, these authors show that black women who grow up in Census tracts with more single parents have significantly higher teenage birth rates. Figure 3a replicates this result 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 those tracts in 2000 (when the women in the sample were approximately 20 years old).

There is a clear positive relationship between the two variables: an OLS regression implies that an 1 pp increase in single parent shares is associated with an 0.136 pp increase in teenage birth rates for black women growing up in low-income families. 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.

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 4) teenage births occur (Washington State Department of Health 2018). A count of 0 is typically not suppressed because it is viewed as posing minimal disclosure risk. We mimic this rule in the Opportunity Atlas by omitting tracts where the implied number of teenage births to black women in a cell is between 1 and 3 (inclusive).

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

9We restrict the sample to tracts with a poverty rate of less than 7% based on the 2000 Decennial Census. See notes to Figure 3.

10Raw 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.

11For simplicity, we conduct this analysis in the publicly available Opportunity Atlas data. To do so, we approximate the count of teenage birth rates as the product of the published predicted teenage birth rate at the 25th percentile of parent income and the count of the number of black women in the cell. Since the published rates include noise added to protect privacy, the resulting product is not generically an integer. To approximate what count-based suppression on the confidential data would have produced, we suppress cells if \( \theta_g \times N_g \in (0.5, 3.5) \).

12We suppress tracts where the numerator of the teenage birth rate for low income Black women lies between 1 and 4, which in practice means suppressing tracts where the implied numerator (calculated as
statistically distinguishable from 0 at the five-percent level. 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. Hence, Chetty et al. [2018] would have entirely missed this result had they used data that followed standard cell-suppression techniques – illustrating that our noise infusion approach has significant advantages in terms of statistical bias not just 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 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 product of the publicly available teenage birth rate, total count of Black women and fraction of Black women below median income) lies in the interval [0.5, 4.5). See notes to Figure 3.
privacy noise variance share is significantly higher, but it is still always smaller than the noise due to sampling error.

Of course, noise infusion could have larger effects on reliability 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 III.

V 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 released at a broader level of aggregation (e.g., at the state or national level). Here, we argued on an intuitive basis that the release of such statistics must come at minimal cost, but formalizing this idea – perhaps by restrictions on distributions or the set of estimators – could provide a way to offer provable guarantees of privacy. More broadly, developing differential privacy techniques that can be applied to many estimators without requiring users to develop new algorithms for each application – as we have done here – may help increase the use of such methods in social science.
References


Zack Cooper, Stuart Craig, Martin Gaynor, and John Van Reenen. The price ain’t right? hospital prices and health spending on the privately insured. 2015.


15


Appendix: Implementation Guide

This guide provides step-by-step instructions for implementing the noise-infusion algorithm in Chetty and Friedman (2019) to publicly 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. We also provide illustrative Stata code that implements the five steps below in the context of a regression estimate.

Step 0. **Estimate the statistic** \( \theta \) you are interested in releasing – e.g., a coefficient or parameter from a regression or other statistical model – in the confidential data (XXRaj edited to say “coefficient or parameter”).

   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.

   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 than ordinary least squares (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 statistic \( \theta_g \) in each cell \( g \) of your data, defined as the largest absolute change in \( \theta \) 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 \( g \) (XXRaj edited to permit adding or removing).
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. (XXRaj added the following two sentences per Salil) In high-dimensional dataspaces, removing a point may be more computationally feasible, although it is especially important to reduce the influence of outliers in such a case. Note that Windsorizing each variable independently (using its marginal distribution) may not eliminate outliers in multiple dimensions.

b. 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 \( \chi \):

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

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

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

Step 3. Determine the privacy parameter \( \varepsilon \) for your release using one of the following methods:

a. Follow established guidelines on \( \varepsilon \) from your data provider.

b. Choose \( \varepsilon \) by plotting the social gain from greater accuracy vs. \( \varepsilon \) and choosing a value of \( \varepsilon \) 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 \( (\tilde{\theta}_g - \theta_g)^2 \) 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 \( \theta_g \) falls below a certain threshold conditional on the noise-infused value \( \tilde{\theta}_g \) falling above that threshold. For example, one might
calculate the probability that \( \theta_g \) falls outside the top 10% of the distribution of \( \{ \theta_g \} \) conditional on observing \( \tilde{\theta}_g \) in the top 10% of the distribution of \( \{ \tilde{\theta}_g \} \).

Step 4. **Add random noise** proportional to maximum observed sensitivity \( \chi \) and the privacy parameter \( \varepsilon \) to each statistic:

\[
\tilde{\theta}_g = \theta_g + \sqrt{2} \frac{\chi}{\varepsilon N_g} \omega,
\]

where \( \omega \) is a random variable with mean 0 and standard deviation 1.

a. The distribution of \( \omega \) 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 \( L(0, \sqrt{2}) \) conforms more precisely to the desired privacy loss limit at all points in the distribution.

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

\[
SE(\tilde{\theta}_g) = \sqrt{SE(\theta_g)^2 + 2 \left( \frac{\chi}{\varepsilon N_g} \right)^2}.
\]

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

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

\[
\tilde{N}_g = N_g + \sqrt{2} \frac{\omega}{\varepsilon}.
\]

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, \( 2E[(\frac{\chi}{\varepsilon N_g})^2]/Var(\tilde{\theta}_g) \), where the expectation is taken over the cells in the dataset.

Step 5. **Release the noise-infused statistics** \( \{ \tilde{\theta}_g \} \) and \( \{ \tilde{SE}(\tilde{\theta}_g) \} \), counts \( \{ \tilde{N}_g \} \), and parameters that control the amount of noise added (\( \varepsilon \) and \( \chi \)) for the groups of interest publicly.

*Questions? Email info@opportunityinsights.org*
Table 1
Variance Decomposition for Tract-Level Estimates: Selected Outcomes and Demographic Groups

<table>
<thead>
<tr>
<th>Panel A. Teenage Birth Rate, for Daughters of Parents at the 25th Percentile</th>
<th>Signal Variance (1)</th>
<th>Sampling Noise Variance (2)</th>
<th>Privacy Noise Variance (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black Females</td>
<td>71.00%</td>
<td>28.18%</td>
<td>0.82%</td>
</tr>
<tr>
<td>White Females</td>
<td>70.58%</td>
<td>28.31%</td>
<td>1.11%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B. Share Incarcerated, for Sons of Parents at the 25th Percentile</th>
<th>Signal Variance (1)</th>
<th>Sampling Noise Variance (2)</th>
<th>Privacy Noise Variance (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black Males</td>
<td>56.39%</td>
<td>40.21%</td>
<td>2.32%</td>
</tr>
<tr>
<td>White Males</td>
<td>33.64%</td>
<td>38.85%</td>
<td>27.51%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C. Household Income, for Children of Parents at the 25th Percentile</th>
<th>Signal Variance (1)</th>
<th>Sampling Noise Variance (2)</th>
<th>Privacy Noise Variance (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Children</td>
<td>90.96%</td>
<td>8.97%</td>
<td>0.08%</td>
</tr>
<tr>
<td>White Children</td>
<td>78.90%</td>
<td>20.82%</td>
<td>0.27%</td>
</tr>
<tr>
<td>Black Children</td>
<td>75.19%</td>
<td>23.59%</td>
<td>1.22%</td>
</tr>
<tr>
<td>Hispanic Children</td>
<td>69.62%</td>
<td>29.08%</td>
<td>1.30%</td>
</tr>
<tr>
<td>Asian Children</td>
<td>69.99%</td>
<td>28.94%</td>
<td>1.06%</td>
</tr>
<tr>
<td>American Indian &amp; Alaska Native Children</td>
<td>81.90%</td>
<td>17.28%</td>
<td>0.82%</td>
</tr>
<tr>
<td>White Males</td>
<td>64.71%</td>
<td>34.60%</td>
<td>0.69%</td>
</tr>
<tr>
<td>White Females</td>
<td>70.36%</td>
<td>28.93%</td>
<td>0.71%</td>
</tr>
<tr>
<td>Black Males</td>
<td>66.58%</td>
<td>31.29%</td>
<td>2.13%</td>
</tr>
<tr>
<td>Black Females</td>
<td>73.16%</td>
<td>25.17%</td>
<td>1.67%</td>
</tr>
</tbody>
</table>

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 who grew up in each tract to parents at the 25th percentile of the parental income distribution. We decompose the total tract-level variance into the signal variance, the noise variance from sampling variation, and the noise variance from infused privacy noise; each row in the table represents this decomposition for a particular outcome variable and demographic group, and the three percentages add to 100% across each row (as the three variance components are independent). To calculate the decomposition, we first decompose the total variance into the total noise variance, equal to the average squared standard error, and the signal variance (reported in Column 1, and equal to the total variance minus total noise variance). We further decompose the total noise variance into two orthogonal components, the privacy noise variance (reported in Column 3, and calculated from the sensitivity scaling parameter used in our application of the MOSE algorithm and the tract-specific counts) and the sampling noise variance (reported in Column 2, and calculated as the total noise variance minus the privacy noise variance). Panel A presents this variance decomposition for teenage birth rate, for the demographic subgroup specified in the first column, following the U.S. Census official definitions of race and Hispanic identity (all non-Hispanic groups refer to non-Hispanic individuals of a particular race). Panels B and C replicate Panel A using the share incarcerated and child household income rank, respectively. The first row (of Panel A) represent the same data depicted in Figure 3.


Notes: This figure demonstrates the calculation of local sensitivity in a hypothetical tract of twenty individuals, each of whom has a parental income rank (x-axis) and child income rank (y-axis). The parameter of interest is the predicted value of child income rank at the 25th 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). The square point denotes this predicted value in the raw data. In order to calculate sensitivity, we add a single point to the data and recalculate the predicted value; local sensitivity is defined by the maximum absolute change in the predicted value from adding a single point. In this example, that occurs when adding the point (0,1), shown as the hollow dot in the upper left corner. With the addition of that point, the estimated regression line shifts to the dashed line, increasing the predicted value.
FIGURE 2: Maximum Observed Sensitivity Envelope

Notes: This figure demonstrates our calculation of the Maximum Observed Sensitivity Envelope (MOSE) for a hypothetical data set. To construct this figure, we calculate the local sensitivity within each tract (see Figure 1 notes), and then plot the local sensitivity (y-axis) against the number of individuals in the tract (x-axis). In this figure, we use log-scaling on both axes. The MOSE is calculated as the function $\frac{\chi}{N}$ with the smallest value of $\chi$ such that the MOSE is weakly greater than local sensitivity in each tract when evaluated at that tract’s count.
FIGURE 3: Association between Teenage Birth and Single Parent Share

A. Noise-Infused Data

\begin{align*}
\text{Slope} &= 0.136 \\
(0.015)
\end{align*}

B. Numerator-Suppressed Data

\begin{align*}
\text{Slope} &= 0.028 \\
(0.017)
\end{align*}

Notes: This figure plots the relationship between the estimated teenage birth rate, for Black women born in 1978-1983 and raised by parents in this tract at the 25th percentile of the parental income distribution, and the fraction of single headed households in that tract according to the 2000 Decennial Census. Each observation represents a single Census tract, and we restrict the sample to tracts with a poverty rate of less than 7% based on the 2000 Decennial Census. The figure presents
a binned scatterplot as a non-parametric representation of the underlying relationship; we bin tracts into 20 groups based on the single-parent share, weighting each tract by the number of Black women used in the calculation of the teenage birth rate. Each dot then plots the average teenage birth rate (y-axis) against the average single-parent share (x-axis) in each of the twenty bins. We calculate the best fit line and raw correlation using a weighted OLS regression on the tract-level data. Panel A shows this relationship using noise-infused data, which is publicly available in the Opportunity Atlas. Panel B shows this relationship using numerator suppression of the Opportunity Atlas data; we suppress tracts where the numerator of the teenage birth rate for low income Black women lies between 1 and 4, which in practice means suppressing tracts where the implied numerator (calculated as the product of the publicly available teenage birth rate, total count of Black women and fraction of Black women below median income) lies in the interval \([0.5, 4.5]\). In Panel A, we calculate the signal correlation by rescaling the raw correlation by the reliability ratio, i.e. the ratio of signal variance to total variance of the teenage birth estimates. See Chetty et al. (2018) for more details on this calculation.
**APPENDIX FIGURE 1: Global Sensitivity**

Notes: This figure replicates Figure 1, except in a contrived case where the addition of a single observation can have an arbitrarily large effect on the predicted value at the 25th percentile of parent income, which occurs as the variance of the dots on the x-axis shrinks to be arbitrarily small.
APPENDIX FIGURE 2: Teenage Birth Rates for Black Women

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 plots a binscatter (see Figure 3 for more details) of the relationship between teenage birth rates for Black women and two measures of tract-specific counts. Panel A plots teenage birth rates for Black women vs. the number of Black women with teenage births in the tract (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 in the tract (i.e., the denominator of the teenage birth rate).