Differentially Private Survey Research: Supplementary (Online) Appendices*

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

September 1, 2021

*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, 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.

Contents

Appendix A Connections to Log-Linear Modeling

To provide additional intuition for how the method in Section 4.2 works, we describe the deep connection between this approach and the classic log-linear modeling literature. 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, may be more naturally represented in tabular form (Agresti, 2007; Awan and Cai, 2020; Christensen, 2006).

This literature shows, without noise, that when survey respondents are selected independently, we can 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), we can then model the counts directly and produce the identical estimate of β as an individual level logistic regression from Equation 4. We give this result here first without noise and then reveal the 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:

$$g_{i(k)} \sim \text{Poisson}\left(\lambda_{i(k)}\right), \qquad \ln \lambda_{i(k)} = X_{i(k)}\gamma + y_{i(k)}(X_{i(k)}\beta). \tag{1}$$

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

$$\ln \frac{\pi_{i(k)}}{1-\pi_{i(k)}} = \ln \lambda_{i(k)} - \ln \lambda_{i(k-1)} = X_{i(k)}\beta \tag{2}$$

which shows that β in the log-linear model representation in Equation 1 is the same quantity as in the individual level logistic regression in Equation 4. Note that the ancillary parameter γ in Equation 1, 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 Forthcoming, 2021, Section 4.1). If we use the score equation to find the maximum likelihood estimator under a log-linear model approach, we find the identical estimate as that with unbiased estimating equations in Section 4.2.[1]

Classically computed Poisson regression model standard errors are too small with noisy counts 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. The solution here is described in Appendix B, the same as we use under unbiased estimating equations.

Appendix B 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

[1] The score equations used to optimize here 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.

noise:

$$\frac{\partial^2 \ln L}{\partial \lambda_m \partial \lambda_l} = -\sum_{k=1}^{K} \tilde{x}_{km}\tilde{x}_{kl}\text{exp}(\tilde{x}_k\lambda) = -X'WX \tag{3}$$

where $W = \text{diag}(\text{exp}(\tilde{x}_1'\hat{\lambda}) \dots \text{exp}(\tilde{x}_K'\hat{\lambda}))$, 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,

$$\tilde{V}(\hat{\lambda}) = \left(X'\hat{W}X\right)^{-1} X'\tilde{W}X \left(X'\hat{W}X\right)^{-1} \tag{4}$$

where $\tilde{W} = \text{diag}((\hat{g}_1 - \text{exp}(\tilde{x}_1'\lambda)^2, \dots, \hat{g}_K - \text{exp}(\tilde{x}_K'\lambda)^2)$, *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

$$\dot{V}(\hat{\lambda}) = \left(X'\hat{W}X\right)^{-1} X'\dot{W}X \left(X'\hat{W}X\right)^{-1} \tag{5}$$

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

Appendix C Full Information

We now develop algorithms for maximizing Equation 11: An EM algorithm in Section C.2, its variance estimator in Section C.3, and a fast approximation in Section C.4.

C.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

$$N_{1k} \sim \text{Binomial}\left(g_k, \frac{\text{exp}(\epsilon_l)}{1+\text{exp}(\epsilon_l)}\right), \qquad N_{0k} \sim \text{Binomial}\left(n - g_k, \frac{1}{1+\text{exp}(\epsilon_l)}\right). \tag{6}$$

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:[2]

$$p(g_k^{\text{dp}}|g_k) = \sum_{n_{1k}=0}^{g_k^{\text{dp}}} \binom{g_k}{n_{1k}} p^{n_{1k}}(1-p)^{g_k-n_{1k}} \cdot \binom{n-g_k}{g_k^{\text{dp}}-n_{1k}} (1-p)^{g_k^{\text{dp}}-n_{1k}} p^{(n-g_k)-(g_k^{\text{dp}}-n_{1k})}$$
$$= \text{Binomial}\left(n, \left[\frac{2p-1}{n}\right] \cdot g_k + (1-p)\right) \tag{7}$$

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

C.2 EM Algorithm

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

$$E_g[\mathcal{L}(\lambda; g^{\text{dp}})] = \sum_{k=1}^{K} \sum_{g} \ln\left[p(g_k^{\text{dp}}|g_k = g)p(g_k = g|\lambda)\right] \cdot p(g_k = g|g_k^{\text{dp}}; \lambda), \tag{8}$$

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.

C.3 Variance Estimation

We denote the final EM estimate from Section C.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

[2] 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):

$$\underbrace{-\ln L''(\lambda; g^{\text{dp}})}_{\text{Observed information}} = \underbrace{E[-\ln L''(\lambda; g, g^{\text{dp}})]}_{\text{Complete information}} - \underbrace{E[-\ln f''(g|g^{\text{dp}}; \lambda)]}_{\text{Missing information}}, \tag{9}$$

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.

C.4 Fast Approximation for Randomized Response

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

$$\begin{aligned} E[v_k] &= E[E(g_k^{dp} - g_k \mid g_k)] = E[(2p-2)g_k + (1-\pi)n] \\ &= (2\pi - 2)\text{exp}(x_k'\lambda) + (1-\pi)n. \end{aligned} \tag{10}$$

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 10: $v_k|x_k, \lambda \sim \text{Pois}((2p-2)\text{exp}(x_k'\lambda) + (1-p)n$.

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

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{k=1}^{K} -(2p-1)\text{exp}(x_k'\lambda) + g_k^{dp} \ln\left[(2p-1)\text{exp}(x_k'\lambda) + (1-p)n\right]. \tag{11}$$

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.

Appendix D Connections between FIML and LLM

The connection between FIML and LLM is easiest to see in the FIML likelihood function in Equation 11 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 1 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 1, 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.

Appendix E Further Details on Abortion Attitudes Study

To illustrate how differential privacy influences the data from the abortion attitudes replication we conduct in Section 5.2, we show how the distribution of counts across the one-hot vector, g, changes as the level of privacy imposed, ϵ, increases.

The variables we used include vote (with 4 categories, including vote no, vote yes, no vote, and do not know), Age (in 6 groups), Sex (2 categories), Education (6 categories), and Party Identification (4 categories). This leaves a one-hot vector with $4 \times 6 \times 2 \times 6 \times 4 =$

$1,152$ elements. Then in the left panel of Figure 1, we plot the distribution of counts associated with each answer combination observed in the raw data with no noise. Only around half of answer combinations are actually observed, as evidenced by the large spike at 0. Few answer combinations are observed more than 5 times.

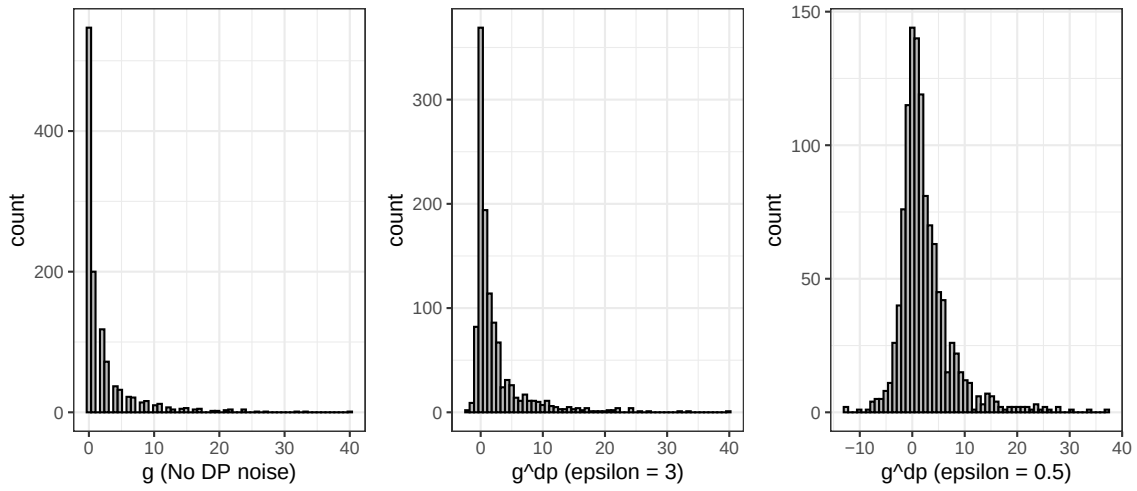

Figure 1: How Differentially Private Noise Affects the Distribution of Counts from the One-Hot Vector (Note: $S^2 = 2 * (1/\epsilon^2)$ and so S^2 is 8 when $\epsilon = 0.5$, and 0.22 when $\epsilon = 3$)

In the middle and right panels, we plot the (now noisy) counts at two different privacy levels. When privacy protection (middle panel, with $\epsilon = 3$), the shape of the distribution is visually similar to the noiseless distribution on the left. With the higher degree of privacy protection we use in our application (right panel, $\epsilon = 0.5$), the distribution is more distorted: it is closer to symmetric and a now substantial proportion is below 0 due to the noise. (To be clear, this graph does not indicate that the *counts* are below zero; only the counts plus noise are, which of course is part of what must be corrected.)

Appendix F Privacy Budgeting

The more information we elicit from any one research subject, the easier it would be to re-identify that person from a dataset, and so the more of the privacy budget we must spend to prevent it. It therefore makes good sense to limit the information we collect to that which we actually need for our analyses and eventual publication. This then leads

us to ask what in fact we need for any one research project. This section elaborates and expands on Section 6.

F.1 Principles

The strategy of limiting data collection to essential information is consistent with ethical principles of data collection, so we limit burdens to our research subjects. It is also consistent with commonly used social science research procedures applied before data is collected — including power analyses, experimental design, survey research sampling procedures, survey instrument design, and preregistration. These venerable procedures work well only when the information we need for our analyses are in fact accurately known ex ante.

However, if we get wrong what we need, and do not collect it, we will obviously miss it. This of course is the nature of the difficult choices researchers make every day in data collection, especially for expensive data collection projects based on sample surveys and large randomized experiments. For example, it would be best to guarantee representativeness (of the sample compared to the population) by careful choice and implementation of our random selection procedures; and similarly in experiments we should ensure statistical balance (between the treated and control groups) by random assignment procedures. Collecting variables to verify these would be wasteful if our procedures are trustworthy, but adding some may be useful just in case something goes wrong. Since this trade off is well known, researchers are commonly drawn to limiting information and also the opposite goal of including extra variables we do not have immediate plans for, so that we can conduct multiple tests of our hypotheses, tests of new hypotheses, purely exploratory analyses, verification of our selection procedures and, in experiments, randomized designs.

The general nature of the trade off we make in differential privacy is thus the same as without privatizing procedures, although we now have additional motivation to do whatever we can ex ante.

F.2 Practices

Ways of limiting information collected for any one individual include eliminating survey questions, asking for only coarsened answers (age group rather than age, degree rather than years of education, etc.; Iacus, King, and Porro 2012), limiting questions to only relevant subsets of respondents, collecting information about an index rather than individual variables making up the index when feasible (such as eliciting the number of times a respondent read a newspaper in the last week, rather than asking seven binary questions about each day of the week), or even applying randomized response when feasible (see Section 3.1).

We do not list in the previous paragraph the commonly used approach of "removing personal identifiers" because, whether a researcher considers a variable to be a personal identifier, is immaterial; the question is whether any variable or combination of variables can in principle identify an individual, regardless of the motive of the researcher in collecting the variable in the first place. Although we know a phone number can identify an individual or household, in many cases a combination of other variables intended for unrelated purposes can be as effective for identifying an individual. The resulting privacy violation would be the same as well. Differential privacy protects respondents regardless of whether a variable is designated as a personal identifier.

In other words, once differential privacy is applied any data (that is to the right of the chosen point among the five choices in our Figure 1), we can be confident the data is secure for *any* threat model – that is, regardless of the motives of and information available to any potential attacker. This means that many intuitive approaches to protecting privacy which may work in partial ways, for some attackers, with some motives, or with certain types of information, does not change the epsilon bound, as it only measures the privacy leakage that is *possible*. If we are optimistic, they may help, but the recent history of data privacy protection suggests a more careful stance.

The same logic also applies to information about respondents but not elicited from them, such as available from (1) background information (such as location), (2) metadata (time, place, or condition of interview), (3) variables created during the interview such

as for randomized assignment in experiments, or (4) derived variables created from 1-3, perhaps in combination with elicited survey responses (such as survey weights). Some of this information seems less sensitive and indeed, for some purposes or some attackers, it is. However, differential privacy is a conservative standard that protects against *all* possible attackers, even if we fail to think of their motives ex ante. For example, the state of residence of respondents seems like such an aggregated quantity as to be innocuous, and revealing whether the researcher randomly assigned some respondent to the treatment or control group seems unlikely to lead to privacy violations since it is by definition unrelated to all other variables. However, if these variables can be combined with some external information, they may make it possible for an attacker to identify a respondent in a different database. Ruling out any one attack because we conclude no one would try it is unsafe. Differential privacy makes it unnecessary.

Finally, it is worth asking when our procedures could be avoided altogether. Consider several situations. First, if the dataset a researcher is seeking is already available publicly, and a determined attacker could merely search the web to find it, privatizing the data might be unnecessary. A counterpoint to this position is Roberts (2018) who shows that secrecy is not a binary concept, where something is private or not private; she instead demonstrates that degrees of privacy violations can be crucial. For example, suppose a research subject has a criminal conviction that is publicly available only by going to the basement of a small town hall in rural Kansas, but a researcher makes that available in a searchable database on the web; even if legal, the respondent may well suffer from publication of this research without privatization.

Second, "data mission creep" is an ongoing problem where data collected for one purpose, and agreed to by the respondent, is used for a purpose not previously envisioned. This is a major issue with corporate data collection, but it is also a concern for creative social scientists who do this routinely. If we know the purpose to which the data will be put, we may opt for fewer privacy protections, but it is not always possible to limit unexpected uses. Differential privacy protects against all possible uses and all possible threat models.

Third, maybe we should not be worried for certain topics of study. Obviously, private information about sexual behavior, national security secrets, unobserved criminal behavior, and many other topics of legitimate study by academics needs to be protected. But where is the line? The lesson from the literature is that outguessing possible attackers based on topic is risky. For example, is it safe to assume that no attacker would be interested in a dataset used to study the well known effect of partisan identification on the vote? Possibly, but would a public official who happened to be in the survey be harmed if she reported voting for a member of the opposition party? Suppose income is elicited in this study as control variable; could a government tax authority, or the respondent's ex spouse, use that information to hurt the respondent as a consequence of our research? Remarkably differential privacy can protect respondents in situations when we are not able to guess the motives of possible attackers.

As scholars, we hope that the answers to when this technology should be used will depend, not only on the level of privacy we can offer respondents, but also on the remarkable public good and knowledge that can come from social science research. To help ensure the latter, we social scientists must take it as our responsibility to guarantee the former. At the same time, we can publicly evaluate the claims of data providers to be making useful data available for public good by reporting the "proportionate loss in effective sample size", L (See Section 5.1). This quantity can be computed and reported without additional privacy budget expenditure, every time our methods are used to analyze differentially private data. The value of L depends on the variance of the underlying estimator (without DP) and how much noise inflates the variance of that estimator. As a result, it will not necessarily decrease monotonically in the sample size for a fixed privacy budget, which equates to fixed variability in the counts. (To see this, take the extreme cases of a very small sample size and a very large sample size. Both can produce a high proportionate loss, but for differing reasons. With a small sample size, the variance without DP will very high, but the variance in the counts from DP will also destroy a large amount of information, thus yielding a relatively high loss in effective sample size. With a very large sample, there is effectively no sampling uncertainty without DP, and so the DP noise will

constitute a large degree of the remaining uncertainty, again yielding a relatively high loss in effective sample size.) As such, providing estimates of L for all empirical analyses is recommended.

At the end of the day, when privatization technology should be used — or choosing one of the points in Figure 1 — is a policy question that could be decided by self-governing academic associations, by companies or other organizations who we are asking for data, by universities where we work, by IRBs in our organizations or those of data providers, or by governments around the world that regulate our research. In other words, the answer to this question is inherently political, and is thus an excellent topic for political science research in its own right. We hope the scholarly community helps guide us through this changing landscape.

Appendix G Understanding and Setting ϵ

How the privacy loss varies with epsilon depends on both the data (e.g., the uniqueness of responses), and the background information an adversary has. Differential privacy protects against the worst case scenario: the most extreme points that could appear in the data, and an adversary with an arbitrary amount of outside information. If the worst case scenario does not apply, then the privacy protection provided by a fixed epsilon in practice will be relatively higher. We present a simple scenario that demonstrates this logic in intuitive terms.

Suppose an adversary knows everything about a person (including that they were in the survey, and that they were uniquely identified in the population by these variables). The adversary would like to know the respondents' answer to a binary question (e.g., how they voted in the referendum). The adversary therefore only has to pay attention to two elements of the one-hot encoding: those that match the background information with either a yes or a no vote. There are therefore 4 response patterns the adversary could observe in these two elements: $\{(1,1),(0,0),(1,0),(0,1)\}$. Only the latter two could be the truth. Suppose, without loss of generality, that the truth is (1, 0) and so we consider the highest privacy loss scenario being that which produces (1, 0), i.e., neither element is

flipped. Under RR, this happens with probability (1 - $1/(1+\exp(\epsilon))^2$. In other words, in the worst case scenario, the likelihood of the truth being (1, 0), given the observed DP response pattern (1, 0) is $(1-1/(1+\exp(\epsilon))^2$. This equates to a likelihood of 0.39 when epsilon = 0.5, and 0.9 when epsilon = 3.

Above we described a worst case scenario where the realized noise was negligible and the adversary had an extreme amount of background information. If the adversary knew less, then the set of relevant cells in the histogram would be larger, and as a result the actual privacy protection would necessarily be better even for the same epsilon.

This is a common situation: differential privacy protects respondents in the face of *any* threat model. The differential privacy bound is thus the *maximum* possible privacy leakage regardless of how much external information an attacker may have and regardless of their motives, computational power, or skills. In contrast, the probability of any one real threat actor that actually exists and actually violates someone's privacy would typically be considerably less than this bound.

Another way to think about this is that the privacy protection produced by a fixed epsilon will vary by context, including the size of the data, the presence of uniquely identifiable individuals in the data, and the background knowledge an adversary may reasonably be expected to possess. As such, there is not one value of epsilon that will be appropriate for all surveys. For most applications, we therefore recommend a simulation study based (if possible) on similar publicly available data, to guide the choice of epsilon. The simulation should involve assuming reasonable background information on the most unique respondent and assessing the extent to which the DP noise obscures and limits the possibility of learning more with high probability. (Of course, any amount of differential privacy prevents an attacker from learning more about the respondent with certainty.) The simulation study can also be used to assess the likely efficiency loss to DP, which should help researchers balance the trade-off between privacy and utility.

Although a simulation of this nature is the best approach, the choice of epsilon can be further guided by precedent and experience. In general, an epsilon below 1 is almost always considered reasonably safe. Nevertheless many practical applications of DP use

values much higher than this. It is also worth emphasizing that differentially private noise decreases exponentially in epsilon. As such, we advise caution in using an epsilon as high as in double digits, since this may entail very little noise in some circumstances.

References

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

Awan, Jordan and Zhanrui Cai (2020): "One Step to Efficient Synthetic Data". In: *arXiv preprint arXiv:2006.02397*.

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

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

Evans, Georgina and Gary King (Forthcoming, 2021): "Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset". In: *Political Analysis*.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2012): "Causal Inference Without Balance Checking: Coarsened Exact Matching". In: *Political Analysis*, no. 1, vol. 20, pp. 1–24. URL: j.mp/woCheck.

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, pp. 762–784.

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.

Roberts, Margaret E (2018): *Censored: distraction and diversion inside China's Great Firewall*. Princeton University Press.