Differentially Private Survey Research[*]

Georgina Evans[†] Gary King[‡] Adam D. Smith[§] Abhradeep Thakurta[¶]

April 3, 2023

Abstract

Survey researchers have long protected the privacy of respondents via de-identification (removing names and other directly identifying information) before sharing data. Although these procedures help, recent research demonstrates that they fail to protect respondents from intentional re-identification attacks, a problem that threatens to undermine vast survey enterprises in academia, government, and industry. This is especially a problem in political science because political beliefs are not merely the subject of our scholarship; they represent some of the most important information respondents want to keep private. We confirm the problem in practice by re-identifying individuals from a survey about a controversial referendum declaring life beginning at conception. We build on the concept of "differential privacy" to offer new data sharing procedures with mathematical guarantees for protecting respondent privacy and statistical validity guarantees for social scientists analyzing differentially private data. The cost of these new procedures is larger standard errors, which can be overcome with somewhat larger sample sizes.

The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science Dataverse* within the Harvard Dataverse Network, at: XXX.

Words: 9970

[*] Our thanks to Kosuke Imai for helpful comments. The current version of this paper and a supplementary appendix is available at GaryKing.org/DPsurvey.

[†] Ph.D. Candidate, Department of Government, Harvard University, 1737 Cambridge Street Cambridge, MA 02138; Georgina-Evans.com, GeorginaEvans@g.harvard.edu.

[‡] Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; GaryKing.org, King@Harvard.edu.

[§] Professor, Computer Science and Engineering, Boston University, ads22@bu.edu. Supported by Cooperative Agreement CB16ADR0160001 with the -Census Bureau, NSF award CCF-1763786, and a Sloan Foundation research award. The views expressed in this paper are those of the -authors and not those of the U.S. Census Bureau or any other sponsor.

[¶] Assistant Professor, Department of Computer Science, University of California Santa Cruz, bit.ly/AbhradeepThakurta, aguhatha@ucsc.edu.

1 Introduction

Survey research constitutes about half of the quantitative evidence base of political science (King, Honaker, et al. 2001; Sturgis and Luff 2020) and an enormous enterprise with data collectors, providers, and analysts spanning numerous areas of academia, government, and private industry. In all these fields, survey researchers have gone to great lengths to protect respondent privacy (Plutzer, 2019; Connors, Krupnikov, and Ryan, 2019), usually by stripping out personally identifiable information, such as name, address, and phone number. Unfortunately, a growing literature now demonstrates that these "de-identification" procedures (and other procedures, such as restricted views, clean rooms, query auditing, etc.) do not protect respondents from intentional re-identification attacks (Dwork and Roth, 2014; Henriksen-Bulmer and Jeary, 2016; Wood et al., 2018). For a vivid example, Sweeney (1997) discovered that 87% of the US population can be individually identified with merely zip code, gender, and date of birth. A quarter century later, modern surveys collect considerably more information and faster computers make re-identification much easier. In fact, the US Census Bureau was able to re-identify the personal answers of 52 million Americans from supposedly anonymous and publicly available 2010 census data (Abowd, 2018). This situation "scares the daylights out of those responsible for curating 'public use' versions of confidential data" (Christensen, Freese, and Miguel, 2019, p.181ff).

Political scientists have a special responsibility to ensure the privacy of our research subjects because political beliefs constitute at once (a) some of the most information respondents seek to keep private, (b) a large fraction of our surveys, and (c) many of our most important scholarly questions. In democracies, privacy legislation, which political scientists also study and which governs our license to operate in some subfields, is based in part on elected representatives' views of these political beliefs. In autocracies, privacy is essential for ensuring the safety of our respondents and their willingness to provide sincere answers. Unfortunately, political science is not immune to this problem, which we demonstrate below by performing our own re-identification attack (on data about attitudes toward a controversial referendum seeking to define life as beginning at conception).

The working solution to this problem, in the law and in social science scholarship, has been *balancing* individual privacy with the benefits to the public good that can come from scholarly access to survey data. In this paper, we introduce new methods to reduce the need for balancing by simultaneously offering mathematical guarantees for the privacy of survey respondents and statistical validity guarantees for researchers analyzing privacy protected data to learn about societal patterns. This is possible by adapting to survey research the fast growing literature on "differential privacy" (Dwork, McSherry, et al., 2006), which may also have the advantage of satisfying regulators (King and Persily, 2020). By adding specially calibrated random "noise" to the data before sharing, differential privacy gives respondents deniability for what they may have said to a pollster and even for whether they took the survey at all. Although differential privacy requires adaptation to new data types, as we do here for surveys, the technique has seen increasing adoptions in other contexts, including Social Science One and Facebook (Messing et al., 2020), US Census Bureau (Abowd et al., 2020), Google (Erlingsson, Pihur, and Korolova, 2014; Wilson et al., 2019), Apple (Tang et al., 2017), and Microsoft (Ding, Kulkarni, and Yekhanin, 2017).

Although theorists have found ways of minimizing the amount of noise necessary to protect the privacy of every person who could be in the dataset, the resulting "noisy" dataset has the equivalent of measurement error, which can bias many statistical inferences in any direction and by any amount (Evans and King, forthcoming; Blackwell, Honaker, and King, 2017; Buonaccorsi, 2010). Fortunately, a principle of differential privacy is that the process generating the noise is always made public. We thus use this public information to design statistical procedures to draw valid inferences from differentially private data. We show that, with the methods proposed herein, the main cost incurred for protecting privacy is larger standard errors or confidence intervals, a cost that can be overcome by increasing the sample size. Of course, for some sensitive surveys, the alternative to paying this "cost" may be no survey data at all. The new procedures also have a surprising benefit in that they also provably reduce the risks of p-hacking and overfitting (Dwork, Feldman, et al., 2015).

We provide a framework to understand where privacy protective noise can be injected into the survey research process in Section 2, and the types of noise for each in Section 3. Although our framework has not appeared before, we use existing differentially private mechanisms from the computer science literature wherever possible (presented in self-contained ways when feasible, because the works we cite are written for a different audience and so would be largely unrecognizable to most social scientists). In Section 4, we then derive our own novel statistical methods to avoid the biases (and uncertainty estimates) induced by the resulting measurement error, and ignored in prior literature. We evaluate our methods with empirical data and simulations in Section 5, and give practical survey design advice in Section 6. Section 7 concludes, with technical details, additional practical advice, and extensions in the Supplementary Appendix.

2 Adding Privacy Protective Noise to Survey Data

For datasets researchers have possession of, privatization is sometimes necessary to convince IRBs to allow us to analyze or share data. The methods described here are even more valuable when researchers negotiate to obtain data from others, especially as data providers, regulators, private companies, governments, nonprofits, and survey respondents continue to increase their expectations for security and privacy.

We now show how to protect survey respondent privacy by adding differentially private noise to the data at one of five different points in the usual data collection process. Each of these points comes with its own requirements, assumptions, and noise distribution. Although we focus primarily on the second and third steps in subsequent sections, we describe all five here to put these two in context and to provide advice for those wishing to use other methods for one of the other steps. We save for Section 3 a description of how the noise for each is calibrated and added.

We begin with Figure 1, which illustrates the typical survey research process. Reading from left to right, we have the text of the survey, administered to a respondent, who then enters his or her answers into a device (e.g., a computer, app, cell phone, or perhaps a tablet provided by an interviewer). The information from all the devices and all the respondents are then continuously sent to the server. The data is then amassed and given to researchers who write up and publish their results for the public or their clients.

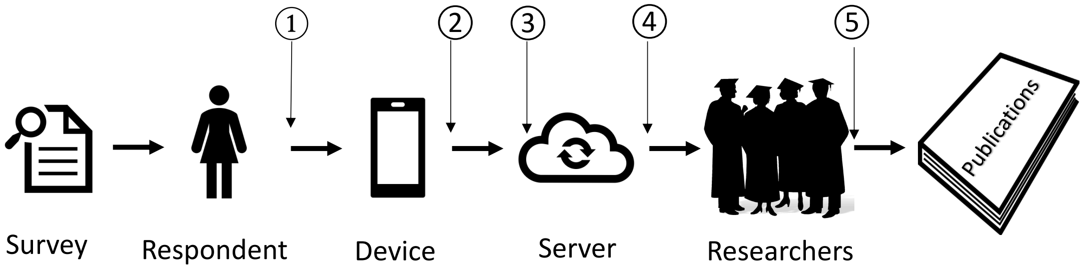

Figure 1: Points In the Survey Process where Privacy Protective Noise Can be Added: ① Randomized response, ② Device output, ③ Server ingest, ④ Server output, and ⑤ Pre-publication. A process described by each point ensures privacy via cybersecurity to its left and differential privacy to its right.

If no privacy protective noise is added, an "attacker" can potentially violate respondents' privacy at any stage, including even from aggregated results that are published. Figure 1 also identifies points at which noise can be added (indicated by numbers in circles at the top of the figure). For each point, ensuring respondent privacy requires trusting one's cybersecurity procedures for every step to its left and differential privacy guarantees for each step to its right.

Consider, for example, Point ⑤, which we call *pre-publication*. Here, researchers are trusted and can run any statistical procedure on the private data they wish, but noise is added, or other privacy protective changes are made, to their statistical results prior to publication, providing a replication data file with the publication, or other public release. This approach requires trusting researchers not only to avoid publishing private information, but also to avoid inadvertently leaking information through data-dependent choices of which analyses to publish (Dwork and Ullman, 2018), such as via preregistration. For this point, we must assume that cybersecurity is good enough to prevent an attacker from obtaining information from the respondent, the device, the server, or the researchers. In contrast, differentially private results that appear in publications come with mathematical guarantees of deniability for any survey respondent, regardless of how hard an attacker may try or what external information they may have.

If instead we add noise at Point ④, *server output*, the researchers need not be trustworthy since they cannot learn anything about any one individual. With noise injected at this point, we are still reliant on cybersecurity to prevent an attacker from breaking into the server where data are stored, the device, or the respondent, but we can guarantee that neither the researchers nor the public reading publications will be able to identify any respondent.

We could alternatively use Point ③, *server ingest*, where noise is added in the server immediately upon receiving data from the device. Because no private data is stored on the server, a one-time break-in to the server cannot violate anyone's privacy. (This strong guarantee is known in the computer science literature as "one-intrusion panprivacy" Dwork, Naor, et al. 2010).

If we are concerned about the security of the server even for the simple task of aggregating and adding noise upon receipt, we can move to Point ② to add noise, which only requires ensuring that the respondent's device is secure, since all data leaving it has privacy protective noise and comes with the mathematical guarantees of deniability.

Finally, we add ①, the well known case of *randomized response*, where noise is added before the respondent chooses and reports an answer, using a physical randomization device under the respondent's control (such as a spinner or a pair of coins, as we describe in Section 3.1). Only after this randomization of the survey question does the respondent enter information into their cell phone or device. This procedure, which can use the same randomization mechanism as ②, protects the respondent so long as an attacker is unable to break cybersecurity by spying on the (randomized) survey question or somehow reading the respondent's mind. Any other use of the data — including by the device, server, researcher analyses, or publications and combination with any external information not in the data — is guaranteed to protect respondent privacy.

3 Differential Privacy

Points ④ and ⑤ in Figure 1 enter after the dataset is amassed and so can be approached using a range of specific procedures fine tuned to each statistical analysis method. See Dwork and Roth (2014) and Vadhan (2017) for overviews and Evans, King, et al. (2020) for a more generic approach. Point ① is the well known randomized response, which we describe in Section 3.1 for its use as designed and because we will use it as a building block for defining differential privacy. Our focus then is on methods for Points ② (device output) and ③ (server ingest) in Sections 3.3 and 3.4 respectively, and their combination in Section 3.5. We will show that mixing these levels will be especially useful in a wide range of applications.

3.1 A Special Case: Randomized Response

Consider the goal of estimating the proportion μ of a population that has engaged in some highly sensitive activity, such as protesting against an authoritarian government, participating in oral sex, or committing a serious crime. For a simple random sample of n observations from this population, each individual i ($i = 1, \ldots, n$) either did, which we denote $y_i^* = 1$, or did not, $y_i^* = 0$, engage in this activity, but may not be willing to reveal this information honestly to a researcher. To be more precise, let y_i denote a binary response to a direct question about participation in the same activity. Then, $\bar{y} \equiv \sum_{i=1}^{n} y_i/n$ is likely a biased estimate (and probably underestimate) of $\mu \equiv \sum_{i=1}^{n} y_i^*/n$.

Randomized response is a way of obtaining a plausibly unbiased estimate of this population parameter, while giving the respondent deniability about their actual answer (Warner, 1965; Blair, Imai, and Zhou, 2015). For simplicity, suppose the survey contains only one sensitive question, although the technique can be extended to any number. Thus, the interviewer (or device) presents each respondent with a spinner that has p ($0 \leq p < 0.5$) proportion of an area where the arrow can stop labeled "I did this" (i.e., claiming that $y_i^* = 1$) and the rest labeled "I did not do this" (i.e., $y_i^* = 0$). The value p (and the fact that the spinner follows a uniform distribution) is known publicly, but each respondent spins privately and does not disclose where the spinner stops. The respondent is then asked only whether the message where the spinner stopped is correct, which we denote $y_i^{(p)}$, with values 1 for "yes" and 0 for "no". More formally, let $z_i \sim \text{Bernoulli}(p)$ be a Bernoulli random draw (observed only to the respondent when they spin the spinner) with known proportion p, and let $y_i^{(p)} = (1 - z_i)y_i + z_i(1 - y_i)$, an expression which flips the value of y_i from 0 to 1 or 1 to 0 with probability $1 - p$. The special case of $p = 0$, indicating that the spinner always returns the same value, is equivalent to the direct survey question: $y_i^{(0)} \equiv y_i$.

Randomized response can be viewed from three perspectives. First, the *privacy* of the respondent is protected because they have deniability: no one can determine whether their observed response is their revelation of their participation in the sensitive activity or the action of the spinner changing the answer. The farther p is from 0, the more privacy is ensured (and as $p \to 0.5$, no information is conveyed at all). Second is the *social psychological assumption* which is that, because of the privacy protections, the respondent is more likely to give an honest answer when $p > 0$ (for which some evidence exists; see Rosenfeld, Imai, and Shapiro 2016).

And finally, conditional on the social psychological assumption, we have a *statistical theory*: Although the mean of the observed responses $\bar{y}$ is biased, we can construct an unbiased estimate by writing $\bar{y} = p\mu + (1 - p)(1 - \mu)$ and solving for μ:

$$\hat{\mu} = \frac{\bar{y} - (1 - p)}{2p - 1}. \tag{1}$$

That is, $E(\hat{\mu}) = \mu$, with variance

$$V(\hat{\mu}) = \frac{\mu(1 - \mu)}{n} + \frac{\frac{1}{16(p-\frac{1}{2})^2} - \frac{1}{4}}{n}. \tag{2}$$

In the special case with $1 - p = 0$, we have the familiar results $\hat{\mu} = \bar{y}$ and $V(\hat{\mu}) = \mu(1 - \mu)/n$.

The advantages of this procedure include (1) a quantification of privacy protection by the choice of p (the closer to 0.5, the more privacy); (2) a reduction in bias, indicated empirically by how sensitively the respondent views the question and how much that would affect their answer; and (3) a plausible (but to-be-empirically-validated) increase in the likelihood that a potential survey respondent will participate in the survey at all. The disadvantage of the procedure is the introduction of noise, which we quantify in terms of an increase the variance or, more specifically, the second term in Equation 2. We build on all these features in the following sections.

3.2 Basic Definition

We now define differential privacy and then give randomized response from Section 3.1 as a special case. First consider two $\underline{d}$atasets D and D' that differ by, at most, one respondent (with mnemonic notation indicated by the corresponding underline). In a standard rectangular survey dataset with one row per respondent, D' is the same as D except that one row may have been swapped out with the data from another respondent or removed entirely. Then define a $\underline{m}$echanism $M(D)$ to be a statistical estimator (i.e., a function of the data) that also includes random noise somewhere in the calculation. A mechanism $M(\cdot)$ is said to be *ϵ-differentially private* if $M(D)$ is indistinguishable from $M(D')$, in the following sense (Dwork, McSherry, et al., 2006):

$$\frac{\Pr[M(D)=m]}{\Pr[M(D')=m]} \leq e^{\epsilon}, \tag{3}$$

for any value m (in the range of $M(D)$), for a discrete sample space, and where ϵ is a policy choice made by the data provider that quantifies the maximum level of privacy leakage allowed, with smaller values potentially giving away less privacy. For small values of ϵ, Equation 3 can be written more intuitively as $\Pr[M(D)=m]/\Pr[M(D')=m] \in 1 \pm \epsilon$ (because $e^{\epsilon} \approx 1+\epsilon$).

The probabilities in this expression stem from the randomness in the mechanism (treating the data as fixed); thus, the choice of ϵ determines the amount of noise required. In fact, the similarity between ϵ in this general case and p in randomized response from Section 3.1 (where smaller values of p imply less randomness and thus less privacy) is not accidental: given a choice for ϵ, a randomized response mechanism is ϵ-differentially private if the spinner area p is computed as $p = 1/(1+e^{\epsilon})$.[1]

[1] Denote p_{jk} as the probability of truth $y^* = j$ and observed response $y^{(p)} = k$. Then "the randomized response mechanism" satisfies Equation 3 and so is ϵ-differentially private if $\max(p_{00}/p_{10}, p_{11}/p_{01}) \leq e^{\epsilon}$, with m fixed. The solution to this expression is $p_{10} = p_{01} = 1/(1+e^{\epsilon})$.

3.3 Local Differential Privacy

One broad distinction commonly made in the differential privacy literature is that between the "local model" — where noise is added at Points ① or ② in Figure 1 — and the "central model" — where noise is added at Points ③, ④, or ⑤ (Vadhan, 2017, Section 7.9.2). In this section we focus on the local model, with noise added as data leaves the respondent's device (Point ②). We do this by first representing a survey dataset in a useful way for this problem and then, second, generalizing the classical randomized response mechanism by applying it to all observed survey responses in a dataset to the respondent's nominal answers, at ② rather than as part of the survey question in ①.

First, we introduce three data representations, along with an example of each in Table 1. Panel (a) gives the most common representation of raw data with three dichotomous survey questions y, x, and z, coded for n individuals, one in each row. With three dichotomous survey questions, only $2^3 = 8$ data patterns can be found in Panel (a) and so we more compactly represent the same data in Panel (b) with these 8 rows and a count for each. Finally, we present the same data a third way by reforming it into a traditional contingency table in Panel (c). Subscripts of the counts in Panels (b) and (c) provide the crosswalk between different data forms by defining the function $i(k)$ as returning the first row from individual level data in Panel (a) with the same values of the variables y, x, and z as a corresponding row in Panel (b). The choice of the first row is an arbitrary choice to remove ambiguity, as all rows of the same type have the same values.

(a) Respondents

i	y	x	z
1	0	0	1
2	0	1	0
3	0	1	0
4	1	0	1
5	1	1	0
⋮	⋮	⋮	⋮
n			

(b) Weighted

Counts	y	x	z
$g_{i(1)}$	0	0	0
$g_{i(2)}$	1	0	0
$g_{i(3)}$	0	1	0
$g_{i(4)}$	1	1	0
$g_{i(5)}$	0	0	1
$g_{i(6)}$	1	0	1
$g_{i(7)}$	0	1	1
$g_{i(8)}$	1	1	1

(c) Tabular

		y	
z	x	0	1
0	0	$g_{i(1)}$	$g_{i(2)}$
	1	$g_{i(3)}$	$g_{i(4)}$
1	0	$g_{i(5)}$	$g_{i(6)}$
	1	$g_{i(7)}$	$g_{i(8)}$

Table 1: Three Representations of the Same Data. (The subscript $i(k)$ of the counts in Panels (b) and (c) is a function that returns the first index value i from (a) with the same values of x, y, and z as the corresponding row in (b) or cell in (c).)

We now generalize these data representations. Consider a survey where question q ($q = 1, \ldots, Q$) has c_q possible response categories. (That is, we use the fact that in most surveys almost all questions are have discrete responses, or can be recoded into discrete categories without loss of much information.) Next, form the Cartesian product of all possible answers to all the survey questions, which has cardinality $K = \prod_{q=1}^{Q} c_q$. Then define a $K \times Q$ matrix R with columns q referring to survey questions and rows k denoting possible response patterns across all questions (R is represented in Panel (b) in Table 1). All of the survey responses for a respondent (more commonly represented by the set of Q responses) can be represented by the matching row of R. We do this via a *one-hot encoding* for individual i, where $r_{ik} = 1$ for row k of R when all survey responses match and $r_{ik'} = 0$ for all $k \neq k'$. "One-hot" means that only one element of the K-vector r_i is 1 and so $\sum_{k=1}^{K} r_{ik} = 1$. (We can picture a one-hot encoding in Table 1, Panel (b), for an individual as an extra column of this table with a 1 in the row that matches all the respondents answers and a zero for all other rows.)

Second, we now add noise by applying the randomized response mechanism to each element in each respondent's one-hot vector: With probability $1-p$, we flip each bit from 0 to 1 or 1 to 0, and keep it the same with probability $p = 1/(1+e^{\epsilon/2})$. That is, we draw z_{ik} from Bernoulli(p) for all i and k, and then compute a differentially private one-hot vector, r_i^{dp}, with $r_{ik}^{\text{dp}} = (1 - z_{ik})r_{ik} + z_{ik}(1 - r_{ik})$ for each k. The device then sends this vector (or a compact version of it) to the server without fear that the respondent's privacy could be violated. The server then sums up all the one-hot vectors, resulting in differentially private or "noisy" counts: $g_{i(k)}^{\text{dp}} = \sum_{i=1}^{n} r_{ik}^{\text{dp}}$.

Although we show how to analyze the data in Section 4, we pause here to emphasize that, just as in Section 3.1 for classical randomized response, this noise biases even the individual counts. However, we can compute unbiased estimates of the true counts, which we label $\hat{g}_{i(k)}$, in the same way as for classical randomized response. We merely substitute in $1/(1 + e^{\epsilon/2})$ for p and $g_{i(k)}^{\text{dp}}$/n for $\bar{y}$ into Equation 1, which gives the unbiased estimate. The same substitutions in Equation 2 give the variance.

3.4 Central Differential Privacy

To add noise on ingest to the server (③, Figure 1), we use the "central model" of differential privacy applied to a one-hot encoding of all the data. The server begins by creating a K-vector of all zeros. Then we add noise to each element by adding a draw from the Laplace distribution (which it turns out makes it easy to satisfy Equation 3). Then each device sends its own raw one-hot encoded vector r_i (see Section 3.3) to the server and it is added one at a time to this vector. If any identifying information comes along, such as the IP address, it is deleted. This means that the raw data from each respondent is only available on the server for long enough to be aggregated with the noise and other respondents' answers.

More formally, this centralized noisy mechanism, applied to counts in the one-hot vector, produces $g^{\text{dp}}_{i(k)} = \sum_{i=1}^{n} r_{ij} + e_k$, with $e_k \sim \text{Laplace}(1/\epsilon)$ which is ϵ-differentially private for all k (because the "sensitivity", the maximum change in the count due to any one respondent, is 1) (Dwork and Roth, 2014, p.32ff).

Since Laplace noise is mean zero, the differentially private count of each element of the one-hot vector is an unbiased estimate of the true count: $E(g^{\text{dp}}_{i(k)}) = g_{i(k)}$. The variance of $g^{\text{dp}}_{i(k)}$, conditional on the private data, is simply the variance of the chosen Laplace distribution, $V(g^{\text{dp}}_{i(k)}|g_{i(k)}) = 2/\epsilon^2$.

3.5 Mixed Local and Central Differential Privacy

Local and central differential privacy are fundamentally connected in ways we can use to our advantage. In particular, a differentially private mechanism evaluated at the local level on many devices (device output, ② in Figure 1) produces stronger privacy guarantees after aggregating when evaluated at the central level (server ingest, ③). That is, an ϵ_ℓ-differentially private mechanism for device output implies an ϵ_c-differentially private mechanism after server ingest, with $\epsilon_c \ll \epsilon_\ell$.

In fact, we can make a choice for ϵ_c and deduce the (larger) implied value for ϵ_ℓ (Erlingsson, Feldman, et al., 2020), meaning that we can use the modified randomized response mechanism for device output in Section 3.3 to achieve a chosen central guarantee. The advantage of this strategy is that for the same ϵ_c at server ingest — that normally comes with no privacy protections at device output — we can also offer some additional (albeit small) privacy guarantee there without any additional cost in terms of noise. We do this by adding a small amount of noise to each of the n local devices. We illustrate this result in the left panel of Figure 2, which shows that for any level of local privacy guarantee (on the horizontal axis), we can obtain a much tighter privacy guarantee at the central level (on the vertical axis), with the effect increasing in n (different lines).

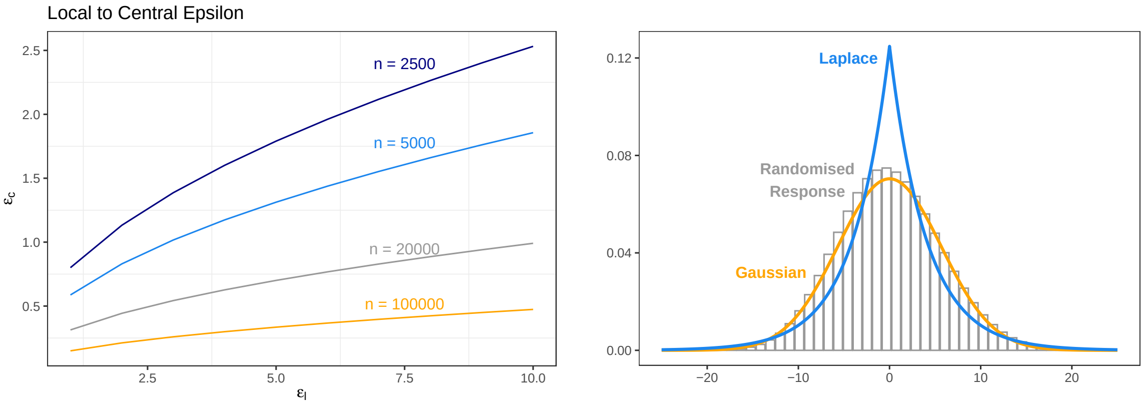

Figure 2: Local Epsilon for Central Guarantees (left panel) and different noise distributions, with Gaussian for comparison (right panel).

Although the exact value of ϵ_c can thus be ensured by either using Laplace noise following Section 3.4, or the modification of randomized response noise following Section 3.3, the statistical consequences are not identical since the randomized response and Laplace mechanisms and noise distributions differ. We convey these differences in the right panel of Figure 2. This panel includes a Laplace distribution, the implied distribution of error in the bias corrected count from modified randomized response, and a Gaussian distribution for comparison.[2]

[2] We can also construct a mechanism that gives the same central guarantee as the Laplace mechanism but with the addition of some local protection as well (see Balle et al., 2019). Thus, on each device, for each of the K elements of the one-hot vector, add a random variable $v_{ik} \equiv X_{ik} - Y_{ik}$, where $\{X_{ik}, Y_{ik}\}$ are independent Polya$(1/n, \alpha)$ variates. Since $\sum_{i=1}^{n} v_{ik} \equiv Z_k \sim \text{DLap}(\alpha)$, this approach is equivalent to adding Laplace noise on the server (i.e., the sum of the local noise is distributed discrete Laplace), with a central privacy guarantee of $\epsilon = \log(1/\alpha)$.

4 Statistical Methods

As methods for randomized response in Point ① are familiar to social scientists (e.g., Blair, Imai, and Zhou, 2015), and methods for implementing differential privacy for Points ④ and ⑤ in Figure 1 are already available outside the survey context (e.g., Evans, King, et al., 2020), we focus here on developing methods for Points ② and ③, and their combination. Most importantly, these are also points where privacy protection can provide the most value for the vast majority of sample surveys in current use throughout academia, government, and private industry. We return to the others in Section 6.

Our specific goal here is a method that can estimate the same quantities of interest from the same statistical models as we would if we had observed the private data (i.e., without noise). Thus, consider an analysis of data in the form of Table 1, Panel (a), which is the usual $n \times Q$ data matrix of respondents by survey questions (or recodes from these questions), from which we construct a binary outcome variable y_i and a vector of explanatory variables x_i. Assume, as is typically used survey data, that rows are independent (conditional on covariates and any hierarchy or clustering), and $y_i \sim \text{Bernoulli}(\pi_i)$, with a logistic regression expressing the relationship between the two:

$$\Pr(y_i = 1) \equiv \pi_i = \frac{1}{1 + e^{-X_i\beta}}. \tag{4}$$

Then the unknown parameter β (or, given chosen values of X, a derived quantity like a probability, risk ratio, or risk difference) is the quantity of interest. (Equation 4 could be easily generalized to multinomial or ordinal logit.)

Without noise, we would estimate β by simply maximizing its log-likelihood:

$$\begin{aligned}
\ln L(\beta) &= \sum_{i=1}^{n} y_i \ln \pi_i + (1 - y_i)\ln(1 - \pi_i) \\
&= -\sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)X_i\beta}\right) && (5) \\
&= -\sum_{k=1}^{K} g_{i(k)} \ln\left(1 + e^{(1-2y_{i(k)})X_{i(k)}\beta}\right) && (6) \\
&= \sum_{k=1}^{K} g_{i(k)} f_k(\beta) && (7)
\end{aligned}$$

where Equation 5 uses the individual data representation portrayed in Panel (a) of Table 1 with index i; Equation 6 uses Panel (b) with weights defined as the cell values $g_{i(k)}$ with index k; and Equation 7 simplifies (in a way that will enable us to generalize to other statistical models below) by letting $f_k(\beta) = -\ln\left(1 + e^{(1-2y_{i(k)})X_{i(k)}\beta}\right)$ for the logit model.

We now provide three methods to estimate β from a differentially private dataset in the form of Panels (b) or (c). We begin with a simple intuitive approach that turns out to have limited usefulness, then describe a simple, fast, and approximately unbiased approach that is inefficient in certain circumstances, and finally introduce a full information approach that is computationally more intensive but approximately unbiased and efficient even in these circumstances.

4.1 Nonparametric Reconstruction

Without noise, we could estimate the logistic regression coefficients of interest by reconstructing the individual-level $n \times Q$ dataset directly from each of the counts, $g_{i(k)}$ and then maximizing the log-likelihood in Equation 4 directly (cf. Liu, 2016; Quick, 2019).

To use this idea in a dataset with noise, we must first address the issue of negative and non-integer valued noisy counts $g^{\text{dp}}_{i(k)}$. We describe an intuitive but naive approach to this problem here by first replacing each $g^{\text{dp}}_{i(k)}$ with the best nonparametric estimate of $g_{i(k)}$ (for all k), which we obtain by simply rounding each to its nearest non-negative integer value and then reconstructing an estimate of the full $n \times Q$ matrix.

The problem with this approach is that rounding negative values up to zero induces systematic bias for most statistics involving the whole dataset, even though rounding to zero is the best estimate for each observation considered on its own. Because the sample space of the true counts $g_{i(k)}$ is asymmetric, only noise can cause us to observe, and hence correct for, $g^{\text{dp}}_{i(k)} < 0$. And we have no indication for any one observation of how to adjust for noisy counts where $g^{\text{dp}}_{i(k)} \geq 0$.

Although the approach is intuitive, nonparametrically optimal at the individual observation level, and can thus sometimes be advantageous in studying small numbers of counts, it is not recommended in most situations due to the bias.

4.2 Log-Linear

We now develop a consistent and approximately unbiased method of estimating β in a logistic regression model. We also show how the same technique can be used for estimating many other types of statistical models. We chose the log-linear model (LLM) name for this methodology based on its surprising connections to the classic log-linear modeling literature for contingency tables (see Appendix A).

Although the maximum likelihood estimates can be found in a variety of ways, an approach that will also prove convenient for differential privacy is the score equation. That is, without noise, we merely find the value of β in a logistic regression that satisfies

$$\frac{\partial \ln L(\beta)}{\partial \beta} = 0. \tag{8}$$

However, under differential privacy, $g_{i(k)}$ is not observed and we instead disclose its noisy, unbiased estimate $g^{\text{dp}}_{i(k)} = g_{i(k)} + v_{i(k)}$, where $v_{i(k)}$ is the mean zero noise. In this situation, instead of maximizing Equation 7 using Equation 8, we maximize

$$\ln L^{\text{dp}}(\beta) = \sum_{k=1}^{K} g^{\text{dp}}_{i(k)} f_k(\beta) \tag{9}$$

which is intuitive and, because it can be recognized as a set of "unbiased estimating equations" (Desmond, 2014), is easy to show that it is consistent. We can obtain point estimates by using the score equation, after taking expected values over the noise,

$$E\left[\frac{\partial \ln L^{\text{dp}}(\beta)}{\partial \beta}\right] = \sum_{k=1}^{K} g_{i(k)} \frac{\partial f_k(\beta)}{\partial \beta} + \sum_{k=1}^{K} E(v_{i(k)}) \frac{\partial f_k(\beta)}{\partial \beta} = \frac{\partial \ln L(\beta)}{\partial \beta},$$

setting it to zero, and solving.

This result also easily generalizes to many statistical models beyond logit: Any statistical model with a likelihood function that can expressed (as in Equation 7) as the counts, to which noise will be added to make it differentially private, multiplied by a function $f_k(\beta)$, without the need for noise, can be estimated by satisfying Equation 9. Appendix B shows how to compute standard errors.

4.3 Full Information

Without noise, LLM (Section 4.2) and the full information maximum likelihood (FIML) approach described here are identical, when the model can be expressed as a model for the counts. They are also identical with or without noise under a "fully saturated" model specification (i.e., with all possible higher order interactions). With both added noise and some dimensions of the contingency table omitted in the logistic regression, however, LLM is no longer full information, meaning that information is available to improve our estimates of β in Equation 4.

We begin with the *complete-data likelihood*, which is the likelihood we would use if we had observed not only the observed noisy counts $g_{i(k)}^{\text{dp}}$ but also with the unobserved true counts, $g_{i(k)}$:

$$\mathcal{L}(\lambda; g, g^{\text{dp}}) = \prod_{k=1}^{K} p(g_{i(k)}^{\text{dp}} \mid g_{i(k)}) p(g_{i(k)} | \lambda_{i(k)}), \tag{10}$$

where the second factor, $p(g_{i(k)} | \lambda_{i(k)})$, is the distribution of the data without noise, a Poisson distribution from the log-linear model approach. The first factor, $p(g_{i(k)}^{\text{dp}} | g_{i(k)})$, is the noise distribution given the true counts. Under our centralized model, this distribution is Laplace$(1/\epsilon)$ and under the local model Appendix C.1 proves that it is Binomial$(n, g_{i(k)}(2p-1)/n+1-p)$ where $p = 1/(1+e^{\epsilon})$. For either process, we integrate over the complete-data likelihood in Equation 10 to derive the *likelihood* (which, in this context, is sometimes called the "observed data likelihood"):

$$\mathcal{L}(\lambda; g^{\text{dp}}) = \prod_{k=1}^{K} \sum_{g=0}^{\infty} p(g_{i(k)}^{\text{dp}} | g) p(g | \lambda_{i(k)}) \tag{11}$$

where g is a the dummy variable used in the summation to denote one of the logically possible values that $g_{i(k)}$. Letting $\lambda_{i(k)} = e^{x_{i(k)}\beta}$, our FIML estimator involves maximizing Equation 11 with respect to β.

Because maximizing this expression directly is computationally expensive, Appendix C gives a faster Expectation-Maximization (EM) algorithm and an even faster approach based on an approximation that gives almost identical answers (in all simulated and real examples we have studied). Of course, LLM is faster than either of these approaches, at the cost of some efficiency when extra variables are collected but not used. See Appendix B for variance estimation and Appendix D for intuition.

5 Analyses

We now use the methods described in Section 4 to analyze differentially private simulated data in Section 5.1 and a real survey which we make differentially private in Section 5.2.

5.1 Simulations

To evaluate the methods introduced in Section 4, we generate $x_i \sim \text{Bernoulli}(0.8)$, and $y_i \sim \text{Bernoulli}(\pi_i)$, with $\pi_i = [1 + \exp(-0.5 - \beta x_i)]^{-1}$, with the quantity of interest set to $\beta = 1.5$. We also modeled other variables to be collected by generating z_i from a discretized Beta distribution with either 23 or 53 categories (since our methods only use the set of categories created by the Cartesian product of all responses to all variables, this is equivalent to multiple variables with fewer categories, so long as $K = \{92, 212\}$). (We also studied numerous other simulation designs and found similar results and so only present this one version.)

We begin with Figure 3, which we construct by running all three methods on 200 simulated datasets, with $\epsilon_c = \epsilon_\ell = 1$, $n = 5,000$, for $K = 92$ (in light blue) and $K = 212$ (in dark blue). Each point in this figure is an estimate of β (represented vertically), with the truth indicated by a dashed horizontal line at 1.5. Each of the boxes summarize the individual results with the mean indicated by a line at the midpoint and endpoints at the 25th and 75th percentiles.

The two plots at the left of Figure 3, showing simulations under nonparametric reconstruction (NP), clearly indicate the large biases of this approach, as most of the dots fall below the line, with the degree of bias increasing in K. Although both show attenuation, we could have used other specifications resulting in bias in any direction.[3] In contrast, both log-linear modeling (LLM in the center) and full information maximum likelihood (FIML, at the right) are approximately unbiased for both levels of K, centered as they are around the truth of 1.5. The figure also reveals that the log-linear model approach has a variance that increases in K — which can be seen by the length of the boxes — but the

[3] For example, consider a simulation with an additional covariate that was an essential confounder, meaning that dropping it would flip the estimated sign on the quantity of interest. As K increases, the power of this confounder drops and the sign will switch. The same result occurs as ϵ decreases.

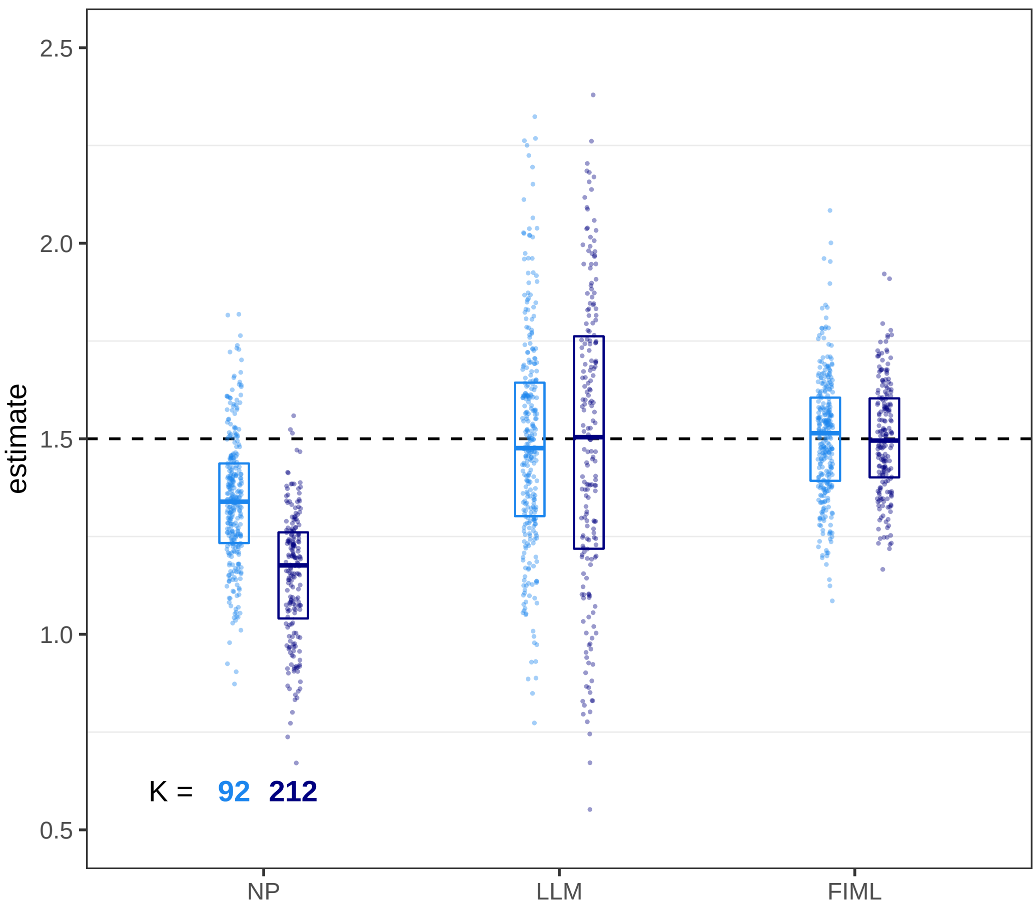

Figure 3: Comparison of Methods Under Different Numbers of Survey Question Categories: Nonparametric reconstruction (NP), Log-linear (LLM), and Full Information (FIML). Each dot is the result of the analysis of one simulated data set with an estimate of β on the vertical axis. Each plot includes slight horizontal jitter for graphical clarity.

variance (and unbiasedness) of FIML remains steady in K.

We summarize the conclusion from Figure 3 in Figure 4, Panel (a). This panel plots histograms of estimates from each of the three methods in 300 simulated datasets with $n = 1,000$, $\epsilon_\ell = 3.5$, and the truth ($\beta = 1.5$) marked with a vertical dashed line. As can be seen, nonparametric reconstruction (NP in black) is biased, FIML (in orange) and LLM (in blue) are each unbiased (centered around the truth), and LLM has higher variance than FIML.

We now analyze LLM and FIML in more detail. Panel (b) plots root mean square error vertically by the variance of the differentially private noise (as a function of ϵ, horizontally). This panel shows that FIML has lower root mean square error for all levels of

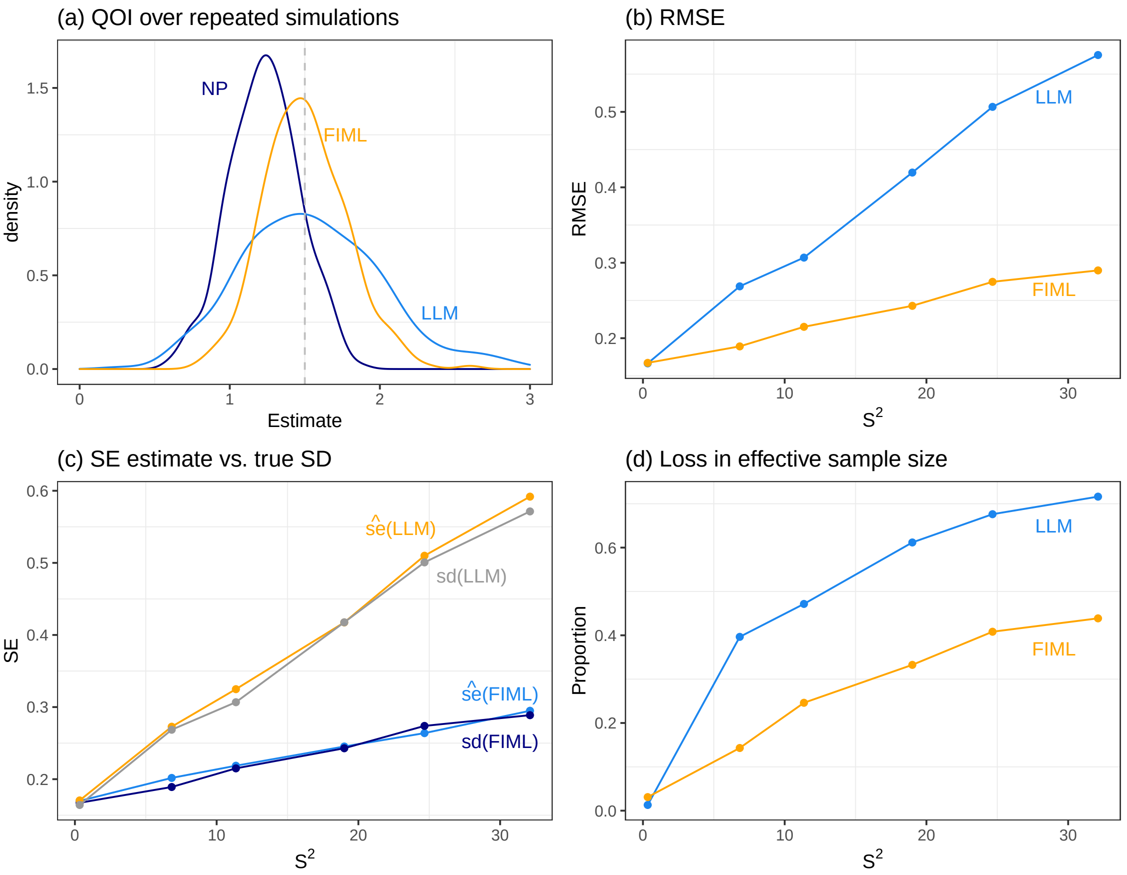

Figure 4: Statistical Properties of Full Information (FIML) and Log-Linear (LLM) approaches. The horizontal axes of panels (b), (c), and (d) is S^2, the variance of the counts induced by the noise controlled by ϵ.

noise (and privacy protection) than LLM, with the advantage growing as the noise grows.

Because our estimators are approximately unbiased, the cost of the noise induced to protect privacy under our framework is limited to increased uncertainty. This uncertainty is accurately measured by our standard errors, as can be clearly seen in Panel (c) of Figure 4, which plots the standard error vertically by the variance of the noise horizontally. In this panel, the true standard deviation of both LLM and FIML increase as the noise increases (the slope is positive for both in the graph, but much more slowly for FIML), and it is closely tracked by the estimated standard error.

We also quantify the cost due to adding noise by translating the increase in standard errors into the proportionate loss in *effective* sample size. That is, the increase in the standard error due to privacy protective noise is equivalent to not adding noise but discarding L proportion of observations (see Evans, King, et al., 2020, Section 4.2). Or, to put it positively, we can overcome the cost of this privacy protective procedure by increasing the number of survey respondents by this "lost" proportion of observations. This analysis appears in Panel (d). The horizontal axis is the variance of the counts induced by the noise; the vertical axis is the proportion of observations discarded. Graphs like this, calculated from the data, should provide important guidance for those designing, planning to share, and reporting on survey research. Of course, if the number of observations is large enough to begin with so that even with larger standard errors one's inferences are sufficiently precise, we may not need to collect additional data.

5.2 Breaking and Protecting Privacy of Abortion Attitudes

In this section, we use real political science data and verify that it is possible to break the privacy protections of the current best practices of survey researchers. We then compare the results of an analysis of (normally unobserved) private data to the corresponding analysis of privatized (i.e,. noisy) data with our statistical methods.

To accomplish these tasks, we begin with a publicly available dataset collected for a landmark article offering the "first comprehensive validation study" of randomized response methods (Rosenfeld, Imai, and Shapiro, 2016). This is especially useful because the survey contains some highly sensitive questions in need of protection. (That the same survey dataset also includes classic randomized response questions, which we cover, may also be especially useful for pedagogical purposes related to this paper.) Rosenfeld, Imai, and Shapiro (2016) surveyed 2,655 voters in Mississippi's 2011 General Election about an anti-abortion ballot initiative called the "personhood amendment," seeking to declare in the state constitution that life begins at conception. The survey asked individuals to report how they voted on this controversial initiative, along with demographic traits such as their age, gender, party ID, and education. Although numerous polls ask about support for abortion, this survey is unusual because it is about an event in which citizens were asked to act on their views by directly voting on the legal status of abortion.

First, we show that the commonly used and best practice de-identification procedures in Rosenfeld, Imai, and Shapiro (2016) in fact did not protect respondents' privacy. We do this through what is known as a "re-identification attack," with which we were able to show that information publicly available was sufficient to unambiguously identify individuals in the survey data. To do this, we followed now standard ethical practices from the cybersecurity literature by privately informing the authors of our discovery prior to publishing this paper. We then helped the authors with an alternative data sharing methodology that protected their respondents' privacy, and which they graciously implemented. We also went further than the literature's ethical standards by limiting ourselves to developing the procedure for our re-identification attack, and demonstrating that re-identification was possible, so that we never needed to actual view or distribute the names of those we showed it was possible to identify.

For obvious reasons, we do not provide replication information for this part of our analysis, but we can give a sense of how re-identification attacks work. The idea is to find survey respondents uniquely characterized by the full cross-classification of a set of variables that are available in both the survey and an outside data source describing the population from which the survey was drawn. Finding an isolated person in one of the cells of this cross-classification indicates that it is possible to find the identity of that person (a step we skipped). Although most such attacks marshal multiple datasets (Henriksen-Bulmer and Jeary, 2016), we used only the data in the article's replication file and one publicly available data source. Thus, without other external information, unusual detective work, or resorting to probabilistic methods, we were able to uniquely re-identify more than a dozen individuals in this dataset — thus making it possible to learn their names, addresses, private answers to sensitive questions about abortion preferences, and more.

Second, we show how, if we had distributed a noise-infused dataset rather than the original, we would have been able to fully protect the privacy of every respondent in the dataset. This procedure would thus enable researchers to substitute mathematical privacy guarantees for mere attestations of how hard they may have tried to protect respondent privacy. And importantly, we also demonstrate that, despite the bias-inducing noise, our statistical methods generate unbiased estimates of quantities of interest to political scientists, along with accurate uncertainty estimates.

To illustrate this result, we focus on support for legal abortion, as indicated by reporting to have voted “No” on the ballot initiative. Surprisingly, prior research in public opinion polls — outside the context of a specific ballot initiative — reveal few gender differences in support for abortion. Our quantity of interest is therefore the difference between men and women in the probability of voting “No” — both the raw descriptive result and the same estimate adjusting for partisan identification as an obvious confounder.

We now treat the original data as “private” and also create a differentially private dataset by adding noise as described in Section 3.3 (with a value of the privacy parameter of $\epsilon = 0.5$). Our first simple descriptive analysis appears in Panel (a) of Figure 5. The vertical dashed blue line indicates that the “private” data shows that men vote “No” 0.31 (31 percentage points) more in favor of legal abortion than women. We then privatize 1,000 data sets (adding noise each time following the same procedure) and, in each, estimate the same quantity from the noisy data with our corrections. The distribution of these estimates reveals the distribution of noise around this result, but centered around the truth, revealing that our estimator is unbiased. Since this distribution does not reflect sampling uncertainty, the cost of the privacy protection is seen (only) in the extra noise illustrated by this distribution. See Appendix E for further details.

Finally, we also give results adjusting for partisan identification with logistic regression; see Panel (b) of Figure 5. The vertical axis in this Panel is the quantity of interest, with a point estimate as a dot and the 95% confidence interval given as a vertical bar. At the left of this figure is the (normally unobserved) result of analyzing the private data, using a logistic regression. Here, we see that men have a predicted probability of voting “No” 0.175 higher than women, which is smaller than the raw 0.31 figure but still much larger than in traditional opinion polls outside the context of a referendum, with a sufficiently narrow confidence interval. We also analyze the privacy protective data and reveal approximately the same point estimate (indicating unbiasedness) with slightly larger confidence intervals — the cost of privacy protection. The confidence intervals here reflect both sampling uncertainty and uncertainty due to the differentially private mechanism. Of course, if smaller confidence intervals were required, we could have collected additional observations, but the substantive conclusion about the large gender gap in attitudes toward abortion is clear either way.

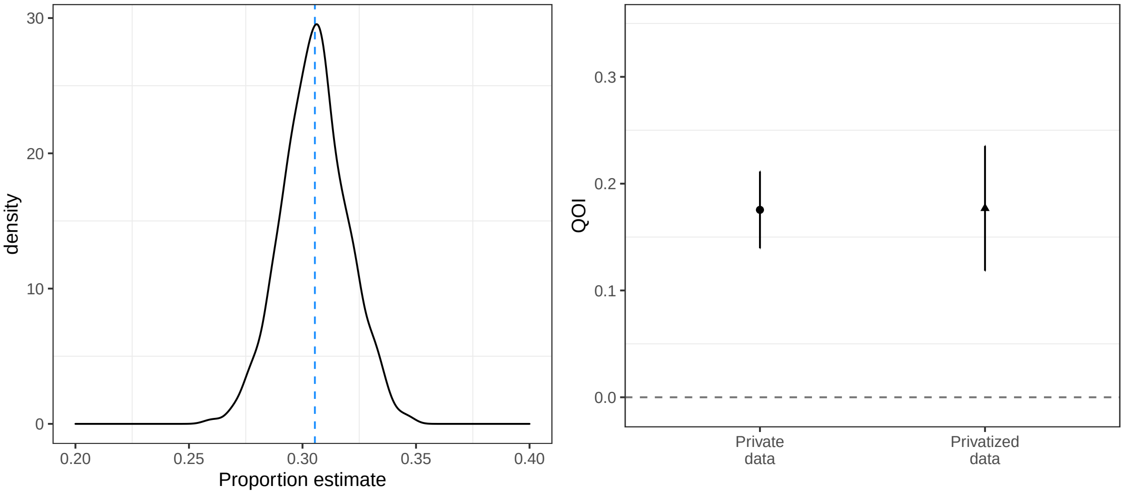

Figure 5: Comparison of Data Analysis: Private to Privatized

6 Practical Issues

The best practices of privacy protected survey research includes all the best practices of classical survey research fine tuned over many decades (e.g., Rossi, Wright, and Anderson, 2013). To these, we now add practices related to data analysis protected by differential privacy. The idea of differential privacy, described formally in Section 3.2, is to give any one person plausible deniability for having taken the survey at all. For any given amount of privacy (i.e., for any ϵ), deniability is enhanced and thus the noise we need to add relative to the signal is reduced with (1) more respondents and (2) fewer or less informative survey questions. A survey with a larger n makes finding, distinguishing, and thus re-identifying any one respondent more difficult; similarly, less informative survey questions makes it easier for one person to be confused with a larger number of others. We discuss these issues here and elaborate with other approaches in Appendix F.

It turns out that satisfying each requirement for privacy protection can be thought of as offering new reasons to follow existing best practices: First, survey researchers have always tried to collect as many observations as logistically and financially feasible, and so this practice should continue. In fact, the increase in the standard errors resulting from noise added can be eliminated by increasing the n even further.

Second, designing surveys typically involves a conflict between our expansive creativity in thinking of questions and limited space on any survey instrument. Longer surveys cost more and also tax each respondent's patience and so may increase inattention, measurement error, nonresponse, and attrition. The move to online surveys, and apparently decreasing attention spans in the general public, may be accelerating these effects. In fact, so strong is the tendency to include unneeded questions in surveys that we have long found it helpful, in advising those designing surveys for the first time, to construct simulated datasets and run simulated analyses prior to finalizing the survey instrument. Survey questions that are not used can then be dropped. Similarly, eliciting coarsened values of variables is a valuable way of limiting unneeded information collected; e.g., if an indicator for being below the poverty level is needed analysis, then eliciting a respondent's income may be wasteful (Iacus, King, and Porro, 2012).

Finally, the methods discussed in this paper can also be used for experiments, which are often survey based and conducted on platforms like Amazon's Mechanical Turk. One advantage of any fully randomized experiment, in this context, is that the assignment variable is unrelated in expectation to any other pre-treatment variables. Thus, if the researcher trusts the that the random assignment is conducted correctly and the sample size is large enough to avoid chance imbalances, we can remove these pre-treatment variables from the survey and save further on the privacy budget.

Although dropping unneeded variables saves on the privacy budget, researchers may choose to include these variables for exploratory analyses, to check on population representativeness, or in experiments to validate the success of random assignment procedures. This is the same type of decision as with the choice of pre-registering a study vs. searching for discoveries in uncharted areas.

Large numbers of or highly informative questions has an additional impact in privacy protective survey research because the amount of noise added increases with the number of questions (and their responses) to be revealed to the researcher (at Points ①, ②, or ③ in Figure 1). One way around this problem is to not reveal the (noisy) survey answers to a researcher, and instead to add noise only after computing the quantity of interest, using Points ④ or ⑤ — an approach well designed for large omnibus surveys with many questions, such as the American National Election Study, Cooperative Congressional Election Survey, the General Social Survey, and the British Election Study. In contrast, Points ①, ②, or ③ are more appropriate for more specific single-purpose surveys, as they have fewer questions and are designed to address one or a small number of research topics. For these surveys, ensuring that only needed questions are included will be helpful.

Finally, our paper adapts only the most commonly used statistical methods used in surveys in the social sciences for use in privatized data analysis. Certain other methods, such as some IRT or independent multinomial logit models, can be adapted directly from our methods. Some approaches, like methods of missing data, can be made to work within our approach, such as by coding missing values as extra categorical responses. More sophisticated methods may require extending our results or developing new approaches.

7 Concluding Remarks

Survey researchers have a long history of trying to ensure privacy for their survey respondents. Not only do respondents deserve these protections from ethical, legal, and moral perspectives, but failure to protect them will reduce incentives to participate in future research and hurt our ability to create public good from survey data.

Unfortunately, classical procedures designed to protect privacy are far less protective than the social science community had realized, just as the threats to privacy are increasing. In contrast, with the differentially private methods offered here, survey researchers can offer respondents mathematical privacy guarantees while knowing that their analytical results will remain statistically valid, both in approximate unbiasedness and accurate uncertainty intervals.

We hypothesize that, as differential privacy and the methods discussed in this paper become more widely known, respondents may be more likely to participate in our surveys and give sincere answers to sensitive survey questions. They may eventually insist on these assurances from many data collectors, from academic researchers to commercial companies. We encourage further research into the conditions that may make this outcome likely. We also hope that scholars will be able to push forward our methods and provide tighter privacy bounds, a wider range of valid statistical methods, and other convenient approaches for creating and analyzing differentially private datasets.

References

Abowd, John M (2018). “Staring-Down the Database Reconstruction Theorem”. In: *Joint Statistical Meetings, Vancouver, BC*. URL: bit.ly/census-reid.

Abowd, John M. et al. (2020). “The modernization of statistical disclosure limitation at the U.S. Census Bureau”. In: URL: bit.ly/DPcensus20.

Balle, Borja, James Bell, Adria Gascon, and Kobbi Nissim (2019). “Differentially private summation with multi-message shuffling”. In: *arXiv preprint arXiv:1906.09116*.

Blackwell, Matthew, James Honaker, and Gary King (2017). “A Unified Approach to Measurement Error and Missing Data: Overview”. In: *Sociological Methods and Research* 46.3, pp. 303–341.

Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou (2015). “Design and analysis of the randomized response technique”. In: *Journal of the American Statistical Association* 110.511, pp. 1304–1319.

Buonaccorsi, John P (2010). *Measurement error: models, methods, and applications*. CRC press.

Christensen, Garret, Jeremy Freese, and Edward Miguel (2019). *Transparent and reproducible social science research: How to do open science*. University of California Press.

Connors, Elizabeth C, Yanna Krupnikov, and John Barry Ryan (2019). “How Transparency Affects Survey Responses”. In: *Public Opinion Quarterly* 83.S1, pp. 185–209.

Desmond, Anthony F (2014). “Estimating equations, theory of”. In: *Wiley StatsRef: Statistics Reference Online*.

Ding, Bolin, Janardhan Kulkarni, and Sergey Yekhanin (2017). “Collecting telemetry data privately”. In: *Advances in Neural Information Processing Systems*, pp. 3571–3580.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth (2015). “The reusable holdout: Preserving validity in adaptive data analysis”. In: *Science* 349.6248, pp. 636–638.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006). “Calibrating noise to sensitivity in private data analysis”. In: *Theory of cryptography conference*. Springer, pp. 265–284.

Dwork, Cynthia, Moni Naor, Toniann Pitassi, Guy N Rothblum, and Sergey Yekhanin (2010). “Pan-Private Streaming Algorithms.” In: *ICS*, pp. 66–80.

Dwork, Cynthia and Aaron Roth (2014). “The algorithmic foundations of differential privacy”. In: *Foundations and Trends in Theoretical Computer Science* 9.3–4, pp. 211–407.

Dwork, Cynthia and Jonathan Ullman (2018). “The fienberg problem: How to allow human interactive data analysis in the age of differential privacy”. In: *Journal of Privacy and Confidentiality* 8.1.

Erlingsson, Úlfar, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Shuang Song, Kunal Talwar, and Abhradeep Thakurta (2020). “Encode, Shuffle, Analyze Privacy Revisited: Formalizations and Empirical Evaluation”. In: *arXiv:2001.03618*.

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*. ACM, pp. 1054–1067.

Evans, Georgina and Gary King (forthcoming). “Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset”. In: *Political Analysis*. URL: GaryKing.org/dpd.

Evans, Georgina, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta (2020). "Statistically Valid Inferences from Privacy Protected Data". In: URL: GaryKing.org/dp.

Henriksen-Bulmer, Jane and Sheridan Jeary (2016). "Re-identification attacks—A systematic literature review". In: *International Journal of Information Management* 36.6, pp. 1184–1192.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2012). "Causal Inference Without Balance Checking: Coarsened Exact Matching". In: *Political Analysis* 20.1, pp. 1–24. URL: j.mp/woCheck.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve (Mar. 2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation". In: *American Political Science Review* 95.1, pp. 49–69.

King, Gary and Nathaniel Persily (2020). "A New Model for Industry–Academic Partnerships". In: *PS: Political Science & Politics* 53.4, pp. 703–709.

Liu, Fang (2016). "Statistical Properties of Sanitized Results from Differentially Private Laplace Mechanism with Univariate Bounding Constraints". In: *arXiv preprint arXiv:1607.08554*.

Messing, Solomon, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins (2020). *Facebook Privacy-Protected Full URLs Data Set*. Version V2. DOI: 10.7910/DVN/TDOAPG. URL: https://doi.org/10.7910/DVN/TDOAPG.

Plutzer, Eric (2019). "Privacy, Sensitive Questions, and Informed Consent: Their Impacts on Total Survey Error, and the Future of Survey Research". In: *Public Opinion Quarterly* 83.S1, pp. 169–184.

Quick, Harrison (2019). "Generating Poisson-Distributed Differentially Private Synthetic Data". In: *arXiv preprint arXiv:1906.00455*.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N Shapiro (2016). “An empirical validation study of popular survey methodologies for sensitive questions”. In: *American Journal of Political Science* 60.3, pp. 783–802.

Rossi, Peter H, James D Wright, and Andy B Anderson (2013). *Handbook of survey research*. Academic Press.

Sturgis, Patrick and Rebekah Luff (2020). “The demise of the survey? A research note on trends in the use of survey data in the social sciences, 1939 to 2015”. In: *International Journal of Social Research Methodology*, pp. 1–6.

Sweeney, Latanya (1997). “Weaving technology and policy together to maintain confidentiality”. In: *The Journal of Law, Medicine & Ethics* 25.2-3, pp. 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”. In: *arXiv preprint arXiv:1709.02753*.

Vadhan, Salil (2017). “The complexity of differential privacy”. In: *Tutorials on the Foundations of Cryptography*. Springer, pp. 347–450.

Warner, Stanley L (1965). “Randomized response: A survey technique for eliminating evasive answer bias”. In: *Journal of the American Statistical Association* 60.309, pp. 63–69.

Wilson, Royce J, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson (2019). “Differentially Private SQL with Bounded User Contribution”. In: *arXiv preprint arXiv:1909.01917*.

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”. In: *Vand. J. Ent. & Tech. L.* 21, p. 209.