Differentially Private Survey Research*

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

August 23, 2020

Abstract

Survey researchers have long sought to protect the privacy of their respondents via de-identification (removing names, addresses, and other directly identifying information) before analyzing or sharing data. Although these procedures obviously help in important circumstances, recent research demonstrates that they fail to protect survey respondents from intentional attempts at re-identification, a problem that threatens to undermine vast survey enterprises in academia, government, and industry. This is especially a problem for political science because political beliefs are not only the subject of our survey questions and scholarship; they are key information respondents seek to keep private and elected representatives use to write privacy legislation. In this paper, we build on the concept of "differential privacy" to offer new survey research 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 or confidence intervals, which can be overcome with somewhat larger sample sizes.

*Our thanks to Kosuke Imai for helpful comments. The current version of this paper 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, cs-people.bu.edu/ads22, ads22@bu.edu.

¶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, Footnote 1) 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 (Connors, Krupnikov, and Ryan, 2019; Plutzer, 2019), usually by "de-identifying" data before sharing by stripping it of readily identifying information, such as name, address, phone number, and other personal identifiers. Unfortunately, a growing literature now demonstrates that these procedures (and many others, 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 one vivid example, Sweeney (1997) discovered that 87% of the US population can be individually identified with merely a zip code, gender, and date of birth. That modern surveys, even stripped of names and addresses, collect considerably more informative data than this "scares the daylights out of those responsible for curating 'public use' versions of confidential data" (Christensen, Freese, and Miguel, 2019, p.181ff).

Privacy in survey research is a special responsibility of political scientists. Political beliefs are at the nexus of our survey questions, our scholarship, and the most important information respondents seek to keep private. In democracies, privacy legislation, which we also study and which governs our ability to do our work, is based in part on elected representatives' views of these political beliefs. In autocracies, privacy of our survey respondents is essential for ensuring their safety and willingness to give sincere answers.

In this paper, we aim to begin to make it possible for survey researchers to switch from diligently *trying* to protect respondent privacy to *guaranteeing*, where possible, that respondent privacy is in fact protected. To do this, we design new data sharing procedures that simultaneously offer mathematical guarantees for the privacy of survey respondents and statistical validity guarantees for researchers analyzing privacy protected data to learn about societal patterns. We adapt to survey research procedures from the fast growing

literature on "differential privacy" (Dwork, McSherry, et al., 2006) in ways that may even satisfy regulators (King and Persily, In press). By adding specially calibrated "noise" (i.e., random numbers drawn from a specific probability distribution) to the data before sharing, differential privacy gives respondents *deniability* — not only for what they may have said to a pollster but even whether their information is in the dataset at all — and data providers and the public a rigorous quantification of possible privacy leakage.

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 statistical inferences in any direction and by any amount (Blackwell, Honaker, and King, 2017; Buonaccorsi, 2010; Evans and King, 2020). 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 new statistical 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 this "cost" may be no survey data at all. Adopting the new procedures also has a surprising benefit in that they provably prevent aspects of p-hacking and overfitting (Dwork, Feldman, et al., 2015).

We identify points where privacy protective noise can be injected into the survey research process in Section 2 and some different types of noise for each in Section 3. We introduce statistical methods that avoid noise-induced measurement error bias in Section 4, which can be interpreted conveniently in the same way as we would analyze data without noise. We evaluate and illustrate these new methods with Monte Carlo evaluations and empirical illustrations in Section 5 and give practical survey design advice in Section 6. Section 7 concludes and the appendices provide technical details.

2 Adding Privacy Protective Noise to Survey Data

We can 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.

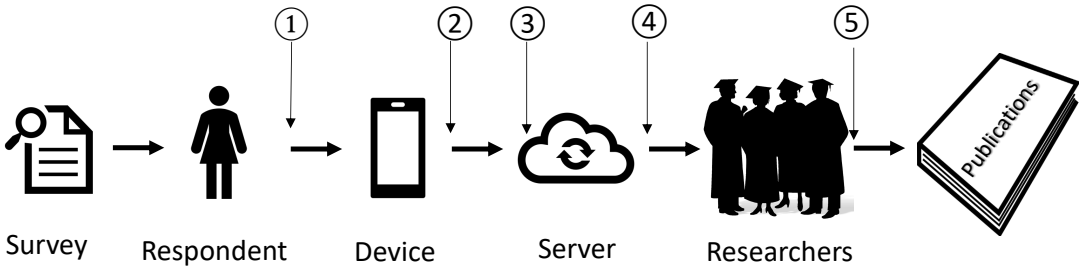

Figure 1: Points In the Survey Process where Privacy Protective Noise Can be Added: ① Randomized response, ② Device output, ③ Server ingest, ④ Server output, and ⑤ Pre-publication. A process described by each point ensures privacy via cybersecurity to its left and differential privacy to its right.

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 or other public release. This approach requires trusting researchers not only to avoid publishing private information, but also to avoid leaking information through data-dependent choices of which analyses to run (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 appear in publications come with mathematical guarantees of deniability for any survey respondent. From a statistical point of view, this stage is easy to ensure statistical validity for researchers, because mean-zero noise does not bias estimates.

Suppose instead we add noise at Point ④, *server output*. Then we only need to ensure that the server and the device are secure from cyber attacks. The researchers need not be trusted and 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 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 will not 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 adding noise at Point ②, 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 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 (Blair, Imai, and Zhou, 2015; Warner, 1965). 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 believed more likely to give an honest answer when $p > 0$ (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 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 datasets D and D' that differ by, at most, one respondent. 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 mechanism $M(D)$ 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]

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 (a) and so we more compactly represent the same data in (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.

We now generalize these data representations. Consider a survey where question q

[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 $p_{10} = p_{01} = 1/(1 + e^{\epsilon})$.

(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
$\vdots$	$\vdots$	$\vdots$	$\vdots$
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: Data Representations, for three dichotomous variables

($q = 1, \ldots, Q$) 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, where $r_{ik}^{\text{dp}} = (1 - z_{ik})r_{ik} + z_{ik}(1 - r_{ik})$ for all 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 $\hat{g}_i$, in the same way as for classical randomized response. We merely substitute in $1/(1+e^{\epsilon/2})$ for p and $g^{\text{dp}}_{i(k)}/\text{n}$ 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. 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 count mechanism 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 (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 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).

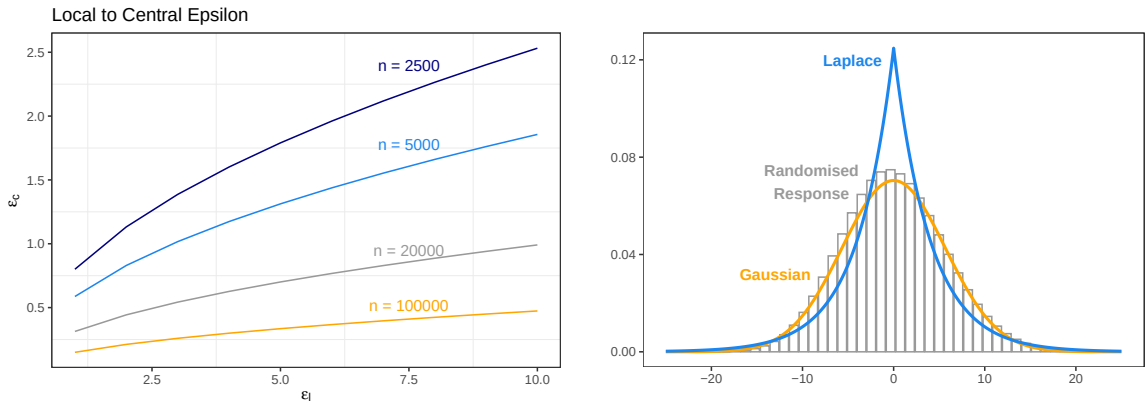

Figure 2: Local Epsilon for Central Guarantees (left panel) and different noise distributions, with Gaussian for comparison (right panel)

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 $\{X_{ik}, Y_{ik}\}$

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 (e.g., Evans, King, et al., 2020), we develop here statistical 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 appropriate for survey data, that rows are independent, 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:

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

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 $\epsilon = \log(1/\alpha)$.

where Equation 5 uses the individual data representation portrayed in Panel (a) of Table 1 with index i, and Equation 6 uses Panel (b) with weights defined as the cell values g_k with index k.

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 an unbiased but approach that is inefficient in certain circumstances, and finally introduce a full information approach that is approximately unbiased and efficient.

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.

To use this idea in a dataset with noise, we must first to 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 would induce systematic bias for most statistical quantities involving the whole dataset, even though rounding to zero is the best estimate available 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 would not be recommended in most situations due to the bias.

4.2 Log-Linear

When, as usual, survey respondents are selected independently, we can build on a result from the literature on log-linear models for contingency tables (Agresti, 2007; Christensen, 2006) to aggregate individual level Bernoulli variables, connected by a logistic regression model (using data in the form of Panel (a), Table 1), into a Poisson regression model with the counts as the unit of analysis (in the form of Panel (c)) — without any additional assumptions (Jing and Papathomas, 2020). This enables us to model the noisy counts directly, and more conveniently for our statistical purposes, and to produce the same estimate of β in Equation 4. Log-linear models were popular decades ago because of their computational advantage when a large n logistic regression was burdensome or infeasible, and also because certain types of information, such as occupational mobility tables in sociology, is more naturally represented in tabular form. We give this result here first without noise and then with modifications needed when adding noise.

To connect the two models, we write $\Pr(Y = 1|X_i) \equiv \pi_i$ as the proportion of observation counts with $y = 1$ as $\pi_{i(k)} = \lambda_{i(k)}/(\lambda_{i(k)} + \lambda_{i(k-1)})$ for expected count $E(g_{i(k)}) = \lambda_{i(k)}$, with even values of k, as in Panel (c). We then consider this count-level log-linear model:

$$g_{i(k)} \sim \text{Poisson}\left(\lambda_{i(k)}\right), \qquad \ln \lambda_{i(k)} = X_{i(k)}\gamma + y_{i(k)}(X_{i(k)}\beta). \tag{7}$$

Noting that Equation 4 can be written as $\ln[\pi_i/(1-\pi_i)] = X_i\beta$, we write

$$\ln \frac{\pi_{i(k)}}{1-\pi_{i(k)}} = \ln \lambda_{i(k)} - \ln \lambda_{i(k-1)} = X_{i(k)}\beta \tag{8}$$

which shows that β in the log-linear model representation in Equation 7 is the same quantity as in the individual level logistic regression in Equation 4. Note that the ancillary parameter γ in Equation 7, which indicates how imbalanced are the marginal values of X, is included in the individual level logistic regression representation and is orthogonal to β. It must be included in the log-linear model but estimates of it can be ignored. From a data analyst's point of view, the count-level *interaction* $(y_{i(k)} \cdot Y_{i(k)})$ between two right hand side variables in this expression — which the logit model regards substantively as

explanatory and dependent variables, respectively — enables us to estimate the effect of a *noninteracted* individual-level explanatory variable $x_{i(k)}$ on $y_{i(k)}$.

When differentially private noise is added to the counts, $g_{i(k)}$ becomes unobserved, and so we replace it with an unbiased estimate, which we call $\hat{\lambda}$ and define as either $g^{\text{dp}}_{i(k)}$ under the central model (Section 3.4) or $\hat{g}_{i(k)}$ under the local model (Section 3.3). However, even with noise added to the counts, $x_{i(k)}$ and $y_{i(k)}$ are measured without error, since indicator variable values are known exactly for each of the K rows. This fact is especially useful because then, under the log-linear model representation, random noise only appears in the outcome variable where it is less likely to bias parameter estimates (unlike error in right side variables, which always induce bias; see Evans and King 2020, Section 4.1).

Although the noisy counts — fed into the log-linear model as the outcome variable — are unbiased estimates of the true counts, we require two adjustments to the standard Poisson regression estimation procedure stemming from the added noise. The first arises because $\hat{g}_{i(k)}$ can be negative or non-integer valued. This means the paramaters cannot be estimated by directly maximizing the log-likelihood, since we obviously cannot take the log of a negative value. We avoid this problem by maximizing the likelihood via the score equations.[3]

The second adjustment is to the uncertainty estimates. Classically computed Poisson regression standard errors are too small because the outcome variable is Poisson plus noise. Thus, even if the mean specification is correct, the count will be overdispersed (i.e., unlike the Poisson, the variance will be larger than the mean; see Cameron and Trivedi 1998; King 1989). Overdispersed count data can sometimes be corrected by robust variance estimation, but in this case we know the noise process exactly and so can do substantially better, which we show how to do in Appendix A.

[3]The score equations are $\frac{\partial \ln L}{\partial \lambda_m} = \sum_{k=1}^{K} \tilde{x}_{km} \left(g_k - e^{\tilde{x}_k \lambda}\right) = 0$, where we use $\tilde{x}$ to generically represent the chosen model matrix of log-linear model specification. This equation is easily solved even when $\hat{g}_k$ takes on negative and non-integer values.

4.3 Full Information

Without noise, the LLM approach in Section 4.2 and the full information maximum likelihood approach (FIML) described in this section are identical. They are also identical with or without noise under a "fully saturated" model specification (that is a specification with all possible higher order interactions). With both added noise and some dimensions of the contingency table not needed in the logistic regression, however, LLM is no longer a full information approach, meaning that more information is available to improve our estimates of β in Equation 4.

One way to think about this extra information is in terms of a trade off: The log-linear approach in Section 4.2 is unbiased, but works by intentionally not adjusting for negative noisy counts, even though we have strong (nonparametric) information about the direction of the bias for these observations, resulting in higher variance relative to approaches that adjust unbiased count estimates known to be wrong, such as with nonparametric reconstruction in Section 4.1. Of course, nonparametric reconstruction has the disadvantage of substantial bias. The FIML approach avoids these trade offs with an estimator that is approximately unbiased and with lower variance.

We begin with estimation theory and then discuss interpretation.

4.3.1 Theory

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

$$\mathcal{L}(\lambda; g, g^{\text{dp}}) = \prod_{k=1}^{K} p(g^{\text{dp}}_{i(k)} \mid g_{i(k)}) p(g_{i(k)} | \lambda_{i(k)}), \tag{9}$$

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, $p(g^{\text{dp}}_{i(k)}|g_{i(k)})$, is the noise distribution given the true counts. Under our centralized model, this distribution is Laplace$(1/\epsilon)$ and under the local model Appendix B.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 9 to derive the *likelihood* (which, in this context, is sometimes

called the observed data likelihood):

$$\mathcal{L}(\lambda; g^{\text{dp}}) = \prod_{k=1}^{K} \sum_{g=0}^{\infty} p(g^{\text{dp}}_{i(k)}|g) p(g|\lambda_{i(k)}) \tag{10}$$

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 10 with respect to β. Because maximizing this expression directly is computationally difficult, we develop a faster EM algorithm for point and variance estimates. We also develop an even faster approach based on a distributional approximation that gives almost identical answers in all simulated and real examples we have studied. We give these results in Appendix B.2.

4.3.2 Interpretation

The connection between FIML and LLM is easiest to see in the FIML likelihood function in Equation 10 which, without noise (i.e., $\epsilon \to \infty$), simplifies to the same Poisson regression model as LLM simplifies to: $\mathcal{L}(\lambda) = \prod_{k=1}^{K} p(g_{i(k)}|\lambda_{i(k)})$. Thus, FIML will outperform LLM when (1) the underlying count estimates contain noise — meaning that $p(g^{\text{dp}}_{i(k)}|g_{i(k)})$ does not collapse to a spike at the true value, the estimated counts are overdispersed, and as a result the LLM estimates are inefficient — and (2) the FIML estimate of the probability distribution $p(g_{i(k)}|\lambda_{i(k)})$ is informative.

Condition (1) occurs when noise is added to protect privacy. Condition (2) is satisfied when the LLM estimator ignores information that FIML can take advantage of. To see where this information arises, consider again the log-linear specification on the estimated counts in Equation 7 used in both FIML and LLM and note that it is more general than the logistic specification because of parameter γ. For example, suppose we construct the estimated counts with three variables, as we do in Table 1, by also collecting survey variable z_k. This variable would not seem to be material because whether or not it is included as an additional term in Equation 7, it would not appear in the corresponding logistic model and does not change the interpretation of the other parameters or their estimates — so long as no noise is added to the counts. However, with noise added,

these extra variables that do not appear in the logistic specification can be quite important. Our LLM estimate ignores this information, but our FIML estimate extracts whatever information is available from it.

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.[4] In contrast,

[4]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.

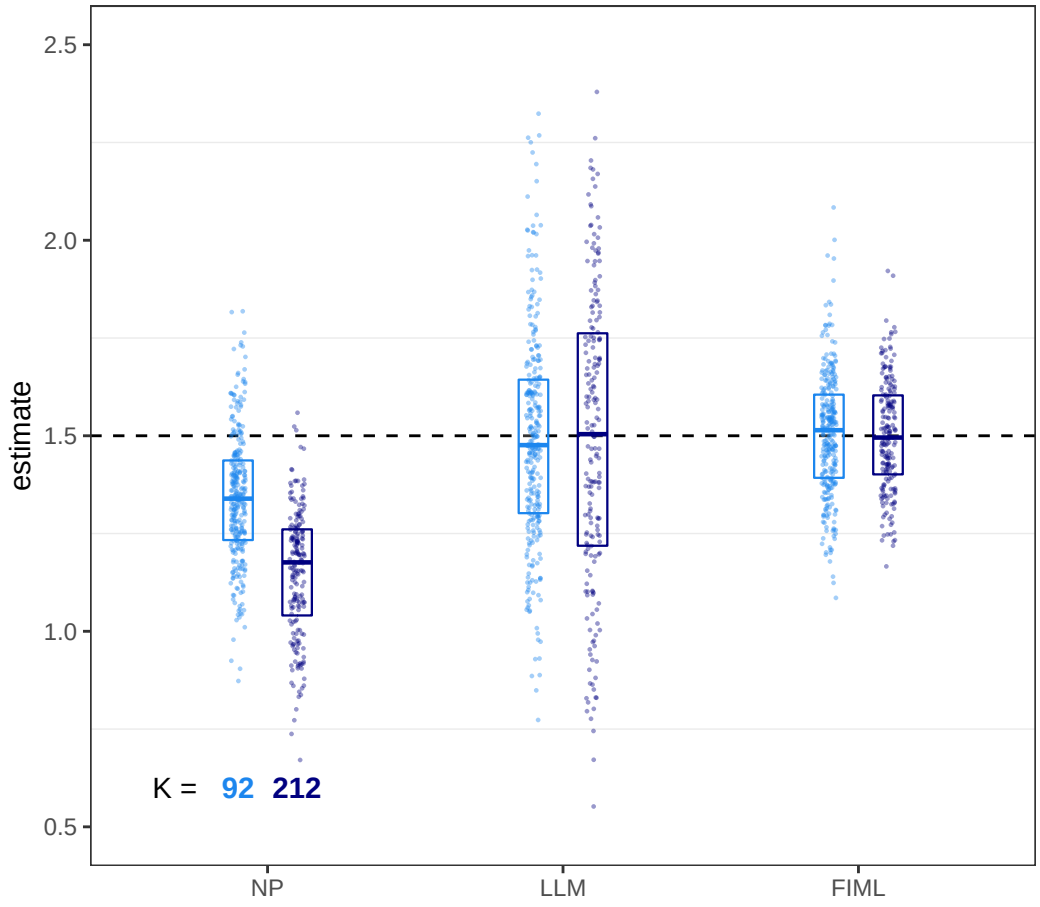

Figure 3: Comparison of Methods Under Different Numbers of Survey Question Categories: Nonparametric reconstruction (NP), Log-linear (LLM), and Full Information (FIML). Each dot is the result of the analysis of one simulated data set with an estimate of β on the vertical axis. Each plot includes slight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

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 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.

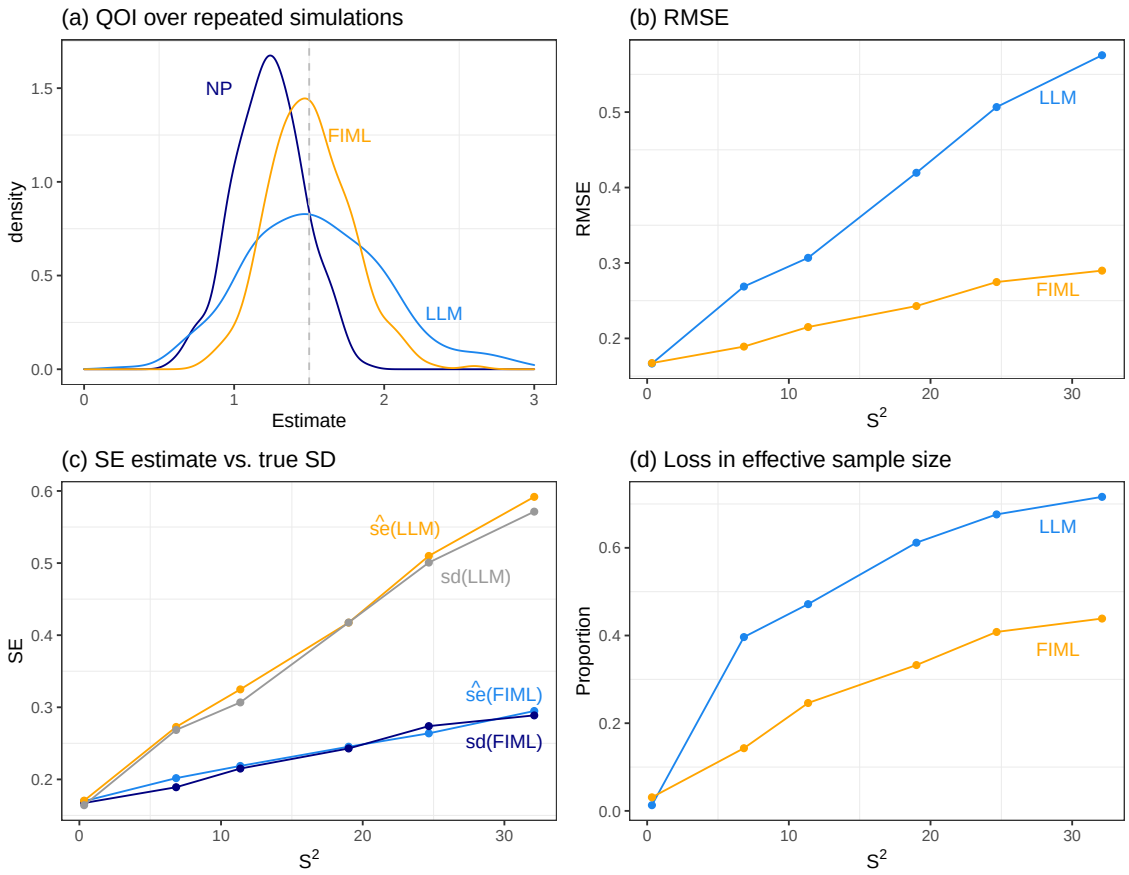

Figure 4: Statistical Properties of Full Information (FIML) and Log-Linear (LLM) approaches. The horizontal axes of panels (b), (c), and (d) is S^2, the variance of the counts induced by the noise controlled by ϵ.

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 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, a fact that 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), 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. This graph should provide important guidance for those designing, and planning to share, new surveys. 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. With the Laplace mechanism in central differential privacy, the noise also scales with n, which will also make L smaller when n is larger.

5.2 Empirical Illustrations

In this section, we use real political science data and verify the ease with which 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 direct questions in need of protection. (That the same survey dataset also includes classic randomized response questions may also be especially useful for pedagogical purposes.) 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

important 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 de-identification procedures used to protect privacy in this survey (which clearly followed best practices at the time) were insufficient. To do this, we conducted a "re-identification attack" using the data in Rosenfeld, Imai, and Shapiro (2016). Although most such attacks marshal multiple datasets (Henriksen-Bulmer and Jeary, 2016), we use only the data in the article's replication file and one publicly available data source. Without other external information, unusual detective work, or resorting to probabilistic methods, we were able to uniquely and unambiguously 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. We could have re-identified others too, but this is unnecessary for present purposes: All respondents deserve privacy protection. (For obvious reasons, we do not provide replication information for this part of our analysis.)

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" for the anti-abortion initiative — 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 of the privatized data 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.

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, which is the cost of privacy protection in estimating this quantity. The confidence intervals here reflect both sampling uncertainty and uncertainty due to the differentially private mechanism.

6 Practical Issues in Survey Design

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

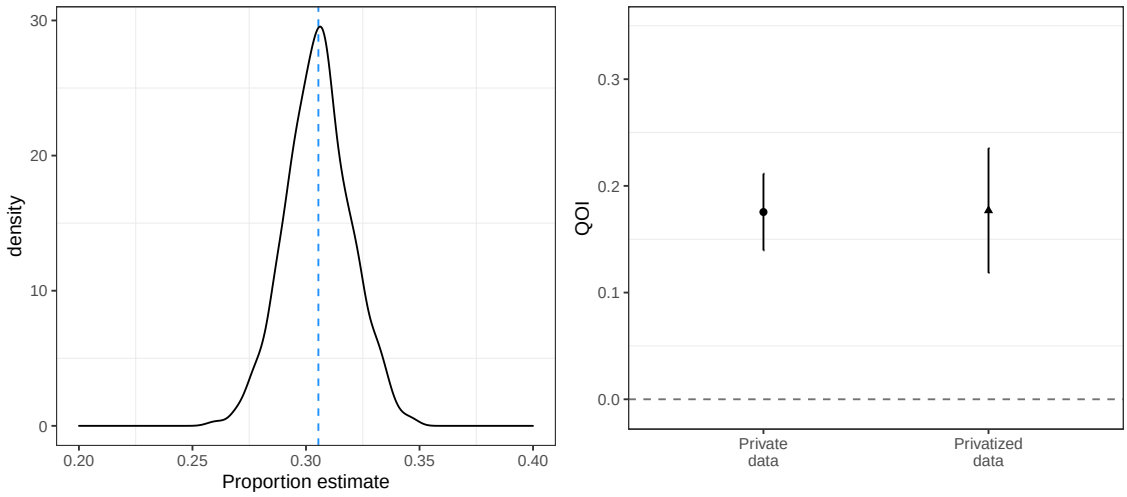

Figure 5: Comparison of Data Analysis: Private to Privatized

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 obviously makes it easier for one person to be confused with a larger number of others.

It turns out that satisfying each requirement for privacy protection merely offers a new reason 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.

And second, designing surveys typically involves a conflict between our expansive creativity in thinking of questions to add and limited space on any survey instrument. The limit on questions is not only due to cost. Longer surveys 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 simu-

lated analyses prior to finalizing the survey instrument. Survey questions that are not used or seemingly unimportant can then easily be dropped.)

Large numbers of questions has an additional impact in privacy protective survey research because the amount of noise added increases with the number questions (or question 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 the vast majority of surveys being conducted today, such as most online, commercial, and research study-specific surveys, as they have fewer questions and are designed to address one or a small number of research topics. For these surveys, ensuring that the only questions asked are those to be used will be helpful.

7 Concluding Remarks

Survey researchers have long tried hard to ensure survey respondents privacy protection. Not only do they deserve it from an ethical, legal, and moral perspective, respondents are much more likely to be cooperative, to participate in our surveys, and to give sincere answers to our questions. But effort in classical methods of privacy protection does not always guarantee success, and recent research indicates that if fails more than we would like. In contrast, with the differentially private methods offered here, survey researchers will be able to offer respondents mathematical guarantees of privacy. At the same time, researchers will be ensured that their analytical results will remain statistically valid, both in approximate unbiasedness and accurate uncertainty intervals.

We also hope and expect that researchers will be able to push forward our methods and provide tighter privacy bounds, a wider range of valid statistical methods, and other convenient methods of creating differentially private datasets.

Appendix A Log-Linear Variance Estimation

We now derive a variance estimator for our log-linear model approach from Section 4.2. Begin with the second order partial derivatives of the Poisson log-likelihood, without noise:

$$\frac{\partial^2 \ln L}{\partial \lambda_m \partial \lambda_l} = -\sum_{k=1}^{K} \tilde{x}_{km}\tilde{x}_{kl}\exp(\tilde{x}_k\lambda) = -X'WX \tag{11}$$

where $W = \text{diag}(\exp(\tilde{x}_1'\hat{\lambda}) \ldots \exp(\tilde{x}_K'\hat{\lambda}))$, and the variance is $\hat{V}(\hat{\lambda}) = (X'WX)^{-1}$.

This approach uses plug in estimators for X and W, and so will be biased under overdispersion. Under classical overdispersion, this problem can be corrected by the robust sandwich estimator,

$$\tilde{V}(\hat{\lambda}) = \left(X'\hat{W}X\right)^{-1} X'\tilde{W}X \left(X'\hat{W}X\right)^{-1} \tag{12}$$

where $\tilde{W} = \text{diag}((\hat{g}_1 - \exp(\tilde{x}_1'\lambda)^2, \ldots, \hat{g}_K - \exp(\tilde{x}_K'\lambda)^2)$, *estimates* the degree of overdispersion. In our case, however, we can do considerably better because we know the degree of overdispersion exactly. We thus instead use

$$\dot{V}(\hat{\lambda}) = \left(X'\hat{W}X\right)^{-1} X'\dot{W}X \left(X'\hat{W}X\right)^{-1} \tag{13}$$

where $\dot{W} = \text{diag}(\exp(\tilde{x}_1'\lambda) + S^2, ..., \exp(\tilde{x}_1'\lambda) + S^2)$ and S^2 is the noise in the counts induced by ϵ.

Appendix B Full Information Maximum Likelihood

We now develop algorithms for maximizing Equation 10: An EM algorithm in Section B.2, its variance estimator in Section B.3, and a fast approximation in Section B.4.

B.1 Randomized Response Distribution

We derive $p(g_k^{\text{dp}}|g_k)$ for randomized response by first recognizing that the differentially private count is the sum of two random variables: the true 1s that *are not* flipped added to the true 0s that *are* flipped. More formally, $g_k^{\text{dp}} = N_{1k} + N_{0k}$, where

$$N_{1k} \sim \text{Binomial}\left(g_k, \frac{\exp(\epsilon_l)}{1+\exp(\epsilon_l)}\right), \qquad N_{0k} \sim \text{Binomial}\left(n - g_k, \frac{1}{1+\exp(\epsilon_l)}\right). \tag{14}$$

Then, since the noise for each element of the one-hot encoded vector occurs independently, we can use the formula for the convolution of independent random variables:[5]

$$p(g_k^{\text{dp}}|g_k) = \sum_{n_{1k}=0}^{g_k^{\text{dp}}} \binom{g_k}{n_{1k}} p^{n_{1k}}(1-p)^{g_k-n_{1k}} \cdot \binom{n-g_k}{g_k^{\text{dp}}-n_{1k}} (1-p)^{g_k^{\text{dp}}-n_{1k}} p^{(n-g_k)-(g_k^{\text{dp}}-n_{1k})}$$
$$= \text{Binomial}\left(n, \left[\frac{2p-1}{n}\right] \cdot g_k + (1-p)\right) \tag{15}$$

where $p = 1/(1+e^{\epsilon})$.

B.2 EM Algorithm

We begin with the *expected complete data log-likelihood*:

$$E_g[\mathcal{L}(\lambda; g^{\text{dp}})] = \sum_{k=1}^{K} \sum_{g} \ln\left[p(g_k^{\text{dp}}|g_k = g)p(g_k = g|\lambda)\right] \cdot p(g_k = g|g_k^{\text{dp}}; \lambda), \tag{16}$$

and then write the *E-Step* by defining

$$\gamma(g; \lambda^{(t)}) \equiv p(g_k^{\text{dp}}|g_k = g) \cdot \frac{\exp(\tilde{x}'_{i(k)}\hat{\lambda}^{(t)})^g \exp(-\exp(\tilde{x}'_{i(k)}\hat{\lambda}^{(t)}))}{g!}$$

and writing

$$p(g|g_k^{\text{dp}}; \hat{\lambda}^{(t)}) = \frac{\gamma(g; \lambda^{(t)})}{\sum_{\tilde{c}_k} \gamma(\tilde{c}_k; \lambda^{(t)})}.$$

Then the *M-step* is

$$\hat{\lambda}^{(t+1)} = \arg\max_{\lambda} \sum_{k=1}^{K} \sum_{g} \frac{\gamma(g)}{\sum_{\tilde{c}_{i(k)}} \gamma(\tilde{c}_k)} \left[g\tilde{x}'_{i(k)}\lambda - \exp(\tilde{x}'_{i(k)}\lambda) - \ln(g!)\right],$$

which reflects the fact that the log-likelihood does not depend on the differentially private counts once we condition on the private counts. We implement the maximization step conveniently via weighted Poissson regression. The algorithm repeats these steps until convergence.

B.3 Variance Estimation

We denote the final EM estimate from Section B.2 by λ^* and now show how to estimate its variance. First, by the properties of maximum likelihood estimation, the limiting distribution of λ^* is $\mathcal{N}(\lambda, I(\lambda)^{-1})$, where $I(\lambda) = -E[\ln L''(\lambda, g^{\text{dp}})]$, which can be estimated by

[5] Define $g_k^{\text{dp}} = N_{1k} + N_{1k}$. N_{1k} and N_{0k} are discrete independent variables, so $P(g_k^{\text{dp}} = z) = \sum_{n_1=0}^{z} P(N_{1k} = n_1)P(N_{0k} = z - n_1)$.

the observed information matrix, $I(\lambda^*) = -\ln L''(\lambda^*; g^{\text{dp}})$. Hence our variance estimator is $[I(\lambda^*)]^{-1}$.

One of the drawbacks of the EM algorithm is that $I(\lambda^*)$ is not produced as a by-product. One option for calculating it is via brute force computation, by finding the hessian of the observed data log-likelihood function evaluated at λ^*. However is this difficult for the same reason we turned to EM rather than maximizing this likelihood directly in the first place — the observed data log-likelihood contains the log of a sum. We therefore estimate $I(\lambda^*)$ by a two-step calculation using Oakes' identity (Oakes, 1999):

$$\underbrace{-\ln L''(\lambda; g^{\text{dp}})}_{\text{Observed information}} = \underbrace{E[-\ln L''(\lambda; g, g^{\text{dp}})]}_{\text{Complete information}} - \underbrace{E[-\ln f''(g|g^{\text{dp}}; \lambda)]}_{\text{Missing information}}, \tag{17}$$

which is advantageous because estimating the complete information, $I_{g,g^{\text{dp}}}(\lambda)$, and missing information, $I_{g|g^{\text{dp}}}(\lambda)$, separately is much faster than estimating the observed information directly. We can then estimate the complete information directly from the M-step in the final iteration. The missing information is approximated using a simple Monte Carlo procedure. First note that

$$I_{g|g^{\text{dp}}}(\lambda) = \text{Var}\left[\frac{\partial \ln f(g|g^{\text{dp}}; \lambda)}{\partial \lambda}\right]$$

can be approximated by simulating datasets $\{g^{(i)}, g^{\text{dp}}\}$ for $i \in 1...N$ and then taking the sample variance over N of $\frac{\partial \ln f(g^{(i)}|g^{\text{dp}};\lambda)}{\partial \lambda}$. More explicitly, we can draw $\{g^{(i)}, g^{\text{dp}}\}$ from the distribution defined by:

$$p(g_k^{(i)} = g) = \frac{\gamma(g; \lambda^*)}{\sum_{\tilde{c}_k} \gamma(\tilde{c}_k; \lambda^*)}$$

Then we take the derivative analytically by recognizing that

$$\ln f(g^{(i)}|g^{\text{dp}}; \lambda^*) = \sum_{k=1}^{K} \ln(\gamma(g; \lambda^*)) + \ln\left(\sum_{\tilde{c}_k} \gamma(\tilde{c}_k; \lambda^*)\right)$$

which gives

$$\frac{\partial \ln f(g^{(i)}|g^{\text{dp}}; \lambda^*)}{\partial \lambda^*} = \underbrace{\sum_k \tilde{x_k}(g_k^{(i)} - \exp(\tilde{x}'\lambda^*))}_{\text{Poisson score equation}} + \frac{\partial \ln\left(\sum_{\tilde{c}_k} \gamma(\tilde{c}_k; \lambda^*)\right)}{\partial \lambda^*}.$$

Since the second term is constant with respect to $g^{(i)}$, it does not influence the sample variance and can be ignored. This conveniently allows us to avoid taking the derivative with respect to the log of a sum.

Now we have estimates of $I_{g,g^{\text{dp}}}(\lambda)$ and $I_{g|g^{\text{dp}}}(\lambda)$, we substitute in our estimates to $[I_{g,g^{\text{dp}}}(\lambda) - I_{g|g^{\text{dp}}}(\lambda)]^{-1}$ to yield our final variance estimate.

B.4 Fast Approximation for Randomized Response

Let $g_k^{dp} = g_k + v_k$, where v_k is the noise, so that

$$\begin{aligned} E[v_k] &= E[E(g_k^{dp} - g_k \mid g_k)] = E[(2p-2)g_k + (1-\pi)n] \\ &= (2\pi - 2)\text{exp}(x_k'\lambda) + (1-\pi)n. \end{aligned} \tag{18}$$

Then approximate by proceeding as if, conditional on $\{x_k, \lambda\}$, v_k is independent of g_k. Then, recognizing that the close relationship between the binomial and Poisson distributions, the distribution of v_k can be well approximated by a Poisson with parameter given by Equation 18: $v_k|x_k, \lambda \sim \text{Pois}((2p-2)\text{exp}(x_k'\lambda) + (1-p)n$.

Under this assumption, the observed data likelihood is given by:

$$\mathcal{L}(\lambda; g^{dp}) = \prod_{k=1}^{K} \sum_{g=0}^{\infty} \frac{\text{exp}(-\gamma_{1k})\gamma_{1k}^{g_k^{dp}-g}}{(g_k^{dp}-g)!} \frac{\text{exp}(-\gamma_{2k})\gamma_{2k}^{g}}{g!}$$

where $\gamma_{1k} = (2p-2)\text{exp}(x_k'\lambda) + (1-p)n$ and $\gamma_{2k} = \text{exp}(x_k'\lambda)$, which simplifies to

$$= \prod_{k=1}^{K} \underbrace{\frac{\text{exp}(-(\gamma_{1k}+\gamma_{1k}))(\gamma_{1k}+\gamma_{2k})^{g_k^{dp}}}{g_k^{dp}!}}_{\text{Poisson pmf}}.$$

We then find the maximum likelihood estimate of λ by maximizing this log-likelihood:

$$\ln \mathcal{L}(\lambda; g^{dp}) = \sum_{k=1}^{K} -(\gamma_{1k}+\gamma_{1k}) + g_k^{dp} \ln\left[\gamma_{1k}+\gamma_{2k}\right]$$

which means our approximate FIML estimator is

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{k=1}^{K} -(2p-1)\text{exp}(x_k'\lambda) + g_k^{dp} \ln\left[(2p-1)\text{exp}(x_k'\lambda) + (1-p)n\right]. \tag{19}$$

Through extensive simulation analyses and empirical tests, we find that any differences in estimates between FIML and this approximate FIML are almost always trivially small.

References

Agresti, Alan (2007): *An introduction to categorical data analysis*. John Wiley & Sons.

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*, no. 3, vol. 46, 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*, no. 511, vol. 110, pp. 1304–1319.

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

Cameron, A. Colin and Pravin K. Trivedi (1998): *Regression Analysis of Count Data*. Cambridge University Press.

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

Christensen, Ronald (2006): *Log-linear models and logistic regression*. Springer Science & Business Media.

Connors, Elizabeth C, Yanna Krupnikov, and John Barry Ryan (2019): "How Transparency Affects Survey Responses". In: *Public Opinion Quarterly*, no. S1, vol. 83, pp. 185–209.

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*, no. 6248, vol. 349, 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*, no. 3–4, vol. 9, 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*, no. 1, vol. 8.

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*.

Evans, Georgina and Gary King (2020): "Statistically Valid Inferences from Differentially Private Data Releases". In: 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*, no. 6, vol. 36, pp. 1184–1192.

Jing, Wei and Michail Papathomas (2020): "On the correspondence of deviances and maximum-likelihood and interval estimates from log-linear to logistic regression modelling". In: *Royal Society Open Science*, no. 1, vol. 7, p. 191483.

King, Gary (Aug. 1989): "Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator". In: *American Journal of Political Science*, no. 3, vol. 33. http://gking.harvard.edu/files/abs/varspecec-abs.shtml, pp. 762–784.

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*, no. 1, vol. 95. http://j.mp/lsZDuW, pp. 49–69.

King, Gary and Nathaniel Persily (In press): "A New Model for Industry-Academic Partnerships". In: *PS: Political Science and Politics*. URL: GaryKing.org/partnerships.

Oakes, David (1999): "Direct calculation of the information matrix via the EM". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, no. 2, vol. 61, pp. 479–482.

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*, no. S1, vol. 83, pp. 169–184.

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*, no. 3, vol. 60, pp. 783–802.

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

Sweeney, Latanya (1997): "Weaving technology and policy together to maintain confidentiality". In: *The Journal of Law, Medicine & Ethics*, no. 2-3, vol. 25, pp. 98–110.

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*, no. 309, vol. 60, pp. 63–69.

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.* Vol. 21, p. 209.