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

by Raj Chetty and John N. Friedman*

Social scientists increasingly use confidential data to publish statistics based on small samples, from descriptive statistics on income distributions and health expenditures in small areas to estimates of the causal effects of specific schools and hospitals. Such statistics allow researchers and policymakers to answer important questions, but also raise concerns about privacy loss – the potential disclosure of information about a specific individual.

In this paper, we develop an easily implementable method to reduce privacy loss when disclosing statistics such as OLS regression estimates based on small samples. 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.[1] Building on ideas from the differential privacy literature (Dwork et al. 2006), 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. We then show that our method outperforms widely used methods of disclosure limitation such as count-based cell suppression both in terms of privacy loss and statistical bias.

This paper briefly summarizes our method and its properties. Further details, including a step-by-step guide and illustrative Stata code to implement the algorithm as well as a discussion of how we used this method to release estimates of social mobility by Census tract in the Opportunity Atlas, are available in the longer version of this manuscript (Chetty and Friedman 2019).

I. 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

*Chetty: Harvard University, 1280 Massachusetts Ave., Cambridge, MA 02138 (e-mail: chetty@fas.harvard.edu); Friedman: Brown University, Robinson Hall 304B (e-mail: john_friedman@brown.edu). 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 Federico Gonzalez, Jamie Gracie, Martin Koenen, Kate Musen, Daniel Reuter, Jesse Silbert, 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.

[1] 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 follow the approach described below.

applied to any statistic, we focus for concreteness on the problem of releasing predicted values (θ_g) from univariate regressions that are estimated in subgroups of the data, indexed by g. For example, Chetty et al. [2018] regress children's income ranks (y_{ig}) in adulthood on their parents' income ranks (x_{ig}) by Census tract g and release predicted values from these regressions. Because each Census tract contains relatively few observations, releasing $\left\{\theta_g\right\}$ 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 $\left\{\theta_g\right\}$. 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. 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 $D_g = \left\{x_{ig}, y_{ig}\right\}$ 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

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

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 almost surely be infinite, 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 (1), effectively providing a "worst case" guarantee on the degree of privacy

loss. A privacy algorithm is "ε-differentially private" if

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

The parameter ε can be interpreted as the maximum risk one is willing to tolerate when releasing the statistic of interest. It is straightforward to show that one can achieve the bound in (2) by adding Laplacian noise $\omega_g \sim L\left(0, \frac{\Delta\theta_g}{\varepsilon}\right)$ to the estimates, where

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

is the "sensitivity" of the statistic θ_g (Dwork et al. 2006). 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 (i.e., sensitivity is high). One must therefore add a large amount of noise to protect her privacy when releasing statistics on mean wealth.

If sensitivity $\Delta\theta_g$ were publicly known, one could obtain differentially private statistics that satisfy any privacy loss threshold ε simply by adding noise $\omega_g \sim L\left(0, \frac{\Delta\theta_g}{\varepsilon}\right)$ to the statistics one seeks to release.[2] In practice, $\Delta\theta_g$ is not known; hence, the key 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. Since the 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. This global-sensitivity approach has

[2] 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).

been used to release simple statistics such as counts and means (Dwork et al. 2006).[3]

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) can grow 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. Thus, global sensitivity is infinite, implying that adding any finite amount of noise will not meet the differential privacy guarantee in (2).

II. Maximum Observed Sensitivity Algorithm

The problem with global sensitivity is that empirically unrealistic but theoretically feasible data configurations drive sensitivity to infinity. We propose an algorithm that instead focuses on values of sensitivity that are empirically relevant. Our approach is conceptually similar 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.

Local Sensitivity. The starting point for our algorithm is measuring *local sensitivity*, 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 $\theta_g = 0.212$.

To compute local sensitivity, we recalculate this predicted value, adding or removing points one by one. 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. When 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

[3]Existing methods to protect privacy when disclosing more complex estimators – such as regression or quasi-experimental estimators – rely on either asymptotic results in large samples or the use of robust statistics such as median regression, limiting their application in social science.

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 itself creates a privacy risk. 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).

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.

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. 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 $\left\{\tilde{\theta}_g\right\}$.

Our maximum observed sensitivity 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). 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 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.[4]

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

$$\omega_g \sim L(0, \frac{\chi}{\varepsilon N_g})$$

follows a LaPlace 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. In addition, release the cell-specific counts as $\tilde{N}_g = N_g + \nu_g$, where $\upsilon_g \sim L(0, \frac{1}{\varepsilon})$.

III. Comparison to Current Methods of Disclosure Limitation

In this section, we compare the properties of our noise infusion approach to 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 from adjacent datasets that differ by a single observation. For example, even when one suppresses cells with a count of fewer than say 100 individuals, one could recover a single individual's income by releasing a mean over 150 individuals and a mean over 151 individuals and differencing the two statistics. 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).

[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.

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. 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 black women growing up in low-income families in low-poverty areas.

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. 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 data by omitting tracts where black women raised in low-income families have between 1 and 4 teenage births (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 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, 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; tracts with a small black population in turn are more likely to be suppressed and also tend to be areas with few single parents. 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.

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. There, 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. For example, just 0.8% of the total variance across tracts in teenage birth rates for black women raised in low-income families comes from the noise that was added to protect privacy. 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. 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.

IV. Conclusion

This paper has developed a practical method for reducing the privacy loss from disclosing statistics based on confidential data that outperforms existing methods of disclosure limitation *both* in terms of privacy loss and statistical bias. The method 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.

References

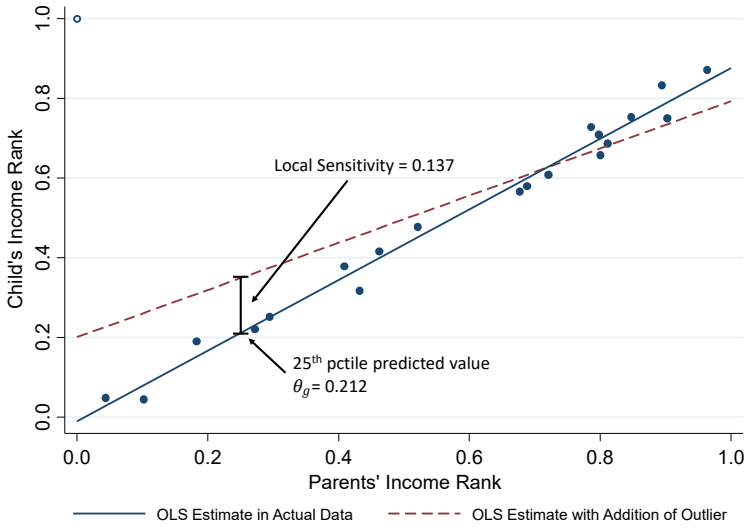
John M. Abowd and Ian M. Schmutte. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. American Economic Review, 109(1):171–202, January 2019.

Raj Chetty and John N Friedman. A Practical Method to Reduce Privacy Loss when Disclosing Statistics Based on Small Cells. Working Paper XX, National Bureau of Economic Research, Forthcoming 2019.

Raj Chetty, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility. Working Paper 25147, National Bureau of Economic Research, October 2018.

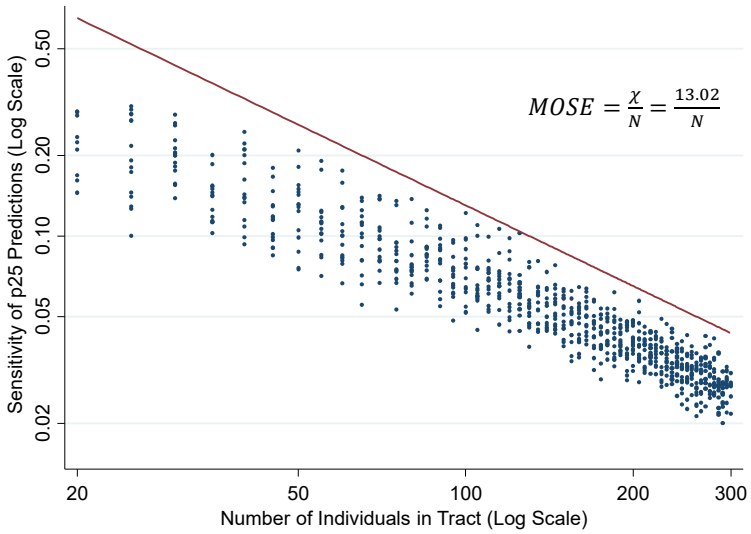
Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In Shai Halevi and Tal Rabin, editors, Theory of Cryptography, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

FIGURE 1: Calculation of Local Sensitivity



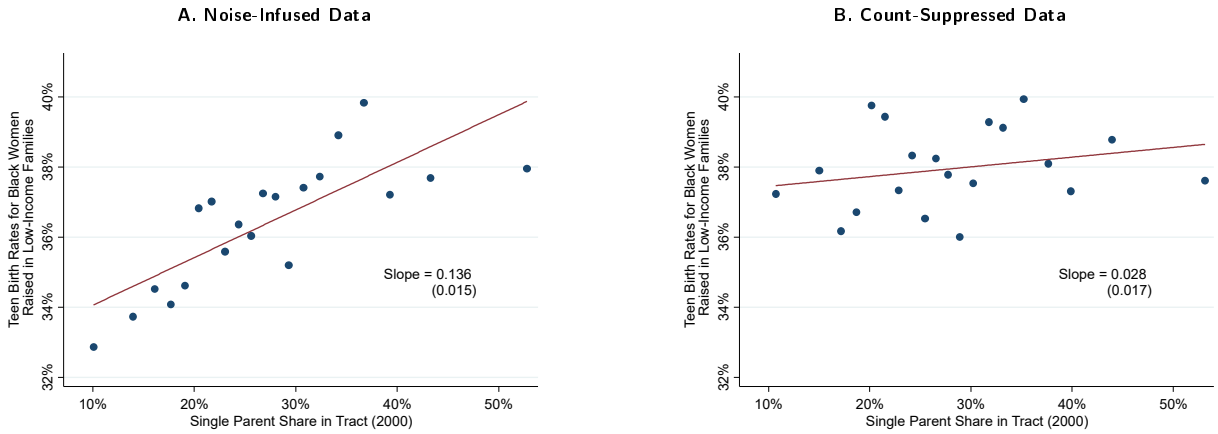
Notes: This figure shows how we calculate local sensitivity in a hypothetical cell (Census tract) with 20 individuals. The figure presents a scatter plot of children's income ranks in adulthood vs. their parents' income rank. The parameter of interest (θ_g) is the predicted value of child income rank at the 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). 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.

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 $\chi = max_g\left[N_g \times LS_{\theta,g}\right] = 13.02$ in this example.

FIGURE 3: Association between Teenage Birth Rates and Single Parent Shares Across Census Tracts



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 25th 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. 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 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 $[0.5, 4.5)$.