American Political Science Review (2023) 1–16
doi:10.1017/S0003055422001411

Statistically Valid Inferences from Privacy-Protected Data

GEORGINA EVANS *Harvard University, United States*
GARY KING *Harvard University, United States*
MARGARET SCHWENZFEIER *Harvard University, United States*
ABHRADEEP THAKURTA *Google Brain, United States*

Unprecedented quantities of data that could help social scientists understand and ameliorate the challenges of human society are presently locked away inside companies, governments, and other organizations, in part because of privacy concerns. We address this problem with a general-purpose data access and analysis system with mathematical guarantees of privacy for research subjects, and statistical validity guarantees for researchers seeking social science insights. We build on the standard of "differential privacy," correct for biases induced by the privacy-preserving procedures, provide a proper accounting of uncertainty, and impose minimal constraints on the choice of statistical methods and quantities estimated. We illustrate by replicating key analyses from two recent published articles and show how we can obtain approximately the same substantive results while simultaneously protecting privacy. Our approach is simple to use and computationally efficient; we also offer open-source software that implements all our methods.

INTRODUCTION

Just as more powerful telescopes empower astronomers, the accelerating influx of data about the political, social, and economic worlds has enabled considerable progress in understanding and ameliorating the challenges of human society. Yet, although we have more data than ever before, we may now have a smaller fraction of the data in the world than ever before because huge amounts are now locked up inside private companies, governments, political campaigns, hospitals, and other organizations, in part because of privacy concerns. In a study of corporate data sharing with academics, "Privacy and security were cited as the top concern for companies that hold personal data because of the serious risk of re-identification," and government regulators obviously agree (FPF 2017, 11; see also King and Persily 2020). If we are to do our jobs as social scientists, we have no choice but to find ways of unlocking these datasets, as well as sharing more seamlessly with other researchers. We might hope that government or society will take actions to support this mission, but we can also take responsibility ourselves and begin to develop technological solutions to these political problems.

Among all the academic fields, political science scholarship may be especially affected by the rise in the concerns over privacy because so many of the issues are inherently political. Consider an authoritarian government seeking out its opponents; a democratic government creating an enemies list; tax authorities (or an ex spouse in divorce proceedings) searching for an aspiring politician's unreported income sources; an employer trying to weed out employees with certain political beliefs; or even political candidates searching for private information to tune advertising campaigns. The same respondents may also be hurt by political, financial, sexual, or health scams made easier by using illicitly obtained personal information to construct phishing email attacks. Political science access to data from business, governments, and other organizations may be particularly difficult to ensure because the subject we study—politics—is also the reason for our lack of access, as is clear from analyses of decades of public opinion polls, laws, and regulations (Robbin 2001).

We develop methods to foster an emerging change in the paradigm for sharing research information. Under the familiar *data sharing regime*, data providers protect the privacy of those in the data via de-identification (removing readily identifiable personal information such as names and addresses) and then simply giving a copy to trusted researchers, perhaps with a legal agreement. Yet, with the public's increasing concerns over privacy, an increasingly hard line taking by regulators worldwide, and data holders' (companies, governments, researchers, and others) need to respond, this regime is failing. Fueling these

Georgina Evans, PhD Candidate, Department of Government, Harvard University, United States, georgieaevans@gmail.com.
Gary King, Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University, United States, King@ Harvard.edu.
Margaret Schwenzfeier, PhD Candidate, Department of Government, Harvard University, United States, schwenzfeier@g.harvard.edu.
Abhradeep Thakurta, Senior Research Scientist, Google Brain, guha@google.com.

Received: January 31, 2021; revised: August 22, 2021; accepted: November 16, 2022.

concerns is a new computer science subfield showing that, although de-identification surely makes it harder to learn the personal information of some research subjects, it offers no guarantee of thwarting a determined attacker (Henriksen-Bulmer and Jeary 2016).

For example, Sweeney (1997) demonstrates that knowing only a respondent's gender, zip code, and birth date is sufficient to personally identify 87% of the U.S. population, and most datasets of interest to political scientists are far more informative. For another example, the U.S. Census was able to re-identify the personal answers of as many as three-quarters of Americans from publicly released and supposedly anonymous 2010 census data (Abowd 2018; see also https://bit.ly/privCC). Unfortunately, other privacy-protection techniques used in the social sciences—such as aggregation, legal agreements, data clean rooms, query auditing, restricted viewing, and paired programmer models—can often be broken by intentional attack (Dwork and Roth 2014). And not only does the venerable practice of trusting researchers to follow the rules fail spectacularly at times (like the Cambridge Analytica scandal, sparked by the behavior of a single social scientist), but it turns out that even trusting a researcher who is known to be trustworthy does not always guarantee privacy (Dwork and Ullman 2018). Fast-growing recent evidence is causing many other major organizations to change policies as well (e.g., Al Aziz et al. 2017; Carlini et al. 2020).

An alternative approach to the data sharing regime that may help persuade some data holders to allow academic research is the two-part *data access regime.* In the first part, the confidential data reside on a trusted computer server protected by best practices in cybersecurity, just as it does before sharing under the data sharing regime. The distinctive aspect of the data access regime is its second step, which treats researchers as potential "adversaries" (i.e., who may try to violate respondents' privacy while supposedly doing their job seeking knowledge for public good) and thus adds mathematical guarantees of the privacy of research subjects. To provide these guarantees, we construct a "differentially private" algorithm that makes it possible for researchers to discover population-level insights but impossible to reliably detect the effect of the inclusion or exclusion of any one individual in the dataset or the value of any one person's variables. Researchers are permitted to run statistical analyses on the server and receive "noisy" results computed by this privacy-preserving algorithm (but are limited by the total number of runs so they cannot repeat the same query and average away the noise).

Differential privacy is a widely accepted mathematical standard for data access systems that promises to avoid some of the zero-sum policy debates over balancing the interests of individuals with the public good that can come from research. It also seems to satisfy regulators and others. Differential privacy was introduced by Dwork et al. (2006) and generalizes the social science technique of "randomized response" to elicit sensitive information in surveys; it does this by randomizing the answer to a question rather than the question itself (see Evans et al. Forthcoming). See Dwork and Roth (2014) and Vadhan (2017) for overviews and Wood et al. (2018) for a nontechnical introduction.

A fast-growing literature has formed around differential privacy, seeking to balance privacy and utility, but the current measures of "utility" provide little utility to social scientists or other statistical analysts. Statistical inference in our field usually involves choosing a target *population* of interest, identifying the *data generation process*, and then using the resulting *dataset* to learn about features of the population. Valid inferences require methods with known statistical properties (such as identification, along with unbiasedness, and consistency) and honest assessments of uncertainty (e.g., standard errors). In contrast, privacy researchers typically begin with the choice of a target (confidential) *dataset*, add *privacy-protective procedures*, and then use the resulting *differentially private dataset or analyses* to infer to the confidential dataset—usually without regard to the data generation process or valid population inferences. This approach is useful for designing privacy algorithms but, as Wasserman (2012, 52) puts it, "I don't know of a single statistician in the world who would analyze data this way." Because social scientists understand not to confuse internal and external validity, they will aim to infer, not to the data they would see without privacy protections, but to the world from which the data were generated.[1]

To make matters worse, although the privacy-protective procedures introduced by differential privacy work well for their intended purpose, they induce severe bias in estimating population quantities of interest to social scientists. These procedures include *adding random error*, which induces measurement error bias, and *censoring* (known as "clamping" in computer science), which when uncorrected induces selection bias (Blackwell, Honaker, and King 2017; Stefanski 2000; Winship and Mare 1992). We have not found a single prior study that tries to correct for both (although some avoid the effects of censoring in theory at the cost of considerable additional noise and far larger standard errors in practice; Karwa and Vadhan 2017; Smith 2011) and few have uncertainty estimates. This is crucial because inferentially invalid data access systems can harm societies, organizations, and individuals—such as by inadvertently encouraging the distribution of misleading medical, policy, scientific, and other conclusions—even if it

[1] In other words, "In statistical inference the sole source of randomness lies in the underlying model of data generation, whereas the estimators themselves are a deterministic function of the dataset. In contrast, differentially private estimators are inherently random *in their computation.* Statistical inference that considers *both* the randomness in the data and the randomness in the computation is highly uncommon" (Sheffet 2017, 3107). As Karwa and Vadhan (2017) write, "Ignoring the noise introduced for privacy can result in wildly incorrect results at finite sample sizes…this can have severe consequences." On the essential role of inference and uncertainty in science, see King, Keohane, and Verba (1994).

successfully protects individual privacy. For these reasons, using existing differentially private systems needs correction before being used in social science analysis.[2]

Social scientists and others need algorithms on data access systems with both inferential validity and differential privacy. We offer one such algorithm that is approximately unbiased with sufficient prior information, has lower variance than uncorrected estimates, and comes with accurate uncertainty estimates. The algorithm also turns censoring from a problem that severely increases bias or inefficiency in statistical estimation in order to protect privacy to an attractive feature that greatly reduces the amount of noise needed to protect privacy while still leaving estimates approximately unbiased. The algorithm is easy to implement, computationally efficient even for very large datasets, and, because the entire dataset never needs to be stored in the same place, may offer additional security protections.

Our algorithm is *generic*, designed to minimally restrict the choice among statistical procedures, quantities of interest, data generating processes, and statistical modeling assumptions. The algorithm may therefore be especially well suited for building research data access systems. When valid inferential methods exist or are developed for more restricted use cases, they may sometimes allow less noise for the same privacy guarantee. As such, one productive plan is to first implement our algorithm and to then gradually add these more specific approaches when they become available as preferred choices.[3]

We begin, in the next section, with an introduction to differential privacy and description of the inferential challenges in analyzing data from a differentially private data access system. We then give a generic differentially private algorithm which, like most such algorithms, is statistically biased. We therefore introduce bias corrections and variance estimators which, together with the private algorithm, accomplishes our goals. Technical details appear in the Supplementary Material. We illustrate the performance of our approach in finite samples via Monte Carlo simulations; we then replicate key results from two recent published articles and shows how to obtain almost the same conclusions while guaranteeing the privacy of all research subjects. The limitations of our methodology must be understood to use it appropriately, a topic we take up throughout the paper, while discussing practical advice for implementation, and in our Supplementary Material. As a companion to this paper, we offer open-source software (called UnbiasedPrivacy) to illustrate all the methods described herein.

DIFFERENTIAL PRIVACY AND ITS INFERENTIAL CHALLENGES

We now define the differential privacy standard, describe its strengths, and highlight the challenges it poses for proper statistical inference. Throughout, we modify notation standard in computer science so that it is more familiar to social scientists.

Definitions

Begin with a confidential dataset D, defined as a collection of N rows of numerical measurements constructed so that each individual whose privacy is to be protected is represented in at most one row.[4]

Statistical analysts would normally calculate a statistic s (such as a count, mean, and parameter estimate) from the data D, which we write as $s(D)$. In contrast, under differential privacy, we construct a privacy-protected version of the same statistic, called a "mechanism" $M(s, D)$, by injecting noise and censoring at some point before returning the result (so even if we treat $s(D)$ as fixed, $M(s, D)$ is random). As we will show, the specific types of noise and censoring are specially designed to satisfy differential privacy.

The core idea behind the differential privacy standard is to prevent a researcher from reliably learning anything different from a dataset regardless of whether an individual has been included or excluded. To formalize this notion, consider two datasets D and D' that differ in at most one row. Then, the standard requires that the probability (or probability density) of any analysis result m from dataset D, $\Pr[M(s, D) = m]$, be *indistinguishable* from the probability that the same result is produced by the same analysis of dataset D', $\Pr[M(s, D') = m]$, where the probabilities take D as fixed and are computed over the noise.

We write an intuitive version of the differential privacy standard (using the fact that $e^{\epsilon} \approx 1 + \epsilon$ for small ϵ) by defining "indistinguishable" as the ratio of the

[2] Inferential issues also affect differential privacy applications outside of data access systems (the so-called central model). These include "local model" systems where private calculations are made on a user's system and sent back to a company in a way that prevents it from making reliable inferences about individuals—including Google's Chrome (Erlingsson, Pihur, and Korolova 2014) and their other products (Wilson et al. 2019), Apple's MacOS (Tang et al. 2017), and Microsoft's Windows (Ding, Kulkarni, and Yekhanin 2017)—and "hybrid" systems to release differentially private datasets such as from Facebook (Evans and King 2023) and the U.S. Census Bureau (Garfinkel, Abowd, and Powazek 2018).

[3] For example, Karwa and Vadhan (2017) develop finite-sample confidence intervals with proper coverage for the mean of a normal density; Barrientos et al. (2019) offer differentially private significance tests for linear regression; Gaboardi et al. (2016) propose chi-squared goodness-of-fit tests for multinomial data and independence between two categorical variables; Wang, Lee, and Kifer (2015) propose accurate p-values for chi-squared tests of independence between two variables in tabular data; Wang, Kifer, and Lee (2018) develop differentially private confidence intervals for objective or output perturbation; and Williams and McSherry (2010) provide an elegant marginal likelihood approach for moderate sized datasets.

[4] Hierarchical data structures, or dependence among units, are allowed within but not between rows. For example, rows could represent a family with variables for different family members. Although N could leak privacy in unusual situations (in which case it could be disclosed in a differentially private manner with our algorithm), for simplicity of exposition, we do not treat N as private.

probabilities falling within ϵ of equality (which is 1). Thus, a mechanism is said to be ϵ-differentially private if

$$\frac{\Pr[M(s,D)=m]}{\Pr[M(s,D')=m]} \in 1 \pm \epsilon, \tag{1}$$

where ϵ is a pre-chosen level of possible privacy leakage, with smaller values potentially giving away less privacy (by requiring more noise or censoring).[5] Many variations and extensions of Equation 1 have been proposed (Desfontaines and Pejó 2019). We use the most popular, known as "(ϵ, δ)-differential privacy" or "approximate differential privacy," which adds a small chosen offset δ to the numerator of the ratio in Equation 1, with ϵ remaining the main privacy parameter. This second privacy parameter, which the user sets to a small value such that $\delta < 1/N$, allows mechanisms with (statistically convenient) Gaussian (i.e., Normal) noise processes. This relaxation also has Bayesian interpretations, with the posterior distribution of $M(s,D)$ close to that of $M(s,D')$, and also that an (ϵ, δ)-differentially private mechanism is ϵ-differentially private with probability at least $1-\delta$ (Vadhan 2017, 355ff). We can also express approximate differential privacy more formally as requiring that each of the probabilities be bounded by a linear function of the other:[6]

$$\Pr[M(s,D)=m] \leq \delta + e^{\epsilon} \cdot \Pr[M(s,D')=m]. \tag{2}$$

Consistent with political science research showing that secrecy is best thought of as continuous rather than dichotomous (Roberts 2018), the differential privacy standard quantifies privacy leakage of statistics released via a randomized mechanism through the choice of ϵ (and δ). Differential privacy is expressed in terms of the maximum possible privacy loss, but the expected privacy loss is considerably less than this worst case analysis, often by orders of magnitude (Carlini et al. 2019; Jayaraman and Evans 2019). It also protects small groups in the same way as individuals, with the maximum risk $k\epsilon$ rising linearly in group size k.

Example

To fix ideas, consider a confidential database of donations given by individuals to an organization bent on defeating an authoritarian leader, with the mean donation in a region as the quantity of interest: $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. (The mean is a simple statistic, but we will use the mean to generalize to most other commonly used statistics.) Suppose also that the region includes many poor, and one more wealthy, individual. Our goal is to enable researchers to infer the mean without any material possibility of revealing information about any individual in the region (even if not in the dataset).

Consider three methods of privacy protection. First, aggregation to the mean alone fails as a method of privacy protection: the differential privacy standard requires that *every* individual is protected even in the worst-case scenario, but clearly disclosing $\overline{y}$ may be enough to reveal whether the wealthy individual made a large contribution.

Second, suppose instead we disclose only the sum of $\overline{y}$ and noise (a draw from a mean-zero normal distribution). This mechanism is unbiased, with larger confidence intervals being the cost of privacy protection. The difficulty with this approach is that we may need to add so much noise to protect the one wealthy outlier that our confidence intervals may be too large to be useful.

Finally, we describe the intuitive *Gaussian mechanism* that will solve the problem for this example and will aid in the development of our general purpose methodology. The idea is to first censor donations that are larger than some pre-chosen value Λ, set the censored values to Λ, average the (partially censored) donations, and finally add noise to the average. Censoring has the advantage of requiring less noise to protect individual outliers, but has the disadvantage of inducing statistical bias. (We show how to correct this bias in a subsequent section, turning censoring into an attractive tool enabling both privacy and utility.)

The bound on privacy loss can be entirely quantified via the noise variance, S^2, even though privacy protection comes from both censoring and random noise. This is because, if we choose to censor less by setting Λ to be high, then we must add more random noise. To set the noise variance, we consider the values of ϵ and δ that would be satisfied in Equation 2. Since we are converting one number, S^2, into two, there are many such (ϵ, δ) pairs. (A related notion of differential privacy, known as zero-concentrated differential privacy, captures the privacy loss of the Gaussian mechanism in a single parameter denoted by ρ, but requires a more complicated definition of differential privacy [Bun and Steinke 2016]. Some applications of differential privacy report the privacy loss in terms of ρ, rather than ϵ and δ, and a simple formula can be used to convert between the two.)

To guide intuition about how the noise variance must change to satisfy a particular ϵ privacy level at a fixed δ, it is useful to write

$$S(\Lambda, \epsilon, N; \delta) \propto \frac{\Lambda}{N\epsilon}. \tag{3}$$

Thus, Equation 3 shows that ensuring differential privacy allows us to add less noise if (1) each person is submerged in a sea of many others (larger N), (2) less

[5] Defining "indistinguishable" via this multiplicative (ratio) metric is much more protective of privacy than some others, such as an additive (difference) metric. For example, consider an obviously unacceptable mechanism: "choose one individual uniformly at random and disclose all of his or her data." This mechanism is not differentially private (the ratio can be infinite and thus greater than any finite ϵ), but it may appear safe on an additive metric because the impact of adding or removing one individual on the difference in the probability distribution of a statistical output is proportional to at most $1/N$.

[6] Our algorithms below also satisfy a strong version of approximate differential privacy known as Rényi differential privacy (see Mironov 2017). See also related privacy concepts from the social sciences (Chetty and Friedman 2019) and statistics (Reiter 2012).

privacy is required (we choose a larger ϵ), or (3) more censoring is used (we choose a smaller Λ).

With censoring, $\hat{\theta}$ is a biased estimate of $\overline{y}$: $E(\hat{\theta}) \neq \overline{y}$. To reduce censoring bias, we can choose larger values of Λ, but that unfortunately increases the noise, statistical variance, and uncertainty; however, reducing noise by choosing a smaller value of Λ increases censoring and thus bias. We resolve this tension in our section on ensuring valid inference.

Inferential Challenges

We now discuss four issues differential privacy poses for statistical inference. For some, we offer corrections; for others, we suggest how to adjust statistical analysis practices.

First, censoring induces selection bias. Avoiding censoring by setting Λ large enough adds more noise but is unsatisfactory because any amount of noise induces bias in estimates of all but the simplest quantities of interest. Moreover, even for unbiased estimators (like the uncensored mean above), the added noise makes unadjusted standard errors statistically inconsistent. Ignoring either measurement error bias or selection bias is a major inferential mistake as it may change substantive conclusions, estimator properties, or the validity of uncertainty estimates, often in negative, unknown, or surprising ways.[7]

Second, it would be sufficient for a proper scientific statements to have (1) an estimator $\hat{\theta}$ with known statistical properties (such as unbiasedness, consistency, and efficiency) and (2) accurate uncertainty estimates. Unfortunately, as Appendix B of the Supplementary Material demonstrates, accurate uncertainty estimates cannot be constructed by using differentially private versions of classical uncertainty estimates, meaning we must develop new approaches.

Third, to avoid researchers rerunning the same analysis many times and averaging away the noise, their analyses must be limited. This limitation is formalized via a differential privacy property known as *composition*: if mechanism k is (ϵ_k, δ_k)-differentially private, for $k = 1, \ldots, K$, then disclosing all K estimates is $(\sum_{k=1}^{K} \epsilon_k, \sum_{k=1}^{K} \delta_k)$-differentially private.[8] The composition property enables the data provider to enforce a *privacy budget* by allocating a *total* value of ϵ to a researcher who can then divide it up and run as many analyses, of whatever type, as they choose, so long as the sum of all the ϵs across all their analyses does not exceed their total privacy budget.

Enabling the researcher to decide how to allocate a privacy budget enables scarce information to be better directed to scholarly goals than a central authority such as the data provider. However, when the total privacy budget is used up, no researcher can run new analyses unless the data provider chooses to increase the budget. This constraint is useful to protect privacy but utterly changes the nature of statistical analysis. To see this, note that best practice recommendations have long included trying to avoid being fooled *by the data*—by running every possible diagnostic and statistical check, fully exploring the dataset—and *by the researcher's personal biases*—such as by preregistration to constrain the number of analyses run to eliminate "p-hacking" or correcting for "multiple comparisons" ex post (Monogan 2015). One obviously needs to avoid being fooled in any way, and so researchers normally try to balance the resulting contradictory advice. In contrast, differential privacy tips the scales: remarkably, it makes solving the second problem almost automatic (Dwork et al. 2015), but it also reduces the probability of serendipitous discovery and increases the odds of being fooled by unanticipated data problems. Successful data analysis with differential privacy thus requires careful planning, although less stringently than with preregistration. In a sense, differential privacy turns the best practices for analyzing a private observational dataset (which might otherwise be inaccessible) into something closer to the best practices for designing a single, expensive field experiment.

In order to ensure researchers can follow the replication standard (King 1995), and to preserve their privacy budget, we recommend that differentially private systems cache results so that rerunning the same analysis adds the identical noise every time and reproduces the identical estimate and standard errors. (Researchers could of course choose to rerun the same analysis with a fresh draw of noise to reduce standard errors at the cost of spending more from their privacy budget.) In addition, authors of politically sensitive studies now often omit information from replication datasets entirely which, instead, could be made available in differentially private ways.

Finally, learning about population quantities of interest is not an individual-level privacy violation (e.g., https://bit.ly/Noviol). A researcher can even learn about an individual from a differentially private mechanism, but no more than if that individual were excluded from the dataset. For example, suppose research indicates that women are more likely to share fake news with friends on social media than men; then, if you are a woman, everyone knows you have a higher risk of sharing fake news. But the researcher would have learned this social science generalization whether or not you were included in the dataset and you have no privacy-related reason to withhold your information.

A key point, however, is that we ensure a chosen differentially private mechanism is inferentially valid or else we will draw the wrong conclusions about social science generalizations. In particular, if researchers use

[7] For example, adding mean-zero noise to one variable or its mean induces no bias for the population mean, but adding noise to its variance induces bias. The estimated slope coefficient in a regression of y on x, with random measurement error in x, is biased toward zero; if we add variables with or without measurement error to this regression, the same coefficient can be biased by any amount and in any direction. Censoring sometimes attenuates causal effects, but it can also exaggerate them; predictions and estimates of other quantities can be too high, too low, or have their signs changed.

[8] Alternatively, if the K quantities are disclosed simultaneously and returned in a batch, then we could choose to set the variance of the error for all together at a higher individual level but lower collective level (Bun and Steinke 2016; Mironov 2017).

privacy-preserving mechanisms without bias corrections, social scientists and ultimately society can be misled. (In fact, it is older people, not women, who are more likely to share fake news! See Guess, Nagler, and Tucker [2019].) Fortunately, all of differential privacy's properties are preserved under post-processing, so no privacy loss will occur when, below, we correct for inferential biases (or if results are published or mixed with any other data sources). In particular, for any data analytic function f not involving private data D, if $M(s, D)$ is differentially private, then $f[M(s, D)]$ is differentially private, regardless of assumptions about potential adversaries or threat models.

Although differential privacy may seem to follow a "do no *more* harm" principle, careless use of it can harm individuals and society if we do not also provide inferential validity. The biases from ignoring measurement error and selection can each separately or together reverse, attenuate, exaggerate, or nullify statistical results. Helpful public policies could be discarded. Harmful practices may be promoted. Of course, when providing access to confidential data, not using differential privacy may also have grave costs to individuals. Data providers must therefore ensure that data access systems are *both* differentially private and inferentially valid.

A DIFFERENTIALLY PRIVATE GENERIC ESTIMATOR

We now introduce an approximately unbiased approach which, like our software, we call UnbiasedPrivacy (or UP); it has two parts: (1) a differentially private mechanism introduced in this section and (2) a bias correction of the differentially private result, using the algorithm in the next section, along with accurate uncertainty estimates in the form of standard errors. Our work builds on the algorithm in Smith (2011), but results in an estimator with a root-mean-square error that, in simulations, is hundreds of times smaller (see Appendix C of the Supplementary Material).

Let $\mathbf{D}$ denote a *population* data matrix, from which N observations are selected to form our observed data matrix D. Our goal is to estimate some (fixed scalar) quantity of interest, $\theta = s(\mathbf{D})$ with the researcher's choice of statistical procedure s. The statistical procedure $s(\cdot)$ refers to the results of running a chosen statistical methodology from among the many statistical methods in widespread use in political science—based on maximum likelihood, Bayesian, nonparametric approaches, or others—and finally computing a quantity of interest computed from them—such as a prediction, probability, first difference, and forecast. Multiple quantities of interest can be estimated by repeating the algorithm (dividing ϵ among the runs).

Our procedure is "generic," by which we mean that researchers may choose among a wide variety of methods, including most now in use across the social sciences, but of course there are constraints. For example, the data must be arranged so that each research subject whose privacy we wish to protect contributes information to only one row of D (see Footnote 4) and, technically, we must select from among the many methods that are statistically valid under bootstrapping.[9]

Let $\hat{\theta} = s(D)$ denote an estimate of θ using statistical estimator $s(\cdot)$, which we would use if we could see the private data; also denote a differentially private estimate of θ by $\hat{\theta}^{\text{dp}}$. To compute $\hat{\theta}^{\text{dp}}$, the researcher chooses a statistical method, a quantity of interest estimated from the statistical method (causal effect, risk difference, predicted value, etc.), and values for each of the privacy parameters (Λ, ϵ, and δ; we discuss how to make these choices in practice below).

Below, we give the details of our proposed mechanism $M(s, D) = \hat{\theta}^{\text{dp}}$ followed by the privacy and then inferential properties of this strategy.

Mechanism

We give here a generic differentially private estimator based on a simple version of the Gaussian mechanism (described above) applied to estimates from subsets of the data rather than to individual observations. To be more specific, the algorithm uses a partitioning version of the "sample and aggregate" algorithm (Nissim, Raskhodnikova, and Smith 2007), to ensure the differential privacy standard will apply for almost any statistical method and quantity of interest. We also incorporate an optional application of the computationally efficient "bag of little bootstraps" algorithm (Kleiner et al. 2014) to ensure an aspect of inferential validity generically, by not having to worry about differences in how to scale up different statistics from each partition to the entire dataset.

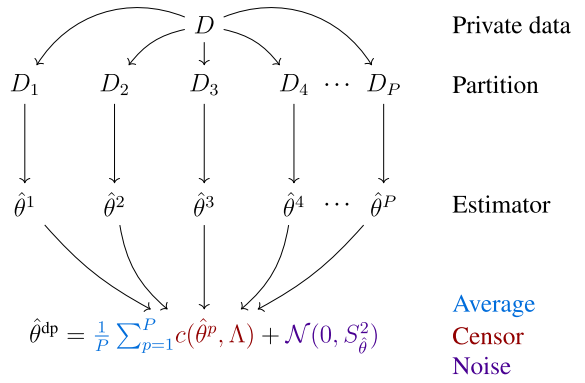
Figure 1 gives a visual representation of the algorithm we now detail. We first randomly sort observations from D into P separate partitions $\{D_1, \ldots, D_P\}$, each of size $n \approx N/P$ (we discuss the choice of P below), and then follow this algorithm.

1. From the data in partition p ($p = 1, \ldots, P$):
   (a) Compute an estimate $\hat{\theta}^p$ (using the same method as we would apply to the full private data, and scaling up to the full dataset, *or* via the general purpose bag of little bootstrap algorithm; see Appendix D of the Supplementary Material).
   (b) Censor the estimate $\hat{\theta}^p$ as $c(\hat{\theta}^p, \Lambda)$.
2. Compute $\hat{\theta}^{\text{dp}}$ by averaging the censored estimates (over partitions instead of observations) and adding mean-zero noise:

$$\hat{\theta}^{\text{dp}} = \hat{\theta} + e, \tag{4}$$

[9] That is, we allow any statistic with a positive bounded second Gateaux derivative and Hadamard differentiability (Wasserman 2006, 35), excluding statistics such as the maximum. We also assume that the parameter value does not fall at a boundary of its support, the distribution of the underlying estimator has a finite mean and variance, and N is growing faster than P. This condition is commonly met, but there are exceptions. For instance, the synthetic control estimator can be nonnormal (Li 2020). It is therefore important to consider whether asymptotic normality applies in any particular application.

FIGURE 1. Differentially Private Algorithm (Before Bias Correction)



where

$$\hat{\theta} = \frac{1}{P}\sum_{p=1}^{P} c(\hat{\theta}^p, \Lambda), \qquad e \sim \mathcal{N}(0, S^2_{\hat{\theta}}), \qquad S_{\hat{\theta}} = S(\Lambda, \epsilon, \delta, P), \tag{5}$$

where S is explained intuitively here and formally in Appendix A of the Supplementary Material.

Privacy Properties

Privacy is ensured in this algorithm by each individual appearing in at most one partition, and by the censoring and noise in the aggregation mechanism ensuring that data from any one individual can have no meaningful effect on the distribution of possible outputs. Each partition can even be sequestered on a separate server, which may reduce security risks.

The advantage of always using the censored *mean* of estimates across partitions, regardless of the statistical method used within each partition, is that the variance of the noise can be calibrated generically to the sensitivity of this mean rather than having to derive the sensitivity anew for each estimator. The cost of this strategy is additional noise because P rather than N appears in the denominator of the variance. Thus, from the perspective of reducing noise, we should set P as large as possible, subject to the constraints that (1) the number of units in each partition $n \approx N/P$ gives valid statistical results in each bootstrap and the estimate being sensible (such as regression covariates being of full rank) and (2) n is large enough and growing faster than P (to ensure the applicability of the central limit theorem or, for a precise rate of convergence, the Berry–Esseen theorem). See also our simulations below. Subsampling itself increases the variance of nonlinear estimators, but usually much less than the increased variance due to privacy-protective procedures; see Mohan et al. (2012) for methods of optimizing P.

Inferential Properties

To study the statistical properties of our point estimator and its uncertainty estimates, consider two conditions: (1) an assumption we maintain until the next section that Λ is large enough so that censoring has no effect ($c(\hat{\theta}^p, \Lambda) = \hat{\theta}^p$) and (2) a method that is unbiased when applied to the private data.

If these two conditions hold, our point estimates are unbiased:

$$E(\hat{\theta}^{\text{dp}}) = \frac{1}{P}\sum_{p=1}^{P} E(\hat{\theta}^p) + E(e) = \theta. \tag{6}$$

In practice, however, choosing the bounding parameter Λ involves a bias-variance trade-off: if Λ is set large enough, censoring has no effect and $\hat{\theta}^{\text{dp}}$ is unbiased, but the noise and resulting uncertainty estimates will be large (see Equation 5). Alternatively, choosing a smaller value of Λ will reduce noise and the resulting uncertainty estimates, but it increases censoring and bias. Of course, if the chosen estimator is biased when applied to the private data, perhaps due to violating statistical assumptions or merely being a nonlinear function of the data like logit or an event count model, then our algorithm will not magically remove the bias, but it will not add bias.

Uncertainty estimators, in contrast, require adjustment even if both conditions are met. For example, the variance of the differentially private estimator is $V(\hat{\theta}^{\text{dp}}) = V(\hat{\theta}) + S^2_{\hat{\theta}}$, but its naive variance estimator (the differentially private version of an unbiased non-private variance estimator) is biased:

$$\begin{aligned} E\left[\hat{V}(\hat{\theta}^{\text{dp}})\right] &= E\left[\hat{V}(\hat{\theta}) + e\right] = V(\hat{\theta}) + E(e) \\ &= V(\hat{\theta}) \neq V(\hat{\theta}^{\text{dp}}). \end{aligned} \tag{7}$$

Fortunately, we can compute an unbiased estimate of the variance of the differentially private estimator by simply adding back in the (known) variance of the noise: $\hat{V}(\hat{\theta}^{\text{dp}}) = \hat{V}(\hat{\theta}) + S^2_{\hat{\theta}}$, which is unbiased: $E\left[\hat{V}(\hat{\theta}^{\text{dp}})\right] = V(\hat{\theta}^{\text{dp}})$. Of course, because censoring will bias our estimates, we must bias correct and then compute the variance of the corrected estimate, which will ordinarily require a more complicated expression.

ENSURING VALID STATISTICAL INFERENCE

We now correct our differentially private estimator for the bias due to censoring, which has the effect of reducing the impact of the choice of Λ. The correction thus allows users to choose smaller values of Λ, and consequentially reduce the variance and use less of the privacy budget (see also our section on practical suggestions below). Because we introduce the correction by post-processing, we retain the same privacy-preserving

properties. It even turns out that the variance of this bias-corrected estimate is actually smaller than the uncorrected estimate, which is unusual for bias corrections. We know of no prior attempt to correct for biases due to censoring in any differentially private mechanism.

We now describe our bias-corrected point estimator, $\tilde{\theta}^{\text{dp}}$, and variance estimate, $\hat{V}(\tilde{\theta}^{\text{dp}})$.

Bias Correction

Our goal here is to correct the bias due to censoring in our estimate of θ. Figure 2 helps visualize the underlying distributions and notation we will introduce, with the original uncensored distribution in blue and the censored distribution in orange, which is made up of an unnormalized truncated distribution and the spikes (which replace the area in the tails) at $-\Lambda$ and Λ. Although Λ and $-\Lambda$ are of course symmetric around zero, the uncensored distribution is centered at θ.

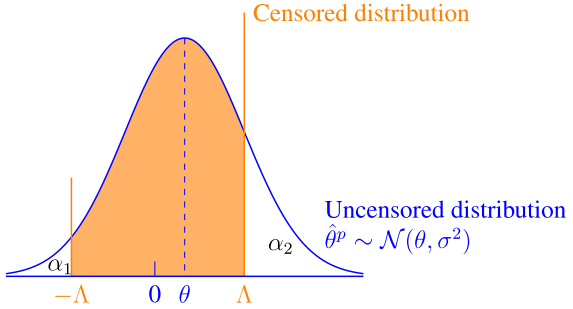
Our strategy is to approximate the distribution of $\hat{\theta}^p$ by a normal distribution, which is correct asymptotically for most commonly used statistical methods in political science. Thus, the distribution of $\hat{\theta}^p$ across partitions, before censoring at $[-\Lambda, \Lambda]$, is approximately $\mathcal{N}(\theta, \sigma^2/n)$, where $n = N/P$.

The proportion left and right censored, respectively (the area under distribution's tails) is therefore approximated by

$$\alpha_1 = \int_{-\infty}^{-\Lambda} \mathcal{N}(t \mid \theta, \sigma^2/n)dt, \qquad \alpha_2 = \int_{\Lambda}^{\infty} \mathcal{N}(t \mid \theta, \sigma^2/n)dt. \tag{8}$$

Under technical regularity conditions (see Appendix E of the Supplementary Material for a proof), we can bound the error on the terms such that $\Pr(\hat{\theta}_n^p < -\Lambda) = \alpha_1 \pm O(1/\sqrt{n})$ and $\Pr(\hat{\theta}_n^p > \Lambda) = \alpha_2 \pm O(1/\sqrt{n})$. This means that the difference between the true proportion censored and our approximation of it, $\alpha_1 + \alpha_2$, is decreasing proportional to $1/\sqrt{n}$, which is typically very small. We thus ignore this (generally negligible) error when constructing our estimator.

FIGURE 2. Underlying Distributions (Before Estimation). The censored distribution includes the orange area and spikes at $-\Lambda$ and Λ



We then write the expected value of $\hat{\theta}^{\text{dp}}$ as the weighted average of the mean of the truncated normal, the spikes at $-\Lambda$ and Λ, and the term we ignore:

$$E\left(\hat{\theta}_n^{\text{dp}}\right) = -\alpha_1\Lambda + (1-\alpha_2-\alpha_1)\theta_T + \alpha_2\Lambda \pm O(1/\sqrt{n}), \tag{9}$$

with truncated normal mean

$$\theta_T = \theta + \sigma/\sqrt{n} \cdot \left(\frac{\mathcal{N}(-\Lambda \mid \theta, \sigma^2/n) - \mathcal{N}(\Lambda \mid \theta, \sigma^2/n)}{1-\alpha_2-\alpha_1}\right).$$

We then construct a plug-in estimator by substituting our point estimate $\hat{\theta}^{\text{dp}}$ for the expected value at the left of Equation 9. We are left with three equations (Equation 9 and the two in Equation 8) and four unknowns (θ, σ, α_1, and α_2). We therefore use some of the privacy budget to obtain an estimate of α_2 (or α_1) as $\hat{\alpha}_2 = \frac{1}{P}\sum_p 1(\hat{\theta}^p > \Lambda)$, which we release via the Gaussian mechanism.[10] How to use the privacy budget is up to the user, but we find that splitting the expenditure equally between the original quantity of interest and this parameter works well.

With the same three equations, and our remaining three unknowns (θ, σ, and α_1), our open-source software gives a fast numerical solution. The result is $\tilde{\theta}^{\text{dp}}$, the approximately unbiased estimate of θ (and estimates of $\hat{\sigma}_{\text{dp}}$ and $\hat{\alpha}_1^{\text{dp}}$). See Appendix F of the Supplementary Material.

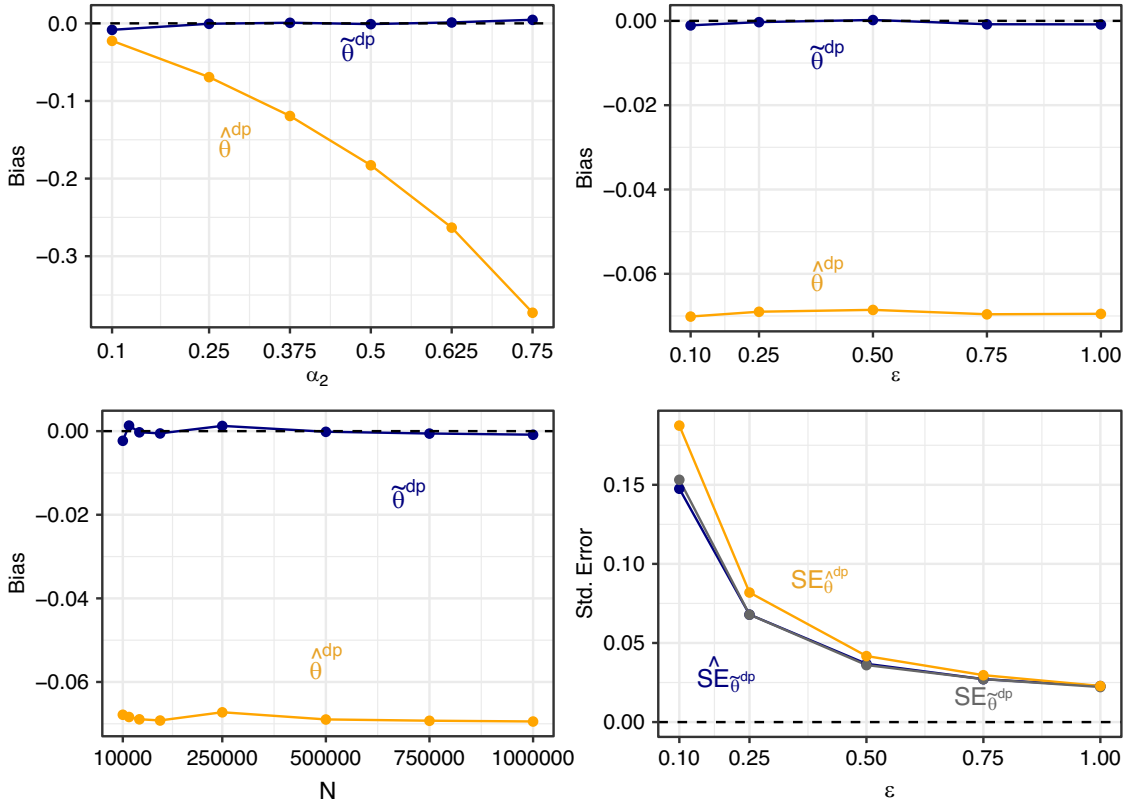
We use the term "approximate unbiasedness" to mean unbiasedness with respect to the private estimator (which itself may not be unbiased), rather than the true parameter. This unbiasedness is approximate for two reasons. First, most plug-in estimators, including ours, are guaranteed to be strictly unbiased only asymptotically. Second, as discussed, our estimating equations have a small error from approximating the partition distribution as normal. And finally, our simulations below show that in finite samples the bias introduced from these two sources is negligible in practice, and considerably smaller than the bias introduced by censoring.

Variance Estimation

We now derive a procedure for computing an estimate of the variance of our estimator, $\hat{V}(\tilde{\theta}^{\text{dp}})$, without any additional privacy budget expenditure. We have the two directly estimated quantities, $\hat{\theta}^{\text{dp}}$ and $\hat{\alpha}_2^{\text{dp}}$, and the three (deterministically post-processed) functions of these computed during bias correction: $\tilde{\theta}^{\text{dp}}$, $\hat{\sigma}_{\text{dp}}^2$, and $\hat{\alpha}_1^{\text{dp}}$. We then use standard simulation methods (King, Tomz, and Wittenberg 2000): we treat the estimated quantities as random variables, bias correct to generate the others,

[10] Since a proportion is bounded between 0 and 1, the sensitivity of this estimator is bounded too and so we can avoid censoring.

FIGURE 3. Monte Carlo Simulations: Bias of the Uncorrected ($\hat{\theta}^{dp}$) and Corrected ($\tilde{\theta}^{dp}$) Estimates, and (in the Bottom-Right Panel) the Standard Error of the True Uncorrected ($SE_{\hat{\theta}^{dp}}$), True Corrected ($SE_{\tilde{\theta}^{dp}}$), and Estimated Corrected ($\widehat{SE}_{\tilde{\theta}^{dp}}$) Estimates (the Latter Two Having Almost Identical Values)



Note: "True" SEs refer to the actual standard deviation of a point estimate.

and take the sample variance of the simulations of $\tilde{\theta}^{dp}$. Thus, to represent estimation uncertainty, we draw the random quantities from a multivariate normal with plug-in parameter estimates, none of which need to be newly disclosed and so the procedure does not use more of the privacy budget. Appendix G of the Supplementary Material provides all the derivations.

SIMULATIONS

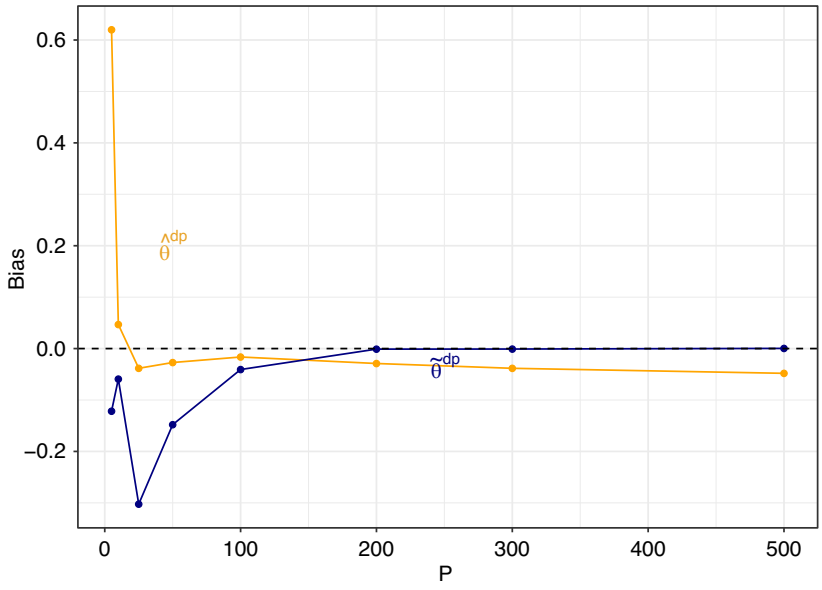
We now evaluate the finite-sample properties of our estimator. We show that while (uncorrected) differentially private point estimates are inferentially invalid, our estimators are approximately unbiased and come with accurate uncertainty estimates. In addition, in part because our bias correction uses an additional disclosed parameter estimate ($\hat{\alpha}_2$), the variance of our estimator is usually lower than the variance of the uncorrected estimator. (Simulations under alternative assumptions appear in Appendix H of the Supplementary Material; replication information is available in Evans et al. [2023].)

The results appear in Figure 3, which we discuss after first detailing the data generation process. For four different types of simulations (in separate panels of the figure), we draw data for each row i from an independent linear regression model: $y_i \sim \mathcal{N}(1 + 3x_i, 10^2)$, with $x_i \sim \mathcal{N}(0, 7^2)$ drawn once and fixed across simulations. Our chosen quantity of interest is the coefficient on x_i with value $\theta = 3$. We study the bias of the (uncorrected) differentially private estimator $\hat{\theta}^{dp}$, and our corrected version, $\tilde{\theta}^{dp}$, as well as their standard errors.[11]

Begin with the top-left panel, which plots bias on the vertical axis (with zero bias indicated by a dashed

[11] We have tried different parameter values, functional forms, distributions, statistical models, and quantities of interest, all of which led to similar substantive conclusions. In Figure 3, for censoring (top-left panel), we let $\alpha_1 \approx 0$, $\alpha_2 = \{0.1, 0.25, 0.375, 0.5, 0.625, 0.75\}$, $N = 100,000$, $P = 1,000$, and $\epsilon = 1$. For privacy (in the top-right panel) and standard errors (bottom-right panel), let $\epsilon = \{0.1, 0.15, 0.20, 0.30, 0.50, 1\}$ while setting $N = 100,000$, $P = 1,000$, and with Λ set so that $\alpha_1 \approx 0$ and $\alpha_2 = 0.25$. For sample size (bottom left), set $N = \{10,000, 25,000, 50,000, 100,000, 250,000, 50,000, 100,000\}$, $P = 1,000$, $\epsilon = 1$, and determine Λ so that $\alpha_1 \approx 0$ and $\alpha_2 = 0.25$. We ran 1,000 Monte Carlo simulations except for $N \leq 50,000$ where we ran 4,000, and 2,000 for $\epsilon = 0.25$. The values of ϵ reported correspond to $\delta = 0.01$ given the Gaussian noise variance; we could instead have set $\delta << 1/N$, in which case the corresponding ϵ would be slightly higher.

FIGURE 4. Performance across *P* (Number of Data Partitions) for Fixed Privacy Budget ($\epsilon = 1$) and Sample Size ($N = 100k$)



horizontal line at zero near the top) and the degree of censoring on the horizontal axis increasing from left to right (quantified by α_2). The orange line in this panel vividly shows how statistical bias in the (uncorrected) differentially private estimator $\hat{\theta}^{\text{dp}}$ sharply increases with censoring. In contrast, our (bias-corrected) estimate in blue $\tilde{\theta}^{\text{dp}}$ is approximately unbiased regardless of the level of censoring.

The top-right and bottom-left panels also plot bias on the vertical axis with zero bias indicated by a horizontal dashed line. The bottom-left panel shows the bias in the uncorrected estimate (in orange) for sample sizes from 10,000 to 1 million, and the top-right panel shows the same for different values of ϵ. Our corrected estimate (in blue) is approximately zero in both panels, regardless of the value of N or ϵ.

Finally, the bottom-right panel reveals that the standard error of $\tilde{\theta}^{\text{dp}}$ is approximately correct (i.e., equal to the true standard deviation across estimates, which can be seen because the blue and gray lines are almost on top of one another). It is even smaller for most of the range than the standard error of the uncorrected estimate $\hat{\theta}^{\text{dp}}$. (Appendix I of the Supplementary Material gives an example and intuition for how to avoid analyses where our approach does not work as expected.)

These simulations suggest that $\tilde{\theta}^{\text{dp}}$ is to be preferred to $\hat{\theta}^{\text{dp}}$ with respect to bias and variance in finite samples.

We also show in Figure 4 how our procedure performs across different partition sizes (P) for fixed ϵ, n, and Λ, using same data generation process. With small P, both $\tilde{\theta}^{\text{dp}}$ and $\hat{\theta}^{\text{dp}}$ are biased because our estimate of the censoring level is noisy in this case (note, however, that in simulations where we use the bag-of-little-bootstraps to estimate the proportion of censored partitions, we perform more favorably with low P, since the asymptotics are in n rather than P). However, when $P > 100$, our procedure corrects the bias from censoring. Our procedure even performs well when $P = 500$ and so $n = 20$. Such a small sample size can be less optimal for other estimators and in skewed data. Analysts should consider the trade-off between noise and partition sample size when choosing P.

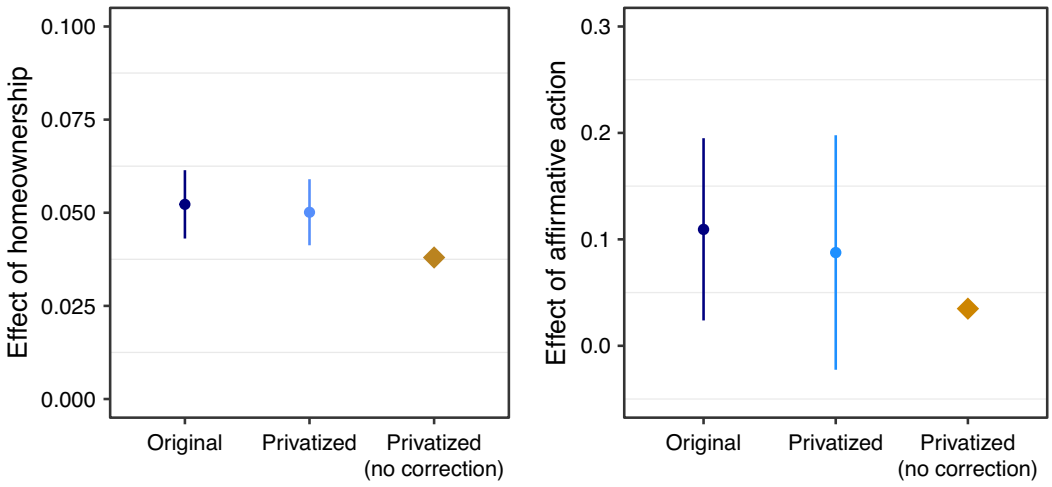
EMPIRICAL EXAMPLES

We now show that the same quantities of interest to political scientists can still be accurately estimated even while guaranteeing the privacy of their respondents. We do this by replicating two important recent articles from major journals. We then treat these datasets as if they were private and not accessible to researchers, except through our algorithm. We then use the algorithm to estimate the same quantities and show that we can recover the same estimates. We also quantify the costs of our approach by the size of the standard errors. See Evans et al. (2023) for replication information.

Home Ownership and Local Political Participation

We begin with Yoder (2020), a study of the effect of home ownership on participation in local politics that uses an unusually informative and diverse array of datasets the author combined via probabilistic matching. Although all the information used in this article is publicly available in separate datasets, combining datasets can be exponentially more informative about each person represented. Some people represented in data like these might well rankle about a researcher being

FIGURE 5. Original versus Privacy-Protected Data Analyses



(a) Yoder (2020) (b) Bhavnani and Lee (2019)

able to easily obtain their name, address, how much of a fuss they made at various city council meetings, all the times during the last 18 years in which they voted or failed to turn out, the dollar value of their home, and which candidates and how much they contributed to each. Moreover, with this profile on any individual, it would be easy to add other variables from other sources.

Yoder (2020) followed current best practices by appropriately de-identifying the data, which meant being forced to strip the replication dataset of many substantively interesting variables, hence limiting the range of discoveries other researchers can make. And yet, we now know that even these procedures do not always protect research subjects, as re-identification remains possible.

In a dataset with $n = 83{,}580$ observations, the author regresses a binary indicator for whether an individual comments at a local city council meeting on an indicator for home ownership, controlling for year and zip code fixed effects, and correcting for uncertainties in the data matching procedure. This causal estimate, which we replicated exactly, indicates that owning a home increases the probability of commenting at a city council meeting by 5 percentage points (0.05) (Yoder 2020, Table 2, Model 1). We focus on this main effect, not the large number of other statistics in the original paper, many of which are of secondary relevance to the core argument. To release all these statistics would require either a large privacy budget (and possibly an unacceptable privacy loss) or excessively large standard errors. Differential privacy hence changes the nature of research, necessitating choices about which statistics are published.

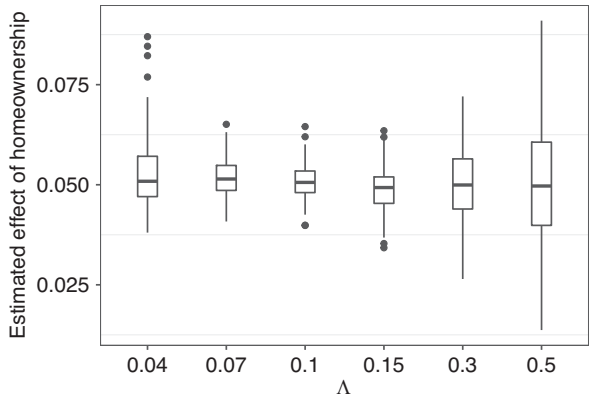
Since there is limited prior research on this question, it is difficult to make an informed choice about the parameter, Λ, which defines where to censor. We therefore dedicate a small portion of the privacy budget to private quantile estimation (Smith 2011). Specifically, we make a private query for an estimate of the 0.6 quantile of the absolute value of the partition-level estimates of our quantity of interest. We dedicate $\epsilon = 0.2$ of our privacy budget to this query, and the returned value was 0.07 (this means approximately 40% of partition-level statistics were greater in magnitude than 0.07). Then, for our main algorithm, we set $\Lambda = 0.075$. Below, we also discuss the sensitivity of our estimator to this choice.

The main causal estimate is portrayed in Figure 5a as a dark blue dot, in the middle of a vertical line representing a 95% confidence interval.

When we estimate the same quantity using our algorithm, the result is a point estimate and confidence interval portrayed in the middle of Figure 5a. As can be seen, the point estimate (in light blue) is almost the same as in the original, and the confidence interval is similar but widened somewhat, leaving essentially the same overall substantive conclusion as in the original about how home ownership increases the likelihood of commenting in a city council meeting. The wider confidence intervals is the cost necessary to guarantee privacy for the research subjects. This inferential cost could be overcome, if desired, by a proportionately larger sample.[12] (For comparison, we also present the biased privatized point estimate without correction on the right side of the same graph.)

[12] For simplicity, we replaced the large number of zip code fixed effects in Yoder (2020) with an equivalent mean differences model. Since we randomly partition the data, we cannot guarantee that the full model with fixed effects will be estimable in every partition, since by chance certain zip codes may not appear. By de-meaning the data instead of including a fixed effect for each zip code, we retain the core logic of comparing outcome variability *within* a zip code, but do so in a way that guarantees the model is estimable in each partition, and significantly reduces the number of parameters to be estimated. The estimated coefficient of interest and its standard error in our simplified model are identical (to three decimal points) to the author's reported estimates. For our algorithm, we use $\Lambda = 0.06, \epsilon_\theta = \epsilon_\alpha = 0.5$, $\delta = 1/N$ where $N = 5{,}978$, and $P = 300$.

FIGURE 6. Distribution of Estimate across 200 Runs of Our Algorithm, at Varying Levels of Λ



In Figure 6, we show how our estimate varies over 200 runs of the algorithm for different values of Λ (which dictates the amount of censoring and noise). We see a trade-off between pre-noise information, which decreases in the level of censoring, and the variance of the noise. In this example, we find that censoring approximately 25% of partitions (corresponding to $\Lambda = 0.1$) induces the lowest estimate variability. However, the results are not significantly or substantively different if we set Λ to somewhat lower (0.07) or higher (0.15) values. But if we set Λ to be very high (0.3 or 0.5), so that censoring is nearly nonexistent, then we have to add large amounts of noise. This demonstrates a key benefit of our estimator that allows us to obtain approximately unbiased estimates in the presence of censoring and with less noise.

Effect of Affirmative Action on Bureaucratic Performance in India

We also replicate Bhavnani and Lee (2021), a study showing that affirmative action hires do not reduce bureaucratic output in India, using "unusually detailed data on the recruitment, background, and careers of India's elite bureaucracy" (5).

The key analysis in this study involves a regression of bureaucratic output on the proportion of bureaucrats who were affirmative action hires. Bureaucratic output was defined as the standardized log of the number of households that received 100 or more days of employment under MGNREGA, India's (and the world's largest) poverty program. The causal estimate, based on $n = 2,047$ observations, is positive and confirms the authors' hypothesis.

We easily replicate these results, which appear in dark blue on the left side of Figure 5b. As above, we now treat the data as private and accessible only through our algorithm and estimate the same quantity. The result appears in light blue in the middle. The overall substantive conclusion is essentially unchanged—indicating the absence of any evidence that affirmative action hires decrease bureaucratic output. The increase in the size of the confidence interval reflects the cost of the privacy protections we chose to apply.[13]

In both applications, our algorithm provides privacy in the form of deniability for any person who is or could be in the data: reidentification is essentially impossible, regardless of how much external information an attacker may have. It also enables scholars to produce approximately the same results and the same substantive conclusion as without privacy protections. The inferential cost of the procedure is the increase in confidence intervals and standard errors, as a function of the user-determined choice of ϵ. This cost can be compensated for by collecting a larger sample. Of course, in many situations, not paying this "cost" may mean no data access at all.

PRACTICAL SUGGESTIONS AND LIMITATIONS

Like any data analytic approach, how the methods proposed here are used in practice can be as important as their formal properties. We discuss here issues of reducing the societal risks of differential privacy, choosing ϵ, choosing Λ, and theory and practice differences. See also Appendix F of the Supplementary Material on suggestions for software design.

Reducing Differential Privacy's Societal Risks

Data access systems with differential privacy are designed to reduce privacy risks to individuals. Correcting the biases due to noise and censoring, and adding proper uncertainty estimates, greatly reduces the remaining risks to researchers and, in turn, to society. There is, however, another risk we must tackle: consider a firm seeking public relations benefits by making data available for academics to create public good but, concerned about bad news for the firm that might come from the research, takes an excessively conservative position on the total privacy budget. In this situation, the firm would effectively be providing a big pile of useless random numbers while claiming public credit for making data available. No public good could be created, no bad news for the firm could come from the research results because all causal estimates would be approximately zero, and still the firm would benefit from great publicity.

To avoid this unacceptable situation, we quantify the statistical cost to researchers of differential privacy or, equivalently, how much information the data provider is actually providing to the scholarly community. To do this, we note that making population-level inferences in a differentially private data access system is equivalent

[13] We used parameters $\Lambda = 0.35$, $\epsilon_\theta = \epsilon_\alpha = 1$, $\delta = 1/N$, and $P = 150$. Notably, this implies that $n \approx 13$ in each of the partitions, which is a relatively small sample size. This makes this a hard test for our procedure, since we rely on an asymptotic approximation. Our results show that the procedure nevertheless performs well in this dataset, but we warn against assuming that the partition-level distributed is well approximated by a normal distribution in all small sample sizes.

to an ordinary data access system with some specific proportion of the data discarded. This means we can provide an intuitive statistic, *the proportion of observations effectively lost due to the privacy-protective procedures.* Appendix J of the Supplementary Material formally defines and shows how to estimate this quantity, which we call L for loss.

Because L can be computed after from the results of our algorithm, without any additional expenditure from the privacy budget, we recommend it be regularly reported publicly by data providers or researchers using differentially private data access systems. For example, in the applications we replicate, the proportion of observations effectively lost is $L = 0.36$ for the study of home ownership and $L = 0.43$ for the study of affirmative action in India. These could have been changed by adjusting the privacy parameters ϵ, δ, and Λ in how the data were disclosed.

Choosing ϵ

From the point of view of the statistical researcher, ϵ directly influences the standard error of the quantity of interest, although with our algorithm this choice will not affect the degree of bias. Because we show above that typically $\hat{V}(\tilde{\theta}^{\text{dp}}) < \hat{V}[c(\hat{\theta}^{\text{dp}}, \Lambda)] < \hat{V}(\hat{\theta}^{\text{dp}})$, we can simplify and provide some intuition by writing an upper bound on the standard error $\text{SE}_{\tilde{\theta}^{\text{dp}}} \equiv \sqrt{\hat{V}(\tilde{\theta}^{\text{dp}})}$ as

$$\text{SE}_{\tilde{\theta}^{\text{dp}}} < \sqrt{V(\hat{\theta}^{\text{dp}}) + S(\Lambda, \epsilon, \delta, P)^2}. \tag{10}$$

A researcher can use this expression to judge how much of their allocation of ϵ to assign to the next run by using their prior information about the likely value of $\hat{V}(\hat{\theta}^{\text{dp}})$ (as they would in a power calculation), plugging in the chosen values of Λ, δ, and P, and then trying different values of ϵ. See also Hsu et al. (2014) and Abowd and Schmutte (2019) for an economic theory approach.

Choosing Λ

Although our bias correction procedure makes the choice of Λ less consequential, researchers with extra knowledge should use it. In particular, reducing Λ increases the chance of censoring while reducing noise, whereas larger values reduce censoring but increase noise. This Heisenberg-like property is an intentional feature, designed to keep researchers from being able to see certain information with too much precision.

We can, however, choose among the unbiased estimators our method produces that have the smallest variance. To do that, researchers should set Λ by trying to capture the point estimate of the mean. Although this cannot be done with certainty, researchers can often do this without seeing the data. For example, consider the absolute value of coefficients from any real application of logistic regression. Although technically unbounded, empirical regularities in how researchers typically scale their input variables lead to logistic regression coefficients reported in the literature rarely having absolute values above about five. Similar patterns are easy to identify across many statistical procedures. A good software interface would thus not only include appropriate defaults, but also enable users to enter asymmetric censoring intervals. Then the software, rather than the user, takes responsibility in the background for rescaling the variables as necessary.

Applied researchers are good at making choices like this as they have considerable experience with scaling variables, a task that is an essential part of most data analyses. Researchers also frequently predict the values of their quantities of interest both informally, when deciding what analysis to run next, and formally for power calculations. If the data surprise us, we will learn this because the $\hat{\alpha}_1^{\text{dp}}$ and $\hat{\alpha}_2^{\text{dp}}$ are disclosed as part of our procedure. If either quantity is more than about 60%, we recommend researchers consider adjusting Λ and rerunning their analysis (see Appendix K of the Supplementary Material) or adapting Smith (2011) procedure to learn the value of Λ where censoring begins (See Appendix C of the Supplementary Material).

Implementation Choices

In the literature, differential privacy theorists tend to be conservative in setting privacy parameters and budgets; practitioners take a more lenient perspective. Both perspectives make sense: theorists analyze worst-case scenarios using mathematical certainty as the standard of proof, and are ever wary of scientific adversaries hunting for loopholes in their mechanisms. This divergence even makes sense both theoretically, because privacy bounds are much higher than expected in practice (Erlingsson et al. 2019), and empirically, because those responsible for implementing data access systems have little choice but to make some compromises in turning math into physical reality. Common implementations thus sometimes allow larger values of ϵ for each run or reset the privacy budget periodically.

We add two practical suggestions. First, although the data sharing regime can be broken by intentional attack, because re-identification from de-identified data is often possible, de-identification is still helpful in practice. It is no surprise that university Institutional Review Boards have rephrased their regulations from "de-identified" to "not readily identifiable" rather than disallowing data sharing entirely. Adding our new privacy-protective procedures to de-identified data provides further protection. Other practical steps can be prudent, such as disallowing repeated runs with new draws of noise of any one analysis.

Second, potential data providers and regulators should ask themselves *Are these researchers trustworthy?* They almost always have been. When this fact provides insufficient reassurance, we can move to the data access regime. However, a middle ground exists by trusting researchers (perhaps along with auxiliary protections, such as data use agreements by university employers, sanctions for violations, and auditing of analyses to verify compliance). With trust, researchers

can be given full access, be allowed to run any analyses, but be required to use the algorithm proposed here to disclose data publicly. The data holder could then maintain a strict privacy budget summed over published analyses, which is far more useful for scientific research than counting every exploratory data analysis run against the budget. This plan approximates the differential privacy ideal more closely than the typical data access regime, as the privacy protections among results published are then protected by mathematical guarantees. There are theoretical risks (Dwork and Ullman 2018), but the advantages to the public good that can come from research with fewer constraints may be substantial.

Limits of Differential Privacy

Finally, differential privacy is a new, rapidly advancing technology. Off-the-shelf methods to optimally balance privacy and utility do not exist for many dataset types. Although we provide a generic method, with an approximately unbiased estimator, methods tuned to specific datasets with better properties may sometimes be feasible. Some applications of differential privacy make compromises by setting high privacy budgets, post-processing in intentionally damaging ways, or periodically resetting the budget (Domingo-Ferrer, Sánchez, and Blanco-Justicia 2021), actions that render privacy or utility guarantees more limited than the math implies. These and other shortcomings of how differentially private has been applied pose a valuable opportunity for researchers to develop tools to reduce the utility-privacy trade-off and develop statistical methods of analyzing these data capable of answering important social science questions. For other more specific limitations of our methodology, see also Appendices F, I, and K of the Supplementary Material.

CONCLUDING REMARKS

The differential privacy literature focuses appropriately on the utility-privacy trade-off. We propose to revise the definition of "utility," so it offers value to researchers who seek to use confidential data to learn about the world, beyond inferences to the inaccessible private data. A scientific statement is not one that is necessarily correct, but one that comes with known statistical properties and an honest assessment of uncertainty. Utility to scholarly researchers involves inferential validity, the ability to make these informative scientific statements about populations. While differential privacy can guarantee privacy to individuals, researchers also need inferential validity to make a data access system safe for drawing proper scientific statements, for society using the results of that research, and for individuals whose privacy must be protected. Inferential validity without differential privacy may mean beautiful theory without data access, but differential privacy without inferential validity may result in biased substantive conclusions that mislead researchers and society at large.

Together, approaches that are differentially private and inferentially valid may begin to convince companies, governments, and others to let researchers access their unprecedented storehouses of informative data about individuals and societies. If this happens, it will generate guarantees of privacy for individuals, scholarly results for researchers, and substantial value for society at large.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055422001411.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available in the APSR Dataverse at https://doi.org/10.7910/DVN/ASUFTY.

ACKNOWLEDGMENTS

Many thanks for helpful comments to Adam Breuer, Merce Crosas, Cynthia Dwork, Max Golperud, Roubin Gong, Andy Guess, Chase Harrison, Kosuke Imai, Dan Kifer, Patrick Lam, Xiao-Li Meng, Solomon Messing, Nate Persily, Aaron Roth, Paul Schroeder, Adam Smith, Salil Vadhan, Sergey Yekhanin, and Xiang Zhou. Thanks also for help from the Alexander and Diviya Maguro Peer Pre-Review Program at Harvard's Institute for Quantitative Social Science.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human subjects.

REFERENCES

Abowd, John M. 2018. "Staring-Down the Database Reconstruction Theorem." In Joint Statistical Meetings, Vancouver, BC. https://bit.ly/census-reid.

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.

Al Aziz, Md Momin, Reza Ghasemi, Md Waliullah, and Noman Mohammed. 2017. "Aftermath of Bustamante Attack on Genomic Beacon Service." *BMC Medical Genomics* 10 (2): 43–54.

Barrientos, Andrés F., Jerome Reiter, Machanavajjhala Ashwin, and Yan Chen. 2019. "Differentially Private Significance Tests for Regression Coefficients." *Journal of Computational and Graphical Statistics* 28 (2): 440–53.

Bhavnani, Rikhil R., and Alexander Lee. 2021. "Does Affirmative Action Worsen Bureaucratic Performance? Evidence from the Indian Administrative Service." *American Journal of Political Science* 65 (1): 5–20.

Blackwell, Matthew, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Overview." *Sociological Methods and Research* 46 (3): 303–41.

Bun, Mark, and Thomas Steinke. 2016. "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds." In *Theory of Cryptography Conference*, eds. Martin Hirt and Adam Smith, 635–58. Berlin: Springer.

Carlini, Nicholas, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks." 28th USENIX Security Symposium (USENIX Security 19).

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, et al. 2020. "Extracting Training Data from Large Language Models." Preprint, arXiv:2012.07805.

Chetty, Raj, and John N. Friedman 2019. "A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples." *Journal of Privacy and Confidentiality* 9 (2): 1–23. https://doi.org/10.29012/jpc.716.

Desfontaines, Damien, and Balázs Pejó. 2019. "SoK: Differential Privacies." Preprint, arXiv:1906.01337.

Ding, Bolin, Janardhan Kulkarni, and Sergey Yekhanin. 2017. "Collecting Telemetry Data Privately." In *Advances in Neural Information Processing Systems*, 3571–80.

Domingo-Ferrer, Josep, David Sánchez, and Alberto Blanco-Justicia. 2021. "The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning)." *Communications of the ACM* 64 (7): 33–5.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Rein-gold, and Aaron Roth. 2015. "The Reusable Holdout: Preserving Validity in Adaptive Data Analysis." *Science* 349 (6248): 636–38.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography Conference*, eds. Shai Halevi and Tal Rabin 265–84. Berlin: Springer.

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

Dwork, Cynthia, and Jonathan Ullman. 2018. "The Fienberg Problem: How to Allow Human Interactive Data Analysis in the Age of Differential Privacy." *Journal of Privacy and Confidentiality* 8 (1): 1–10. https://doi.org/10.29012/jpc.687.

Erlingsson, Úlfar, Ilya Mironov, Ananth Raghunathan, and Shuang Song. 2019. "That Which We Call Private." Preprint, arXiv:1908.03566.

Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova. 2014. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response." In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 1054–67. https://doi.org/10.1145/2660267.2660348.

Evans, Georgina, and Gary King. 2023. "Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset." *Political Analysis*, 1–21. doi:10.1017/pan.2022.1.

Evans, Georgina, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. 2023. "Replication Data for: Statistically Valid Inferences from Privacy-Protected Data." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/ASUFTY.

Evans, Georgina, Gary King, Adam D. Smith, and Abhradeep Thakurta. Forthcoming. "Differentially Private Survey Research." *American Journal of Political Science*.

FPF. 2017. "Understanding Corporate Data Sharing Decisions: Practices, Challenges, and Opportunities for Sharing Corporate Data with Researchers." Technical Report Future of Privacy Forum. https://bit.ly/fpfpriv.

Gaboardi, Marco, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan. 2016. "Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing." *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, PMLR 48: 2111–20.

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. "Issues Encountered Deploying Differential Privacy." In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–37. https://doi.org/10.1145/3267323.3268949.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5 (1): eaau4586.

Henriksen-Bulmer, Jane, and Sheridan Jeary. 2016. "Re-Identification Attacks—A Systematic Literature Review." *International Journal of Information Management* 36 (6): 1184–92.

Hsu, Justin, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C. Pierce, and Aaron Roth. 2014. "Differential Privacy: An Economic Method for Choosing Epsilon." In *2014 IEEE 27th Computer Security Foundations Symposium*, 398–410. Piscataway, NJ: IEEE.

Jayaraman, Bargav, and David Evans. 2019. "Evaluating Differentially Private Machine Learning in Practice." 28th USENIX Security Symposium (USENIX Security 19).

Karwa, Vishesh, and Salil Vadhan. 2017. "Finite Sample Differentially Private Confidence Intervals." Preprint, arXiv:1711.03908.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28 (3): 443–99.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

King, Gary, and Nathaniel Persily. 2020. "A New Model for Industry–Academic Partnerships." *PS: Political Science and Politics* 53 (4): 703–9.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 341–55.

Kleiner, Ariel, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. 2014. "A Scalable Bootstrap for Massive Data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (4): 795–816.

Li, Kathleen T. 2020. "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods." *Journal of the American Statistical Association* 115 (532): 2068–83.

Mironov, Ilya. 2017. "Rényi Differential Privacy." In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–75. Piscataway, NJ: IEEE.

Mohan, Prashanth, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. 2012. "GUPT: Privacy Preserving Data Analysis Made Easy." In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 349–60. https://doi.org/10.1145/2213836.2213876.

Monogan, James E. 2015. "Research Preregistration in Political Science: The Case, Counterarguments, and a Response to Critiques." *PS: Political Science and Politics* 48 (3): 425–9.

Nissim, Kobbi, Sofya Raskhodnikova, and Adam Smith. 2007. "Smooth Sensitivity and Sampling in Private Data Analysis." In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, 75–84. doi.org/10.1145/1250790.1250803.

Reiter, Jerome P. 2012. "Statistical Approaches to Protecting Confidentiality for Microdata and Their Effects on the Quality of Statistical Inferences." *Public Opinion Quarterly* 76 (1): 163–81.

Robbin, Alice. 2001. "The Loss of Personal Privacy and Its Consequences for Social Research." *Journal of Government Information* 28 (5): 493–527.

Roberts, Margaret E. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton, NJ: Princeton University Press.

Sheffet, Or. 2017. "Differentially Private Ordinary Least Squares." In *Proceedings of the 34th International Conference on Machine Learning*, 70: 3105–14.

Smith, Adam. 2011. "Privacy-Preserving Statistical Estimation with Optimal Convergence Rates." In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, 813–22. https://doi.org/10.1145/1993636.1993743.

Stefanski, Len A. 2000. "Measurement Error Models." *Journal of the American Statistical Association* 95 (452): 1353–58.

Sweeney, Latanya. 1997. "Weaving Technology and Policy Together to Maintain Confidentiality." *The Journal of Law, Medicine and Ethics* 25 (2–3): 98–110.
Tang, Jun, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. 2017. "Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12." Preprint, arXiv:1709.02753.
Vadhan, Salil. 2017. "The Complexity of Differential Privacy." In *Tutorials on the Foundations of Cryptography*, ed. Yehuda Lindell, 347–450. Cham, CH: Springer.
Wang, Yue, Daniel Kifer, and Jaewoo Lee. 2018. "Differentially Private Confidence Intervals for Empirical Risk Minimization." Preprint, arXiv:1804.03794.
Wang, Yue, Jaewoo Lee, and Daniel Kifer. 2015. "Differentially Private Hypothesis Testing, Revisited." Preprint, arXiv:1511.03376.
Wasserman, Larry. 2006. *All of Nonparametric Statistics*. New York: Springer Science & Business Media.
Wasserman, Larry. 2012. "Minimaxity, Statistical Thinking and Differential Privacy." *Journal of Privacy and Confidentiality* 4 (1): 51–63.
Williams, Oliver, and Frank McSherry. 2010. "Probabilistic Inference and Differential Privacy." In *Proceedings of the 23rd International Conference on Neural Information Processing Systems* 2: 2451–59.
Wilson, Royce J., Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. 2019. "Differentially Private SQL with Bounded User Contribution." Preprint, arXiv:1909.01917.
Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18: 327–50.
Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, and Salil Vadhan. 2018. "Differential Privacy: A Primer for a Non-Technical Audience." *Vanderbilt Journal of Entertainment and Technology Law* 21 (1): 209–75.
Yoder, Jesse. 2020. "Does Property Ownership Lead to Participation in Local Politics? Evidence from Property Records and Meeting Minutes." *American Political Science Review* 114 (4): 1213–29.