Differential Privacy for Social Science Inference*

Vito D'Orazio† James Honaker‡ Gary King§

Prepared for presentation at the Annual Meetings
of the American Political Science Association,
September 4, 2015.

Abstract

Social scientists often want to analyze data that contains sensitive personal information that must remain private. However, common techniques for data sharing that attempt to preserve privacy either bring great privacy risks or great loss of information. A long literature has shown that anonymization techniques for data releases are generally open to reidentification attacks. Aggregated information can reduce but not prevent this risk, while also reducing the utility of the data to researchers. Even publishing statistical estimates without releasing the data cannot guarantee that no sensitive personal information has been leaked. Differential Privacy, deriving from roots in cryptography, is one formal, mathematical conception of privacy preservation. It brings provable guarantees that any reported result does not reveal information about any one single individual. In this paper we detail the construction of a secure curator interface, by which researchers can have access to privatized statistical results from their queries without gaining any access to the underlying raw data. We introduce differential privacy and the construction of differentially private summary statistics. We then present new algorithms for releasing differentially private estimates of causal effects and the generation of differentially private covariance matrices from which any least squares regression may be estimated. We demonstrate the application of these methods through our curator interface.

*For discussions and comments we thank Natalie Carvalho, Vishesh Karwa, James Lo, Jack Murtagh, Kobbi Nissim, Or Sheffet, Adam Smith, Salil Vadhan, Teppei Yamamoto, and numerous other members of the "Privacy Tools for Sharing Research Data" project http://privacytools.seas.harvard.edu. This work was supported by the NSF (CNS-1237235), the Alfred P. Sloan Foundation and a Google gift.

†Assistant Professor in the School of Economic, Political, and Policy Sciences at the University of Texas-Dallas.

‡(james@hona.kr, http://hona.kr)

§Albert J. Weatherhead III University Professor, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (king@harvard.edu, http://GaryKing.org)

1 Introduction

Social scientists often want to analyze data that contains information that must remain private. This private information might cause economic harm to subjects if leaked publicly, as in data that contains medical information, histories of drug use, or illicit behavior. Or it might result in shame to subjects, if they have revealed to the researcher past victimization or unpopular political opinions. Or even if no shame or harm results, violating the trust of subjects by making confidential information public may make them less likely to continue in future studies. In addition to ethical, disciplinary and professional norms, institutional review boards may impose additional restrictions on researchers to safeguard against legal liability. The increasing ability of big data collections, sensor data, and social media data to measure individual behavior in nuance, ensures that such privacy concerns will only increase. For all these reasons, researchers who generate new sources of data are responsible for using and releasing the data in ways that do not compromise the privacy of the subjects.

In tension with these strong ethical reasons to preserve privacy are the strong scientific imperatives to make data increasingly open and available. Science relies on the open examination of previous findings [1]. Data sharing and data archiving are often requirements of the funding sources to collect the data [2] and the publication venues of the generated research [3]. Studies that release their data are more likely themselves to be cited [4]. In the words of Crosas et al. [5],

> "Accessible and reusable data are fundamental to science in order to continuously validate and build upon previous research. Progressive expansive scientific advance rests upon access to data accompanied with sufficient information for reproducible results, a scientific ethic to maximize the utility of data to the research community, and a foundational norm that scientific communication is built on attribution."

This tension leads researchers to explore methods that they believe allow for data distribution, while maintaining the privacy of research subjects.

However, common techniques for data sharing that attempt to preserve privacy either bring great privacy risks or great loss of information. A long literature has shown that anonymization techniques for data releases are generally open to reidentification attacks[6, 7, 8]. Aggregated information can reduce but not prevent this risk, while also reducing the utility of the data to researchers [9, 10, 11]. Even publishing statistical estimates without releasing the data cannot guarantee that no sensitive personal information has been leaked[11].

At the other end of sharing, researchers who are interested in analyzing sensitive data collected by others often require a great degree of effort before getting any exploratory access to the data. They often require approval from an Institutional Review Board or travel to a secure location, or construction of non-networked environments for data storage. Since there is often no *ex ante* statistical information available for sensitive data, the requester may not know the extent to which the data will be useful for addressing her research question, or

whether the data even contain the correct information to answer the question of interest or test the proposed hypothesis. Thus, she must make a relatively uninformed decision to file for access, with a significant chance that upon access the effort will prove fruitless. The difficult balancing act between privacy and open science on the part of data collectors, and the significant burden on data consumers of even exploratory access to find the correct private data sources, are problems that any capable framework for privacy preservation should alleviate.

Differential Privacy, deriving from roots in cryptography, is one formal, mathematical conception of privacy preservation. It guarantees that any reported result does not reveal information about any one single individual [12, 13]. That is, the distribution of answers one would get with differentially private (DP) algorithms from a dataset that does not include myself must be indistinguishable from the distribution of answers where I have added my own information. These algorithms inject a precisely calculated quantity of noise to any statistical query to mask the possible contribution of any one individual to the result. It is provable that no possible combination of queries or model results can tease out the information of any individual.

While the theoretical literature over the last decade has developed DP algorithms for many different tasks, practical implementations are few and narrowly tailored to specific use cases. Moreover, while all DP algorithms are proven not to leak personal information, few results exist about their utility to provide useful results to researchers, or their effects on bias and efficiency of future inferences.

We present new work to make the theoretical advances of differential privacy practical to social science researchers, and to bridge the mathematical results into broadly useful tools for data sharing.

We make three contributions. First, we tailor the mechanisms of differential privacy to the release of causal estimates, particularly difference of means estimators, their standard errors and confidence intervals, and the consequences of pairwise matching. Second, we present the development of a new algorithm for the generation of differentially private covariance matrices. Since any least squares regression may be estimated from a dataset's covariance matrix, this is particularly useful for statistical inference and for more accurate statistical exploration of datasets containing private information. Third, we detail the construction of a secure curator interface, by which researchers can have access to statistical results from their queries, without gaining any access to the underlying raw data. We demonstrate the application of these methods through our curator interface.

2 Differential Privacy

Private or sensitive data may be loosely defined as data that, if known, can cause harm to an individual. Researchers who collect data on human subjects are often collecting sensitive data and require approval from their Institutional Review Board (IRB) to conduct their research. Once a dataset has been labeled by an IRB as containing sensitive information, it is the responsibility of the researcher to ensure data privacy. That is, the researcher must not take actions the compromise the privacy of his or her research subjects.

In the social sciences, a large and growing number of datasets contain sensitive information. The state of practice for managing data privacy concerns has been to publish accurate statistical estimates and then either (a) make the data inaccessible to others (data jailing) or (b) strip the personally identifiable information (PII) and release the data. Neither approach is both conducive to good scientific practice and sufficiently responsible. Additionally, in both cases aggregate statistical estimates are reported with perfect accuracy, a practice that is not sufficient to guarantee data privacy.

When working with sensitive data, individual-level statistical information is generally avoided, and accurate, aggregate statistical information is generally published. While this will preserve privacy in many cases, it does not guarantee privacy preservation in all cases [9, 10, 11]. In short, every time statistical estimates are released, we increase the likelihood of learning a piece of sensitive information. As results accumulate over time, that likelihood increases. The US Census and Eurostat, the European Union's official statistical agency, only publish select aggregate statistics for precisely this reason.

The risk of releasing sensitive information through the publication of aggregate statistics is not negligible, but it is minimal. In many cases, especially when the data are only mildly sensitive, we might be willing to accept some risk and report the true estimates. More detrimental to scientific progress is the practice of data jailing. Restricting access to data severely limits what may be learned from future analysis. Researchers interested in using the data may find the costs for accessing it prohibitively expensive or the paperwork to be exceedingly time-consuming. It also makes scientific replication difficult, if not impossible. It is a hindrance to scientific progress.

Rather than restricting access to the sensitive data, it is more common for researchers to anonymize a dataset by stripping all PII, and then to release the supposedly sanitized data. Examples of PII include the subject's name, address, or social security number. However, as shown in [6, 7, 8], it is often possible to re-identify individuals through linkage attacks. A linkage attack is when a dataset that contains sensitive information and has had all PII stripped is successfully linked to another dataset that does not contain sensitive information and has not had its PII stripped. For example, one study has found that "87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}" [14, 1].

Due to the potential for reidentification, the Health Insurance Portability and Accountability Act specifies 18 identifiers of Personal Health Information that are required to be

stripped from a dataset before release.[1] However, stripping the 18 HIPAA identifiers provides no provable guarantee of privacy. Furthermore, as more sensor and other individual-level data becomes available, it is increasingly likely that arbitrary attributes can and will be used to identify research subjects–even when all PII is redacted.

An alternative solution to data jailing and de-identifying is to allow statistical queries through a secure curator interface. In such a system, researchers may query the data but never actually have access to the raw data. From the researcher's perspective, this can be equally as advantageous as releasing de-identified data since those who wish to analyze the data may still do so. Statistical analyses are replicable. However, the risk to data privacy remains [15]. Differencing attacks and restructuring attacks are examples of ways in which an adversary may acquire knowledge that should remain private [11]. In short, any system for releasing statistical estimates simply cannot respond with perfect accuracy to any and all queries and still guarantee that privacy is being preserved. Rather, such systems either audit the queries, perturb the data, or perturb the statistical estimate [11]. The safest approach, and the one we further explore and expand, to to perturb the estimate using the guarantee of differential privacy.

Differential privacy is a recently developed notion of privacy preservation that guarantees that no individual's privacy is compromised by the released information. The general idea is to add noise that is proportional to the sensitivity of the true estimate [12]. The *sensitivity* of an estimate is the theoretical range of values that can be observed in neighboring datasets. *Neighboring datasets* are defined as two datasets that differ by *at most* one row. By guaranteeing differential privacy, we guarantee that the perturbed estimate may be observed regardless of the presence or absence of any individual in the data.

2.1 Definition

Consider a mechanism, M, that gets potentially sensitive data about individuals as input and performs some computation over the data. Our notion of computation is very broad and includes any procedure for transforming data into some output. Examples range from the calculation of summary statistics, to regression estimates, to an application of statistical disclosure limitation technique (such as de-identification) aimed at producing a version of the data that is considered safe to share or disclose.

Consider a function, T, whose output is an event, b, from a discrete set of events, B. Each event in B has some probability associated with it, and the probabilities sum to 1. Below, T is discussed in the case where B is simply yes or no, or 0 or 1, but this may be easily generalized. The input to T might include the output of the differentially private mechanism, M, and/or any auxiliary information. We need to know neither the auxiliary information nor how T determines which event to return. Differential privacy provides a guarantee on what can be learned by incorporating the output of M into the input of T. For example, T might be whether a person possesses a trait, Pr[T] is our belief about whether a

[1] For a complete list of identifiers, see the Health Insurance Portability and Accountability Act or visit http://privacyruleandresearch.nih.gov.

person possesses a trait, and M might reveal how many people in the data have that trait.

Given neighboring datasets A and A', the definition of a differentially private mechanism provides the guarantee that:

$$Pr[T(M(A)) = 1] \leq e^{\epsilon} Pr[T(M(A')) = 1] + \delta, \qquad \forall\ T. \tag{1}$$

By *neighboring* we mean precisely that datasets A and A', differ by *at most* one observation. This may be *any* theoretically possible observation. For example, observation i in A' might be redefined to be the minimum (or maximum) theoretical value for each variable. Thus, by masking the contribution of *any* single individual to the output of M, differential privacy guarantees that the results from $M(A)$ and $M(A')$ are nearly indistinguishable.

Differential privacy provides us with a quantifiable "shaded" measure of privacy; epsilon and delta quantify "privacy loss" and can be mathematically related to the excess risk to an individual that results from her data being used (lower values of epsilon and delta guarantee lower risk). Of the two parameters, delta controls the probability that a bad privacy breach event would happen, and should hence be kept so small (e.g., one in a billion) that we will neglect it in our discussion below. The other parameter, epsilon, controls the "allowed" privacy risk. Indirectly, epsilon also controls the accuracy to which the differentially private computation can be performed – lower privacy risk comes with lower accuracy – and hence epsilon cannot be chosen to be negligibly small as delta is. Rather, epsilon should be chosen so as to allow a reasonable compromise between privacy and accuracy. In general, epsilon should be thought of as a small number, say, between .001 and 1.

Epsilon controls the effect of each individual's information on the outcome in the following sense: differential privacy requires that, whether the mechanism is applied to the data with or without John's information included, the probability that $T = 1$ can change by a factor of at most 1+epsilon = 1.01, i.e., each probability changes by at most 1% (this analysis is approximate). To see what this property can be shown to imply, consider John who is worried about the social consequences he would face if his political affiliation became known. In this example, $T = 1$ means John's political affiliation becomes known. Say the probability that John's political affiliation becomes known is .05 ($PR[T = 1] = .05$). Then, if we set epsilon to be .01, differential privacy guarantees that the probability of John's political affiliation becoming known would not increase to more than 5.05 as a result of learning the output of the differentially private mechanism, M. This is true regardless of whether John's information is in the data.

A bound on the change in the probability of T, a function which we know nothing about, is quite counterintuitive. However, it is a very strong guarantee and much of the appeal of differential privacy is a result of this guarantee. If we think about $M(A)$ and $M(A')$, the calibrated noise that has been added the output in both situations ensures that the distribution of answers is nearly identical.

In the following, X_i is whether or not individual i has attribute X. T is the function that an adversary would use to determine if $X_i = 1$. T tells us whether $X_i = 1$. Pr[T] is our belief about whether $X_i = 1$. An application of Bayes' Rule provides the following result:

$$
\begin{aligned}
Pr[X_i = 1|T(M(A)) = y] &= \frac{Pr[T(M(A)) = y|X_i = 1]Pr[X_i = 1]}{Pr[T(M(A)) = y|X_i = 1]Pr[X_i = 1] + Pr[T(M(A)) = y|X_i = 0]Pr[X_i = 0]} \\
&= \frac{Pr[T(M(A)) = y|X_i = 1]Pr[X_i = 1]}{Pr[T(M(A)) = y|X_i = 1](Pr[X_i = 1] + \frac{Pr[T(M(A))=y|X_i=0]Pr[X_i=0]}{Pr[T(M(A))=y|X_i=1]})} \\
&= \frac{Pr[X_i = 1]}{Pr[X_i = 1] + Pr[X_i = 0]\frac{Pr[T(M(A))=y|X_i=0]}{Pr[T(M(A))=y|X_i=1]}} \\
&= \frac{Pr[X_i = 1]}{Pr[X_i = 1] + Pr[X_i = 0]e^{\pm\epsilon}} \\
&\leq \frac{Pr[X_i = 1]}{Pr[X_i = 1] + (1 - Pr[X_i = 1])e^{-\epsilon}}
\end{aligned}
\tag{2}
$$

We are comparing two worlds, one in which we have no information and one in which we have differentially private information. Differential privacy guarantees that $Pr[T = 1]$, considered our *prior*, can change by at most $\frac{100*prior}{prior+e^{-\epsilon}(100-prior)}$.

Table 1: Interpreting Epsilon

$Pr[T = 1]$	epsilon					
	.01	.05	.1	.2	.5	1
1	1.01	1.05	1.1	1.22	1.64	2.67
5	5.05	5.24	5.5	6.04	7.98	12.52
10	10.09	10.46	10.94	11.95	15.48	23.2
25	25.19	25.95	26.92	28.93	35.47	47.54
50	50.25	51.25	52.5	54.98	62.25	73.11
75	75.19	75.93	76.83	78.56	83.18	89.08
90	90.09	90.44	90.86	91.66	93.69	96.07
95	95.05	95.23	95.45	95.87	96.91	98.1
99	99.01	99.05	99.09	99.18	99.39	99.63
	Upper bound on $Pr[T = 1|T(M(A))]$					

Note: Cell values, excluding epsilon values, are percentages. They are calculated as $\frac{100*prior}{prior+e^{-\epsilon}(100-prior)}$.

More generally, Table 1 shows the effect of different epsilon values on our belief that $T = 1$. The left column is our prior belief that $T = 1$. Each column to the right contains an upper bound on our updated belief having learned $M(A)$. For example, if there is a 99% chance of John's political affiliation being known, and then we learn $M(A)$ with an epsilon of 0.5, then our belief about John's political affiliation can become *at most* 99.39%.

The more computations John participates in, the higher the risk is for his privacy. Having a quantified measure of privacy is beneficial in understanding how this risk may accumulate,

and differential privacy provides us with a bound on how risk accumulates across multiple analyses. The exact analysis is beyond the scope of this document, but is known as *composition theorems*. As an example, suppose John participates in two analyses, each providing risk parameter $epsilon = 0.01$. Differential privacy then entails that his overall risk amounts to at most $2 * epsilon = 0.02$. We note that while differential privacy is not the only technique that quantifies risk, it is currently the only framework with quantifiable guarantees on the risk resulting from composition. For example, in k-anonymity one may perceive k as corresponding to risk, but one can demonstrate two k-anonymized datasets that in tandem result in complete revelation of information.

Any differentially private estimates may be used as input to any algorithm and still retain the privacy guarantee.

The conservation of epsilon across mechanisms amounts to the notion of a "privacy budget" where a dataset contains a global epsilon that may be dispersed among all mechanisms one wishes to compute with that dataset. The privacy budget is discussed in more detail in our discussion of the secure curator interface.

3 Privacy Preserving Summary Statistics

Much of the theory of differential privacy is grounded in counting queries on databases[2]. From this, differentially private summary statistics have been developed for a large number of algorithms, including common summary statistics such as the mean, median, mode and quantiles. Prior to turning to our own work on mechanisms to release differentially private causal and regression estimates, we illustrate one such mechanism for releasing a differentially private mean of a variable.

The Laplace distribution provides a commonly used mechanism for creating differentially private versions of simple, univariate, continuously valued statistics, and is a useful demonstration of a privacy preserving mechanism. Let us consider calculating a mean for N observations of a variable, X, in a private dataset. Assume we want to report a version of the mean of that variable that obeys the definition of differential privacy above. Differential privacy requires that no information about any individual can be leaked, so first we determine the *sensitivity* of the released statistic to the value of any one individual in the dataset. Any function of the data, $f(X)$, has a sensitivity, which we denote Δf.

If the data is bounded between x_{min} and x_{max}, then a single individual would decrease the mean the most if they are at the lower bound, and increase the mean the most if they are at the upper bound. The difference between the mean of X when our hypothetical individual is at the lower bound, and at the upper bound, is the largest possibly effect one individual can have on this value, and thus gives us the sensitivity of the statistic. Since any individual contributes x_i/N to the mean, then the sensitivity here is:

$$\Delta f = \frac{x_{max} - x_{min}}{N} \tag{3}$$

as depicted in figure 1. This is the absolute upper bound on how much a change in one individual's data can influence the statistic at hand. Intuitively, we want to guarantee that we do not reveal information about any one individual by adding noise sufficient to mask the largest possible contribution of any one individual.[3]

The Laplace is a convenient distribution to use for this noise, given the construction of the definition of differential privacy. The Laplace has distribution function:

$$f_{Laplace}(x|b,\mu) = \frac{1}{2b}\ \exp\Big(-\frac{|x-\mu|}{b}\Big) \tag{4}$$

with mean μ, and variance $2b^2$. The Laplace is a mirrored, and thus symmetric version of the exponential distribution. The exponential is common to survival and event history

[2]That is, computing the number of observations in a dataset that obey a predicate, such as having a particular combination of attributes.

[3]For example, if we knew that a variable had mean zero, except for John, whose value, x_J, we did not know, then the mean would reveal John's information. John would contribute x_J/N to the mean, and $x_J = \bar{x}N$. We want to add enough noise to the mean so that we no longer learn x_J even in this situation. However, we don't want to add so much noise that we can't learn that the rest of the population has mean close to zero, if we hadn't known that.

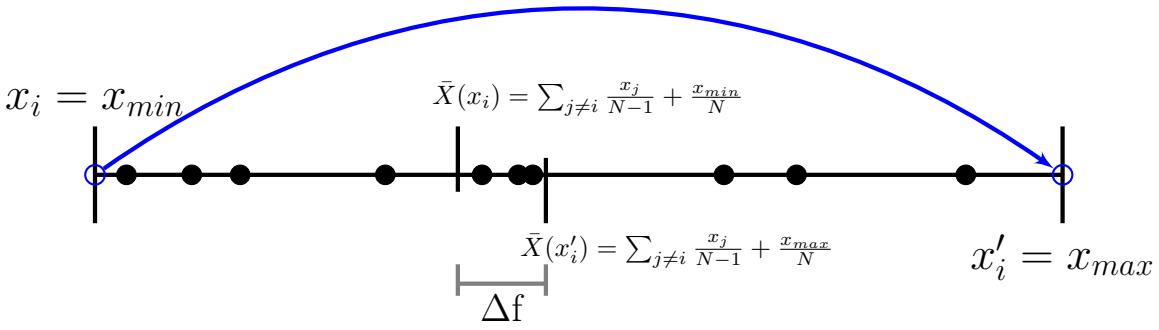

Figure 1: *Sensitivity of the mean for a bounded variable.*

models, which use its *memoryless* distribution[4], which we are also about to exploit. To make a continuous variable differentially private, we add a draw from a mean zero Laplace, with parameter b as:

$$b = \frac{\Delta f}{\epsilon} \tag{5}$$

So our differentially private mean, $M(X)$, which combines the "true" sample mean with Laplace noise, becomes:

$$M(X) = \bar{X} + Y; \qquad Y \sim f_{Laplace}(b = \Delta f/\epsilon, \mu = 0) \tag{6}$$

To check this mechanism meets the definition of differential privacy, consider some probability of any outcome, z. The ratio of this probability between two adjacent datasets, is given by:

$$\frac{pr[M(X) = z]}{pr[M(X') = z]} = \frac{e^{\frac{-\epsilon|\bar{X} - z|}{\Delta f}}}{e^{\frac{-\epsilon|\bar{X}' - z|}{\Delta f}}} = e^{\frac{\epsilon|\bar{X}' - z| - \epsilon|\bar{X} - z|}{\Delta f}} = e^{\frac{\epsilon|\bar{X}' - \bar{X}|}{\Delta f}} \leq e^{\epsilon} \tag{7}$$

the last step following since we know $\Delta f \geq |\bar{X}' - \bar{X}|$ by the definition of the sensitivity. It thus follows that $Pr[M(X) = z] \leq e^{\epsilon} Pr[M(X') = z]$, thus meeting the definition of ϵ-differential privacy (in this case, with parameter $\delta = 0$). For other continuously valued summary statistics, the same Laplace mechanism works for preserving privacy, however, the analytic form for the sensitivity, Δf, will change by statistic.

Dwork and Roth show a more general derivation of 7 which holds for any continuous function, including multidimensional functions. Gaussian distributions can be used in place of Laplace, for some small value of δ. As discussed in section 2.1, if a released statistic is differentially private, then any transformation or post-processing of that statistic is also privacy preserving; a generally useful construction then is to divide the range of a variable into 2^k equal partitions and create a perfect binary tree. Each of the 2^k leaves contains the

[4]The hazard function of an exponential waiting time, as the ratio of two exponentials, is a constant.

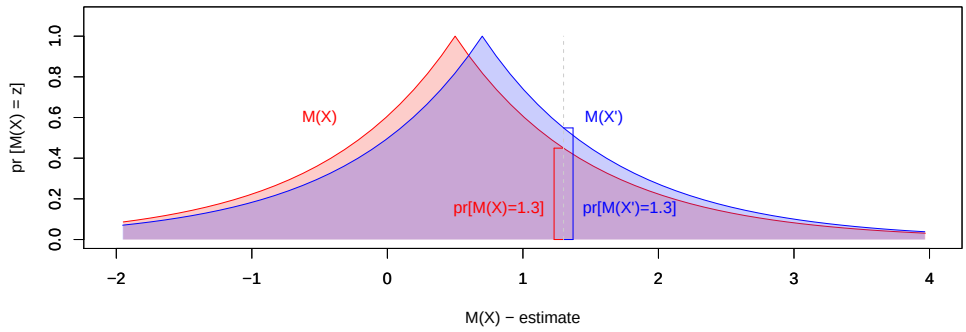

Figure 2: *Two Laplace distributions, for two adjacent datasets X and X'. The definition of ϵ-differential privacy requires the ratio of $M(X)/M(X')$ is not greater than e^{ϵ} for all points along the x-axis. Thus for any realized output z – for example here, $z = 1.3$ – we can not determine that X or X' were more likely to have produced z.*

count of the number of observations in that partition, and each node contains the sum of the two nodes (or leaves) directly below. All the nodes and leaves form a vector of length $2^{k+1} - 1$ which can itself be made differentially private by the Laplace or Gaussian mechanisms. From this tree, different combinations of nodes and leaves can give estimates to many forms of useful summary statistics including means, medians, modes, quantiles, as well as density and cumulative density graphs.

4 Causal Inference

4.1 Randomized Experiments

Assume we wish to learn whether some binary variable, t, has a *causal effect* on a continuous outcome, y. That is, any intervention that changes t would necessarily also result in a change in y. Hypothetically, every individual might have a distribution of outcomes $Y(1)_i$ if t_i is set to 1, and $Y(0)_i$ if t_i is 0. The causal effect we are interested in is the expected effect on Y of t, or $E[Y(1) - Y(0)]$. Unfortunately, for every individual, we only observe one outcome, $y(t = 1)$ or $y(t = 0)$.

If we have a population where individuals are randomly selected for treatment, then the causal effect of treatment can be estimated as the average outcome among those randomly selected for treatment compared to the average among those randomly assigned as control observations. This *difference of means* estimator, is robust to many distributions of Y and relationships to other causal factors. Importantly, while there may exist many other variables that also cause Y, we do not need to control for their effect in our estimation, as randomization allows that for any such auxiliary variable, their distribution among the treatment observations should be approximately equal to their distribution among the control observations. Thus their expected contributions to the two population means cancel.

4.2 Difference of Means

Assume each individual i experiences treatment $t_i \in \{0, 1\}$. The observed outcome after assignment to treatment, $y_i(t_i)$, we will abbreviate here as y_i. We will assume Y is bounded in a known range $y_{\min} \leq y_i \leq y_{\max}$. We can briefly describe summary statistics of two sub-populations, the treated and control observations. The count, mean and standard deviation are given by:

$$n_1 = \sum t_i \qquad n_0 = \sum 1 - t_i \tag{8}$$

$$\bar{y}_1 = \frac{\sum t_i y_i}{n_1} \qquad \bar{y}_0 = \frac{\sum (1 - t_i) y_i}{n_0} \tag{9}$$

$$sd(y_1) = \sqrt{\frac{\sum t_i (y_i - \bar{y}_1)^2}{n_1}} \qquad sd(y_0) = \sqrt{\frac{\sum (1 - t_i)(y_i - \bar{y}_0)^2}{n_0}} \tag{10}$$

In appendix A, we derive the sensitivity of the difference of means estimator (theorem A.1), as well as its standard error (lemma A.5). These are as follows:

Statistic	Formula	Sensitivity
Difference of Means	$\bar{y}_1 - \bar{y}_0$	$\frac{y_{\max} - y_{\min}}{n_1 + 1} + \frac{y_{\max} - y_{\min}}{n_0 + 1}$
Standard Error of DofM	$\sqrt{\frac{sd(y_1)^2}{n_1} + \frac{sd(y_0)^2}{n_0}}$	$\sqrt{\frac{N^* - 1}{N^{*3}}}\Big(y_{\max} - y_{\min}\Big)$ where $N^* = \min(n_0, n_1)$

We briefly provide an intuition about how these sensitivities are derived and show how they are incorporated into a mechanism for preserving privacy, before demonstrating some Monte Carlo experiments illustrating differentially private causal estimates.

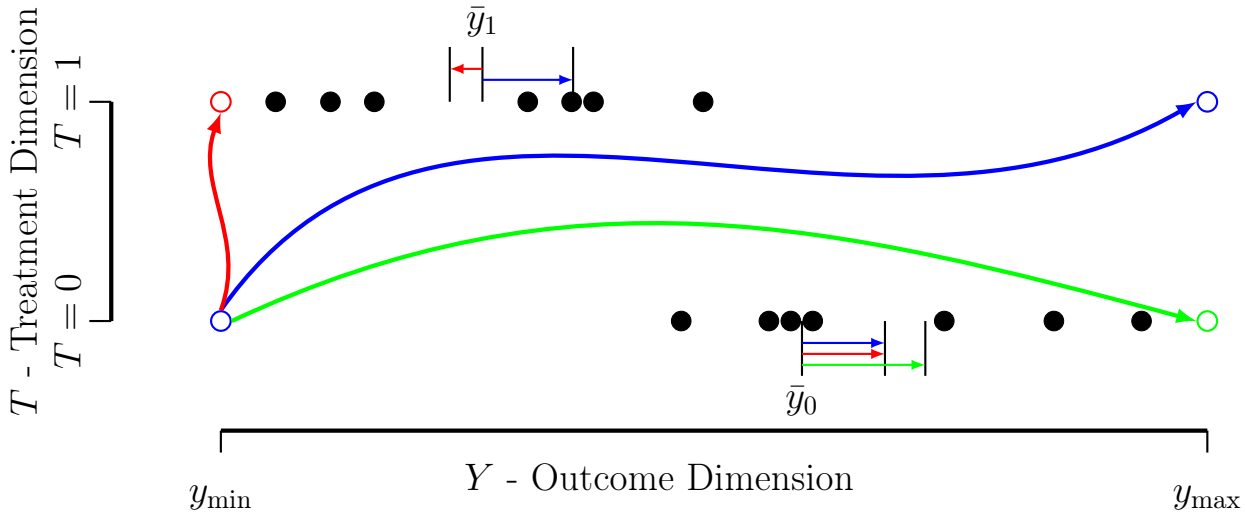

Figure 3: *Sensitivity of difference of means test. The red movement changes $\bar{y}_1$ less than the blue movement. However, the red movement changes the difference in means value by a greater quantity as it shifts in the opposite direction to the change in $\bar{y}_0$.*

As we have seen, sensitivity describes the largest possible change in a function's value that can result from arbitrarily changing any one individual's values in the dataset. In an experimental setting, changing one individual's data means we might change both their treatment t_i and their outcome y_i. In appendix A.2, we derive the sensitivity of the difference of means test, and we show the logic of this proof in figure 3. This figure shows a dataset of y on the horizontal and t on the vertical, with some example fixed data points in black. Now we allow one observation to arbitrarily move across the space of the data, and our goal is understand what possible movement creates the largest change in the difference of means test $\bar{y}_1 - \bar{y}_0$.

In the fixed data in this example, we have control data that generally take on high values of y, and observations under treatment that have low values of y; this suggests the causal effect of treatment is to lower the outcome y. The means of the control and treated values are denoted on the graph. Following section 3, the mean is most effected when we move an observation across the range of the data from one bound to the opposite bound. The green arrow represents movement of a control observation from the lower bound of y to the upper bound. The corresponding change that this creates on the mean of the control observations $\bar{y}_0$ (which we know is $(y_{\max} - y_{\min})/N_0$) is shown by the related green arrow under $\bar{y}_0$. This is the largest effect we can have on $\bar{y}_0$ from any arbitrary movement of one observation. However, our goal is to determine the largest effect on the $\bar{y}_1 - \bar{y}_0$.

The red and blue arrows represent moving this observation not only in y, but reassigning it from control to treatment. When the observation is moved out of control, the mean $\bar{y}_0$ still changes, because an outlier to the far left of the distribution has been removed; the blue and red lines above $\bar{y}_0$ measure how much the mean of the control changes. The blue arrow shows a movement of our observation to both treatment and the upper bound of y. This has a large change on the treatment mean, $\bar{y}_1$ because it adds an outlier. So this movement from $(y_i = y_{\min}, t_i = 0)$ to $(y_i = y_{\max}, t_i = 1)$ results in large changes in both the treatment and control means, because it adds and removes an outlier from each, respectively. However, the net effect on the difference of means estimate $\bar{y}_1 - \bar{y}_0$ is actually very small, because these movements are in the same direction and thus cancel. That is, the movement shifts both $\bar{y}_1$ and $\bar{y}_0$ in the same direction, and changes their difference very little. The larger effect in this data, is the movement denoted by the red arrow. Here we shift the observation in t but not y, and the corresponding changes in the means are in opposite directions, magnifying the effect on the difference in the means. In our proof, we show this is the form of movement that has the largest effect on the difference of means test, and it has the largest effect when the rest of the data is located at the opposite bound of y. This largest possible effect is the sensitivity.

4.3 Differentially Private Mechanism for Difference of Means

With sensitivity for the difference of means test derived, we can construct a provably differential private release of this statistic by using the same mechanism as in section 3. That is, we compute the difference of means in the private data, and then add Laplace noise to this value with standard deviation proportional to the sensitivity we have derived. Here that implies:

$$M(X) = \bar{y}_1 - \bar{y}_0 + Z; \quad Z \sim f_{Laplace}(b = \Delta f/\epsilon, \mu = 0); \quad \Delta f = \frac{x_{\max} - x_{\min}}{N_1 + 1} + \frac{x_{\max} - x_{\min}}{N_0 + 1} \tag{11}$$

For which we can state a very simple algorithm for generating the release:

Algorithm 1: Differentially Private Difference of Means Estimate

1. Calculate $\bar{y}_1 - \bar{y}_0$
2. Calculate $\Delta f = \frac{x_{\max} - x_{\min}}{N_1 + 1} + \frac{x_{\max} - x_{\min}}{N_0 + 1}$
3. Draw $Z \sim f_{Laplace}(\mu = 0, b = \Delta f/\epsilon)$
4. Release $M(X) = \bar{y}_1 - \bar{y}_0 + Z$

4.4 Monte Carlo Example

To demonstrate this privacy preserving mechanism, we simulate data to show the noise that results. We assume outcome Y is bounded and generated from a latent value as:

$$Y(t_i)^* = \beta_0 + \beta_1 * t_i + \nu; \qquad \nu \sim \mathcal{N}(0, 0.1) \tag{12}$$

$$t_i \in \{0, 1\}; \qquad \bar{t} = 0.5; \beta_0 = 0.2; \beta_1 = 0.6 \tag{13}$$

$$Y(t_i) = \begin{cases} 0 & Y(t_i)^* < 0 \\ Y(t_i)^* & 0 \leq Y(t_i)^* \leq 1 \\ 1 & Y(t_i)^* > 1 \end{cases} \tag{14}$$

where t is the experimental treatment. We simulate datasets of 2000 (1000 treated, 1000 control) observations and set $\epsilon = 0.5$. In the top-left of figure 4, in blue we show the distribution of difference of mean statistics estimated across 1000 simulated datasets. These are the values computed in the Monte Carlo datasets with the private data. The distribution comes from the sampling error of the finite sample. In red, is the distribution of differentially private versions of the difference of means, in these same datasets. Comparing the distributions, the differentially private estimates are still unbiased, while the extra noise added by differential privacy to these estimates increases the standard deviation by about 60 percent (.0071 to .0044). A sixty percent increase in the standard error is what we would have seen in the original estimator on the private data if the sample size had been reduced from 2000 to about 800 observations. So an interpretation of the utility cost of differential privacy in this example, is that it results in answers as noisy as an 800 observation dataset, or put another way, results in an effective sample size of 40 percent of the original dataset.

The center-top graph of figure 4 now shows these same two distributions plotted against each other. If there was no Laplace noise added to the differentially private values, then all the points would line up on the $y = x$ line (shown dashed in blue), and so the vertical distance of each point from this line is the draw from the Laplace that the estimate received in that simulation. The distribution of all these Laplace draws is shown as a histogram in the top-right graph, with the density from which they were drawn superimposed in red.

The second row of these plots show the same graphs but now for differentially private versions of the standard error of the difference of mean test. On the bottom right we can see that the variance of the Laplace noise that has been added is smaller (by about a factor of two) than for the difference of mean itself (the graphs are on the same scale). This is because the sensitivity is smaller. However, the center graph shows that even though we are adding less noise, the standard errors are far less useful. The variance of the estimated standard errors across the Monte Carlo samples is very small. Even though we are adding less Laplace noise than before, as a ratio to the underlying variance of the sampling distribution, this noise is overwhelming; The standard deviation of the differentially private standard error is 46 times the standard deviation of the sampling distribution of private values of the standard error. Moreover, the sample standard errors are close to zero, and when we construct differentially private versions with Laplace noise, many of these –here 6.2 percent– become negative. If we were using these standard errors in the denominator of a t-test we would have nonsensical

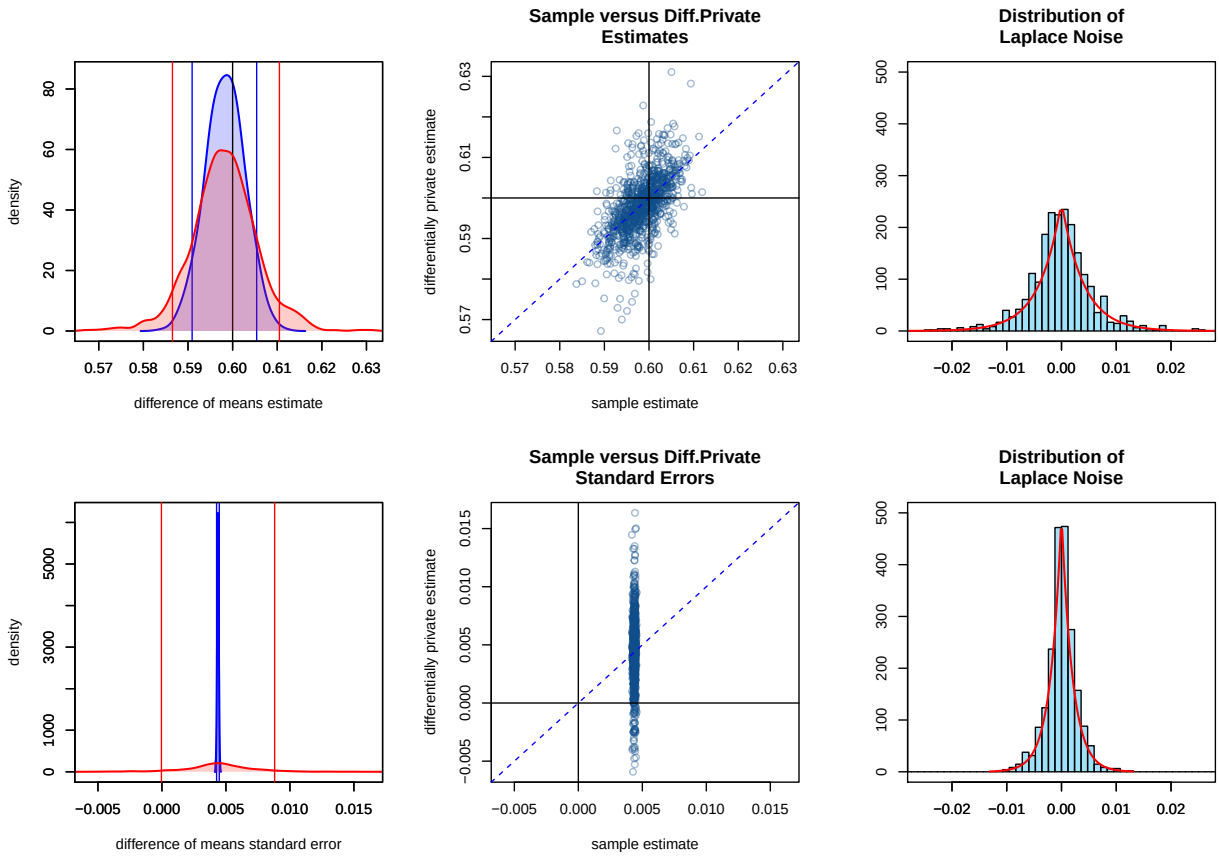

Figure 4: *Distributions of differentially private statistics of the difference of means estimate (top row) and standard error of the difference of means (bottom row).*

answers. Although it is central to many mechanisms for differential privacy, this simulation shows the Laplace is not a good distribution to use for privacy preservation of standard errors. We demonstrate a new mechanism for privacy-preserving standard errors in the next section.

4.5 Differentially Private Mechanism for Standard Errors

As we have seen, the Laplace mechanism creates usably accurate estimates of differences of means, but the same privacy-preserving mechanism does not create usable standard errors. The sampling distribution of the standard error is generally quite small. The sampling distribution of the standard deviation collapses to the population standard deviation at a rate of the order of $1/\sqrt{n}$, while the standard error itself collapses to zero at a rate $SD/\sqrt{n}$ or order of $1/n$. The sensitivity of the Laplace is order of $1/n$, so the noise of the Laplace mechanism is not converging to zero faster than the sampling distribution of the standard error. Thus we see the error of the Laplace mechanism dominate the standard error, while the sampling error dominated the calculation of the mean.

In this case, we want instead to turn to a new privacy mechanism that takes advantage of the fast convergence of the standard error. This property means that quite small subsamples from the dataset would themselves provide accurate estimates of the standard error. The *subsample and aggregate* mechanism [16] can take advantage of the fact that subsamples of the data provide accurate answers.

The subsample and aggregate algorithm [16] divides the dataset X into M subsets, $\{X_1, \ldots X_M\}$, of equal size. In each subset we compute a function $f_m = f(X_m)$. We then choose a method to aggregate the M values of the function into one answer $\bar{f}$. The key insight, is that each observation appears in only one subset, and thus can influence only one f_m. The advantage to exploit is that the original function f might have high sensitivity, but the sensitivity now is the sensitivity of the aggregation method that creates $\bar{f}$ rather than f itself. The difficulty to avoid is that most mechanism reduce noise as a function of the number of units and now $M \ll N$.

We use the Winzoring mean approach of Smith [17], which first computes a differentially private version of a bound that captures some fraction of all the M values of f. It then Winzorizes the values of f (that is, censors values beyond outside these bound to the limits of the bounds) and then averages the Winzorized values. Winzorized means are robust estimates of the means of Normally distributed random variables. As we have seen in section 3, computing a mean of N values over range R has sensitivity R/N. This approach pays off if across the small samples, the Winzorized range R is sufficiently small to compensate for the low value of N, which here is the number of subsamples the dataset can be divided into. (The algorithm, following Smith [17] is given in appendix B).

Figure 5 shows our previous Monte Carlo where instead the privacy-preserving standard errors are generated with the subsample and aggregate mechanism. The densities of the estimates from the private data, and the differentially private estimates closely match, although the differentially private standard errors are slightly attenuated. The top-center and

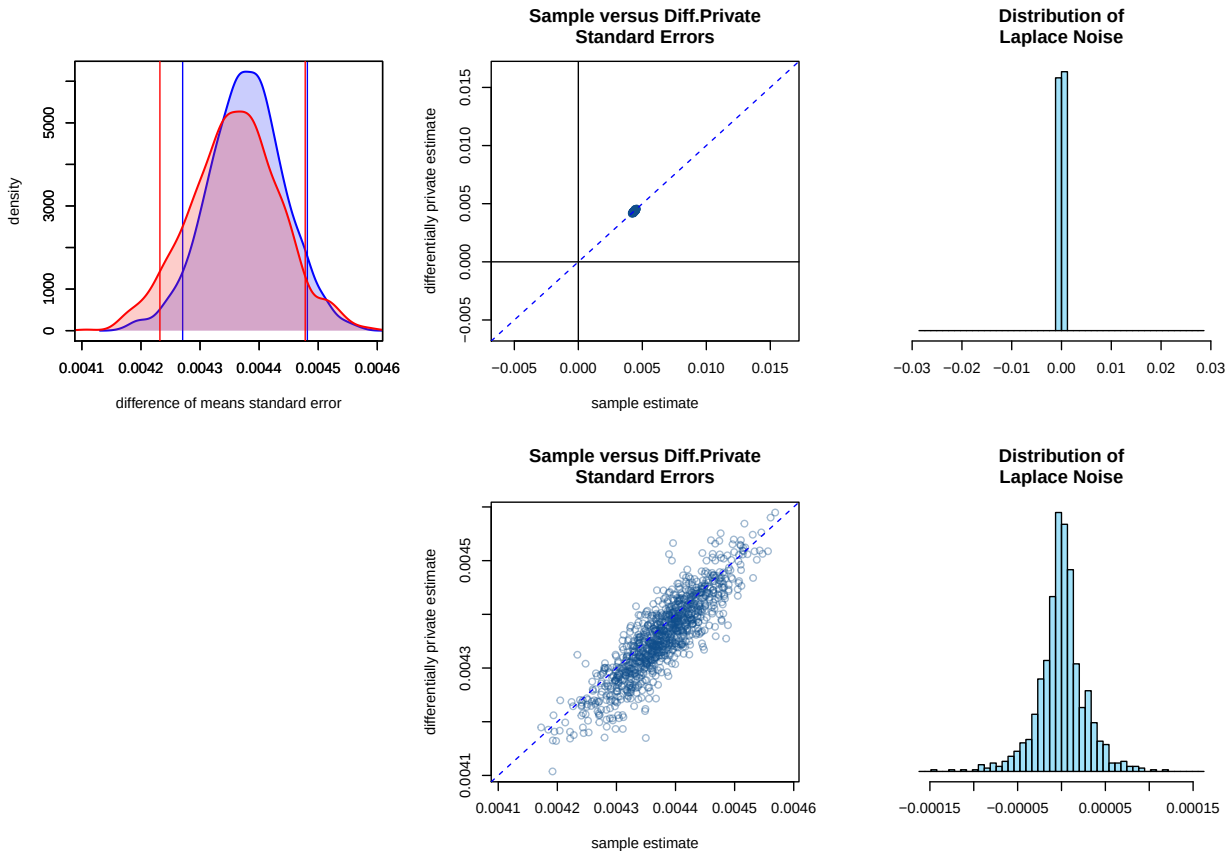

Figure 5: *Distributions of differentially private estimates of the standard error of the difference of means test, using the subsample and aggregate privacy mechanism. The top row shows the distribution of privacy preserving standard errors on the same scale as in the previous figure 4 and shows how dramatically the noise of the distribution has collapsed. The bottom row readjusts the axes so that the distribution can be seen as other than a spike.*

top-right graphs are on the same scale as in figure 4 and shows how dramatically the noise of the distribution has collapsed. On the previous scaling the differentially private values look like a spike. The ratio of the respective standard deviations of the distributions has shrunk from 42 to 1.2 times the standard deviation of the sampling distribtuion. The bottom row of this figure rescales the axes so as to see more clearly the distributions. The slight mass below the $y = x$ line in the center graph is another way to visualize the attenuation of the differentially private standard errors, while the distribution of noise in the bottom-right is now a mixture of Laplaces, that still roughly resembles a Laplace distribution.

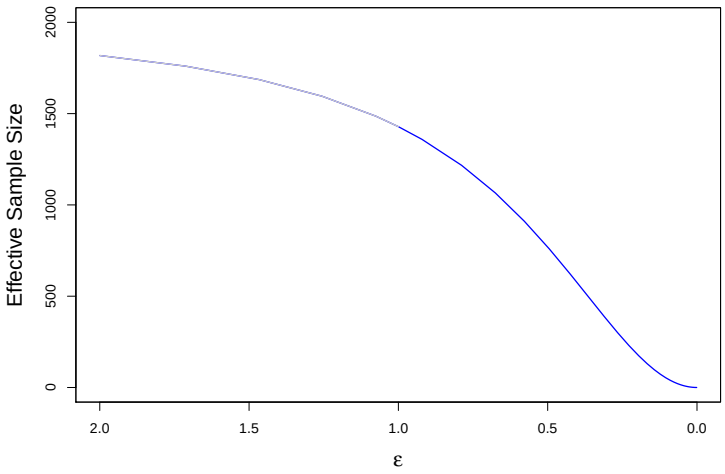

Figure 6: *The equivalent number of observations in a difference of means test, that would create a sampling distribution of the same noise as a sample distribution of 2000 observations under differential privacy, across different ranges of the privacy parameter ϵ. This parameter is commonly below 1, and in that range the effective sample size decreases as approximately $3/4\ \epsilon\ n$.*

4.6 Effective Sample Sizes

We saw in the Monte Carlo example, for these particular parameter values, the variance of the differentially private estimate with 2000 observations had the variance one would expect from a difference of means estimate that did not preserve privacy, of 800 observations. We can generalize this understanding of an *effective sample size* that results from differential privacy. The variance of the difference of means estimator (ignoring the censoring) is:

$$\text{Var}(\bar{x}_{1-0}) = \frac{sd(y_1)^2}{n_1} + \frac{sd(y_0)^2}{n_0} = \frac{4\nu^2}{n} \tag{15}$$

in our example where both the variance and number of observations is the same for both the treatment and control observations. or the differentially private statistic the variance is:

$$\text{Var}(M(x)) = \text{Var}(\bar{x}_{1-0}) + \text{Var}(Z) = \frac{4\nu^2}{n} + 2\frac{\Delta f^2}{\epsilon^2} = \frac{4\nu^2}{n} + \frac{32R^2}{\epsilon^2 n^2} \tag{16}$$

Where R, the range of the data is $y_{\max} - y_{\min} = 1$. Notice that the variance of Z collapses as n^2 while the variance of the difference of means test only collapses as n, thus these distributions become relatively more close as n increases. From this we can solve for an *effective sample size*, n^{eff} that results from differential privacy. The effective sample size is a number of observations that would be required if we permitted access to the results from the private data, so as to get the same standard error as the noisier results from differential

privacy. Equations 15 and 16 combine to solve this as:

$$n^{\text{eff}} = \frac{4\nu^2\epsilon^2 n^2}{4\nu^2\epsilon^2 n + 32R^2} \tag{17}$$

We graph this in figure 6, for our Monte Carlo example ($n = 2000, \nu = 0.1, R = 1$) across all values of ϵ. For $\epsilon < 1$, the section denoted in blue, the curve is roughly approximated by $n^{\text{eff}} = 3/4\ \epsilon\ n$, thus as we decrease ϵ to increase the level of privacy protection, we pay roughly linearly in effective sample size.

4.7 Matching Methods

It is often the case in observational data, that we are unable to randomize the treatment value as we would desire in an experimental study. In such cases, matching methods provide a mechanism by which we can extract a set of treatment and control observations, that attempt to retain the property that other causal variables have the same distribution among both the treated and control subgroups.

In Appendix A we prove that if a dataset, D, is constructed by matched pairs, and then a function f with sensitivity Δf is computed on the pair-matched data, then the sensitivity of the entire operation (matching and computing the function) is at most $3\Delta f$.

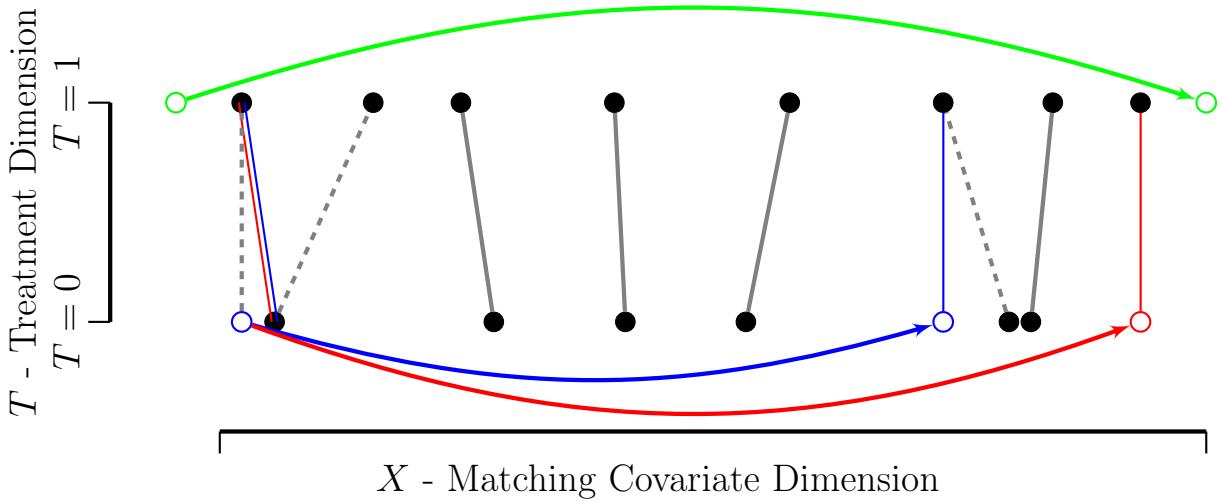

Figure 7: *Sensitivity of paired matching methods. Dashed lines represent matches that are broken by the movement of x to some x'. Colored lines represent new matches that are created by that movement.*

We sketch out the logic of this proof in figure 7. The x axis represents some covariate dimension on which we match observations, while the y dimension is treatment. To create matched pairs, we attempt to find one treated and one control observation that have similar x values. The grey lines (both solid and dashed) represent a set of possible matches in some

dataset. To understand the sensitivity of pair matching, we want to consider how these matches might change if some single observation is allowed to be moved in an arbitrary fashion. Obviously in paired matches, the number of matched observations is necessarily even. In the proof, we show that the number of observations that are matched can only change by -2, 0, or 2 observations, and the total number of observations that change is at most 4. As an example, in figure 7 the blue arrow represents moving one control observation from a low to a high value of x. As it moves, its original match is broken, which causes a knock-on effect of another match to be broken, and the new match (depicted as a red line). However, with all this movement, so far all that has changed is one treatment and one control observation have left the dataset. Following the blue arrow, after movement the new x is a better match than existed in a previous pair, and thus a new match is formed. In total, 3 observations dropped out of the dataset (including the original observation that moved), and one was added (the newly moved observation) for a net loss of two observations. The red arrow shows an example where movement changes the dataset by 4 observations for no net loss. The green arrow shows a movement from an unmatched location to a new location that remains unmatched, and does not change the matched dataset at all.

In the instances where there are changes, half of these changes are occuring because we move one observation in the dataset, and the other half are occuring because their matches may change. However, the effect of changing one observation in the dataset is already accounted for in the sensitivity of the estimator we are going to compute. Matching, potentially doubles this effect, as the matched observations that change can also have a potential effect on the estimator. Since the sensitivity is the upper bound on this effect, the matches that enter or exit the dataset can themselves also effect the estimator by at most, the value of the sensitivity. Thus in total, the largest possible effect of the movement of one observation on a function in a matched dataset is twice the sensitivity of the function itself.

This is a powerful result because it means all no new sensitivities need to be derived for functions as a result of pair matching in the dataset. Instead, we just adjust the sensitivities used in the mechanism appropriate for whatever function we are planning to use after matching. For example, if we set ϵ to 1.5, instead of 0.5, all the simulations in 4 remain valid if the difference of means was being calculated after matching.

4.8 Confidence Intervals from Differentially Private Releases

The differentially private difference of means test, as we saw in equation 16, has variance both from the underlying sampling distribution of the difference of means test, and from the Laplace noise added to preserve privacy. That means confidence intervals that are constructed to provide coverage of the true population value, need to incorporate both sources of error. Fortunately, the variance of the Laplace distribution from which the noise is drawn is public information[5]. Unfortunately, the combination of Laplace noise and (asymptotically) Gaussian sampling error make for a difficult distribution to integrate so as to construct

[5]This accords with the concept that the privacy-preserving mechanism is known and public, and explicitly avoids "secrecy through obscurity".

confidence intervals. It is feasible to use Monte Carlo integration for this problem, that is, draw simulations from a mean zero Gaussian of given variance, and a mean zero Laplace of known variance, and numerically calculate bounds on this distribution that capture the desired fraction of the distribution.

Algorithm 2: Numerical Confidence Interval for Difference of Means

1. **for** i in 1 to k **do**
2. Draw $s \sim f_{Normal}(\mu = 0, \sigma = M_{\text{se}})$
3. Draw $n \sim f_{Laplace}(\mu = 0, b = \Delta f/\epsilon)$
4. Compute $q_i = |s + n|$
5. **endfor**
6. Compute index $b = \lfloor k\alpha \rfloor$
7. Sort the results decreasing and find $q_{(b)}$
8. Calculate $\text{CI}_{1-\alpha} = M_{\text{dom}} \pm q_{(b)}$

However, if an analytical approach is required, it is conservative to calculate the variance of the sum of the Gaussian and the Laplace (which simply sum together linearly), and then compute a confidence interval using this variance and the Laplace distribution The critical value of a Laplace of variance A, is guaranteed to be larger than the critical value of some combination of Gaussian of Variance ρA and Laplace $(1-\rho)A$.[6] This gives us:

$$CI_{95}[\bar{y}_1 - \bar{y}_0] = M_{\text{dom}}(X) \pm 2.996\ \sigma \tag{18}$$

$$\sigma^2 = M_{\text{se}}(X)^2 + 2(\Delta f/\epsilon)^2 \tag{19}$$

$$\Delta f = \frac{x_{\max} - x_{\min}}{N_1 + 1} + \frac{x_{\max} - x_{\min}}{N_0 + 1} \tag{20}$$

Where 2.996 is the critical value of the Laplace, and $M_{\text{dom}}(X)$ and $M_{\text{se}}(X)$ are the differentially private releases of the difference of means estimator and standard error of the difference of means.

[6] While this is conservative, we also saw that the differentially private standard error estimates were attenuated.

4.9 Replication Example: A Policy Analysis of Government Transfers to Promote Women's Health

As an example of the differentially private mechanisms for causal analysis, we replicate part of an analysis by Lim et al. [18] on a policy assessment of India's Janani Suraksha Yojana (JSY), a cash transfer program designed to encourage women to participate in antenatal care and birthing facilities. This study was replicated by Carvalho and Rokicki [19] and we use their constructed data [20] originally from India's District Level Household and Facility Survey (DLHS-3) [21].

We are interested in the causal effect of the cash transfer program on women's health outcomes, therefore, we take the treatment variable to be receiving cash assistance from the JSY program. As an outcome, we measure the incidence of women delivering babies at in-facility care centers, one of the outcomes intended to be incentivised by this program. Carvalho et al. show that structural differences in how the program was run across different states meant that the treatment effect varies by regional government [19, 22]. Thus we have 33 different Indian states with 33 different treatment effects. These states differ in both number of treated observations, and size of treatment effect, even though the quantities of interest, measurement instruments, and covariates in the analysis are all the same. This allows us to witness the effect of our differentially private mechanisms across a range of parameter values. The data used in this analysis are public, thus allowing us to compare the results of our methods to the sample values in the true data. However, the variables used–individual healthcare outcomes and government cash assistance–are very much in the style of personal data that commonly require privacy protection in social science research.

In figure 8 we show the difference of means estimates of the treatment effect of the JSY cash transfer on the probability of delivering at an in-facility birthing center. Estimates and confidence intervals are shown in pairs; each bottom blue line shows the sample estimate and confidence interval directly from the private data, while on top the red version shows the differentially private release of the estimate and its confidence interval.

The top graph orders the results by the estimated treatment effect in the private data. We see, in these examples, generally the privacy preserving results track the results. Pubjab and Dadra are exceptions. Andaman reverses the direction of the effect, although both intervals include zero. Daman has largely exagerated treatment effects.

The bottom graph arranges the estimates by the sample size of the pair-matched dataset. We see that the states in which we have fewer than 50 observations give very inaccurate answers with confidence intervals generally beyond this scale of the x-axis; here is where our most of our troubling examples are located, such as Daman, Dadra and Andaman. The band of observations between 100 and 500 observations have some large added noise, but generally give the same inference as the confidence intervals from the private data. The exception is the Punjab result where the differentially private confidence intervals includes zero and the version from private data does not. The differentially private interval, does however, cover the sample estimate, and also would not include zero at 90 percent confidence. Above 500 observations the confidence intervals visually quite accurately resemble the private values.

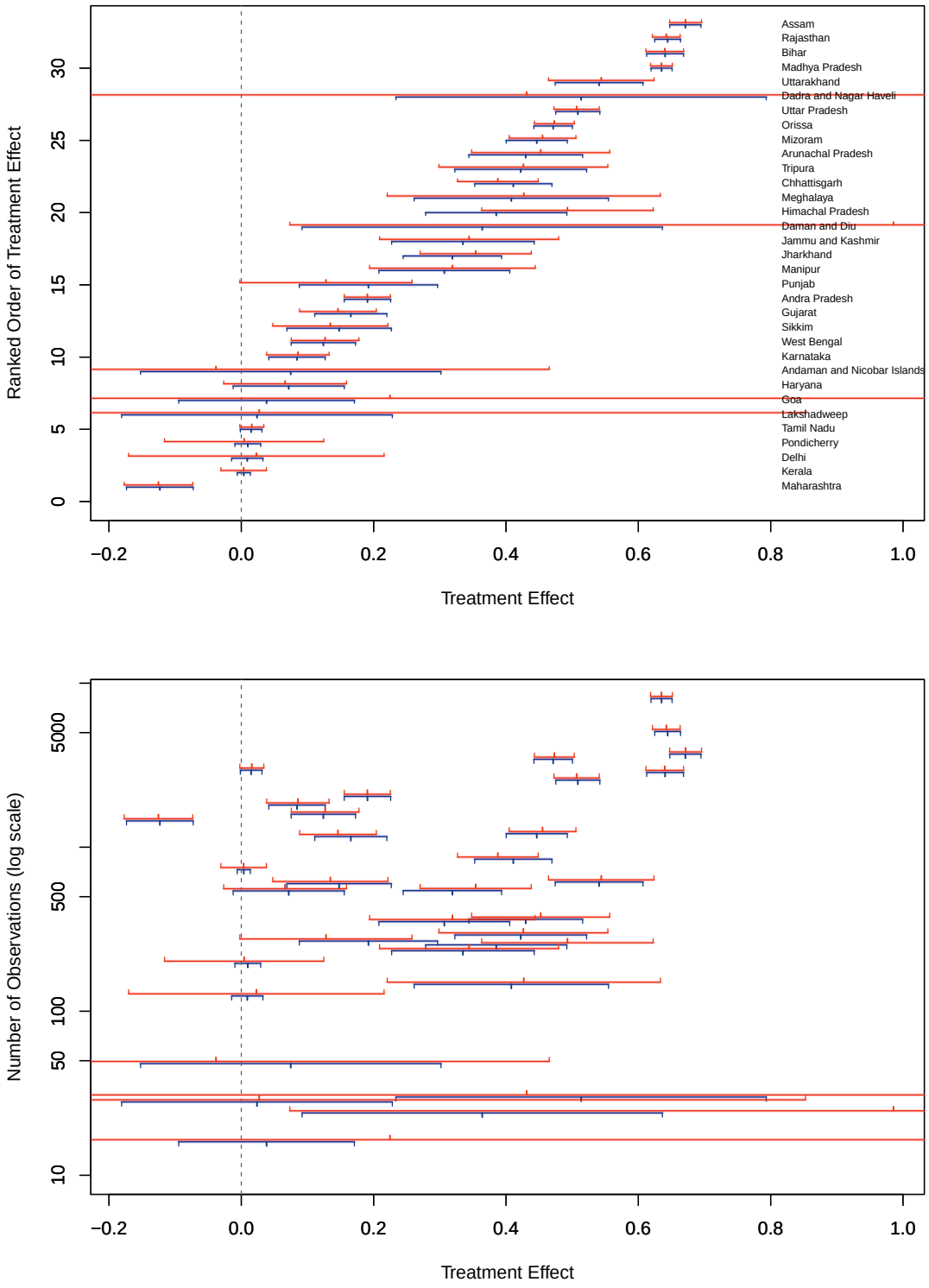

Figure 8: *Difference of means estimates across 33 Indian states for the treatment effect of JSY cash transfers to women on the probability of delivering at an in-facility birthing center. Each bottom blue line shows the sample estimate and confidence interval in private data, while the red version on top shows the differentially private release of the estimate and confidence interval.*

5 Private Regression

Next we consider mechanisms for releasing the coefficients of linear regressions in a differentially private manner. Assume we have some dataset $\mathbf{Z}$, and we want to release the estimates of $\mathbf{Z}'\mathbf{Z}$. By means of the sweep operator, coefficients and standard errors of regressions of any combination of variables in Z can be computed purely from the information in $\mathbf{Z}'\mathbf{Z}$ [23] [24], [25]. This allows us to estimate any regression without detracting from the privacy budget beyond what is necessary to release $\mathbf{Z}'\mathbf{Z}$, making it valuable to mimic an interactive setting where researchers want to analyze sensitive data that sits inaccessible behind secure storage (see section 6).

The general method for releasing a differentially private covariance matrix is to add Gaussian noise to each element of $\mathbf{Z}'\mathbf{Z}$, a notion originally proposed by [26] and refined by [27]. The noise is mean 0 and has a variance proportional to the sensitivity of the element. However, the privatized $\mathbf{Z}'\mathbf{Z}$ is often too noisy to retrieve meaningful regression estimates. This is common when there is the possibility of extreme values in $\mathbf{Z}$, which is typical of social science data where variables such as *income* have a heavy skew. The concept of sensitivity is to calculate the effect the largest change in one observation can have on the function. However, outliers can have notoriously large effects on regression estimates, and thus adding noise proportional to sensitivity, or noise sufficient to drown out any possible outlier, will cause problems for regression estimates.

Our approach is to first trim the data, effectively removing the extreme values and reducing the noise necessary to mask a change from an observation's true values to the new "extremes." Although trimming biases our estimates, least squares' estimates are sensitive to extreme values and thus they are often *intentionally* omitted, sometimes for theoretical reasons, sometimes to improve performance, and sometimes for robustness.[7] Even if we are in the anti-trimming camp, in our application it is beneficial so long as trimming extreme values allows us to estimate a privatized covariance matrix whose regression estimates are closer to the truth than a non-trimmed, privatized covariance matrix.

The privacy preserving algorithm that we propose for $\mathbf{Z}'\mathbf{Z}$ is given in 3. The first step is to define a function, $f(x)$, that calculates the distance from any observation to a defined center of the data. For arbitrary distance metric, we trim all observations that are beyond some ordered threshold, m. Perhaps we set m to 95% of N; we would therefore be dropping the most extreme 5% of our data as per $f(x)$. Let r represent the distance from the origin to x_m, the observation at (or just prior to) m, or the one with the greatest distance *after* trimming.

To trim the data in a way that is, itself, differentially private, we add noise to r using the Gaussian Mechanism. For intuition, Figure 9 is a visual depiction of the sensitivity for

[7] Trimming as a preprocessing step in data exploration is common in exploratory settings, such as [28] who trim data to improve performance of a machine learning algorithm. In Economics, for example, extreme values of the rate of inflation are often trimmed [29]. In studying the effect of regime type on a country's foreign direct investment, [30] notes the effect of outliers and experiments with trimmed data. In Psychology, trimming is used as a robust statistical method in [31, 32].

Algorithm 3: Differentially Private Regression

1. define a function $f(x)$ that calculates the distance from a row to the origin
2. set trimming threshold percentile m
3. calculate $f(x_i)$ for each row
4. define radius r to be $f(x_m)$ where x_m is the observation at the m percentile
5. compute Δf, the sensitivity of the radius, to be $f(x_{m+1}) - f(x_{m-1})$
6. set r^* to be $r + \iota$ where $\iota \sim N(0, \tau^2)$ and $\tau = \Delta f \sqrt{2ln(1.25/\delta)}\epsilon$
7. trim all observations where $f(x_i) > r^*$
8. compute sensitivity of trimmed $\mathbf{Z'Z}$ for each element
9. Add noise $z_j' z_i = z_j' z_i + \gamma$, $\gamma \sim N(0, \tau^2)$, $\tau = \Delta f \sqrt{2ln(1.25/\delta)}\epsilon$

trimming. x_m is the point that is at (or just prior to) the threshold m. Δf, our sensitivity, is the distance from x_{m-1} to x_{m+1}. The reason for this is that if we move any observation from the left of x_m to the right of x_m, as shown in *red*, then r adjusts to x_{m+1}. Likewise, if we move any observation from the right of x_m to the left of x_m, as shown in *blue*, then r adjusts to at most x_{m-1}.

It is possible to move an observation anywhere, not just from one relative extreme to the other; two depictions of this are shown in Figure 10. The *orange* shows a move from the right of x_{m+1} to the right of x_{m+1}. Clearly, the threshold is not affected. The *green* shows a move from the left of x_{m-1} to a position between x_m and x_{m+1}. This redefines x_{m+1} to be closer to x_m, resulting in a smaller Δf. Therefore, the original Δf remains at worst a conservative measure of the sensitivity for trimming.[8] So, regardless of the presence or absence of any individual observation, r ranges from $f(x_{m-1})$ to $f(x_{m+1})$, and our sensitivity is Δf. The differentially private threshold for trimming is $r + \iota$ where $\iota \sim N(0, \tau^2)$ and $\tau = \Delta f \sqrt{2ln(1.25/\delta)}\epsilon$.

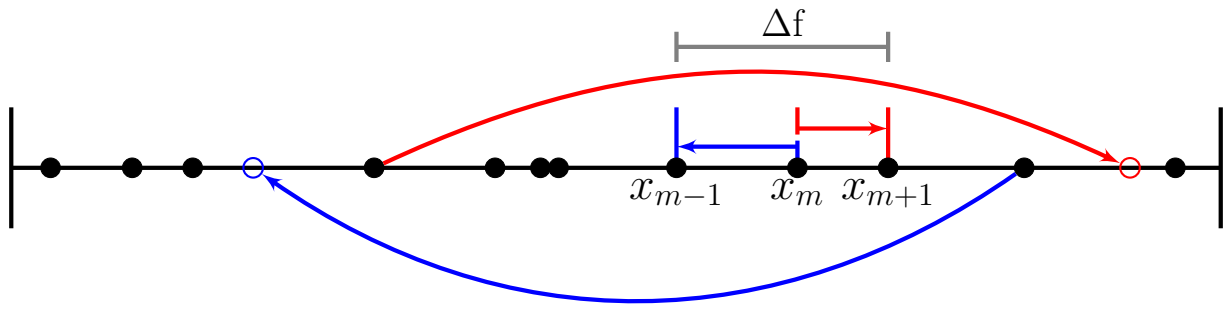

Figure 9: *Observations moving across the trimmed hull.*

[8]This is not true for any distance function f. In our example, we use a unidimensional distance where this is true.

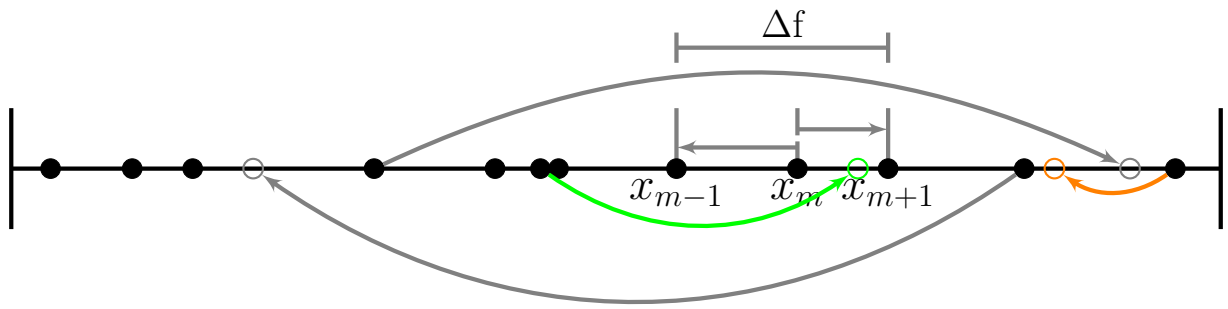

Figure 10: *Sensitivity of the radius/range of a trimmed dataset.*

Finally, we compute the sensitivity of the trimmed $\mathbf{Z}'\mathbf{Z}$ and add noise to each element $z_j'z_i = z_j'z_i + \gamma$, $\gamma \sim N(0, \tau^2)$, $\tau = \Delta f\sqrt{2ln(1.25/\delta)}\epsilon$.

5.1 Monte Carlo Study

We construct Monte Carlo data generated from a joint multivariate Normal distribution as:

$$\{Y, X\} \sim f_{\text{mvn}}(\mathbf{0}, \Sigma) \tag{21}$$

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1k} \\ \sigma_{12} & 1 & \sigma_{23} & \cdots & \sigma_{2k} \\ \sigma_{13} & \sigma_{23} & 1 & & \\ \vdots & \vdots & & \ddots & \\ \sigma_{1k} & \sigma_{2k} & & & 1 \end{pmatrix} \tag{22}$$

We treat the first variable as the outcome of interest, Y, and the remaining variables as possible explanatory factors. We set $\mathbf{Z} = \{\mathbf{1}, \mathbf{Y}, \mathbf{X}\}$, that is, the data above prepended with a column of 1's. For each dataset we are interested in estimates of $\mathbf{Z}'\mathbf{Z}$, from which we can compute any regression coefficients and their standard errors. We assume the equation of interest is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \tag{23}$$

The simplest approach to constructing a differentially private version of $\mathbf{Z}'\mathbf{Z}$ is to add Gaussian noise to each unique term, with variance proportional to *sensitivity*, τ, as:

$$z_j'z_i = z_j'z_i + \gamma; \qquad \gamma \sim \mathcal{N}(0, \tau^2) \tag{24}$$

$$\tau = \Delta f\sqrt{2\,\ln(1.25/\delta)}/\epsilon \tag{25}$$

Here, the minimum distance that could be calculated is 0 and corresponds to an observation of all 0s. The maximum distance is r^* and can be achieved through different combinations of values. In the examples that follow, we compute two versions of local sensitivity,

where the true sensitivity is somewhere between the two.The first sensitivity is calculated as the change in each cell's value when the "closest" row, or that with the smallest $f(x)$, is recoded as the "farthest" row, or that with the largest $f(x)$, and vice versa. This provides us with three values for each cell in $\mathbf{Z}'\mathbf{Z}$: base value, min-to-max value, and max-to-min value. Δf is calculated as the largest of the three values minus the smallest of the three values.

This provides an empirical, or local, sensitivity that is likely smaller than the theoretical sensitivity. We therefore compute another sensitivity score, only this time rather than shift an entire row, we shift variable-by-variable. Thus, each value in the row with the smallest $f(x)$ is recoded as the largest value observed in the data for that variable. Similarly, each value in the row with the largest $f(x)$ is recoded as the smallest value observed in the data for that variable. This will produce an observation whose $f(x)$ is so large that it would be trimmed, and another observation whose $f(x)$ will be at or extremely close to 0. Thus, the sensitivity is larger than would be expected.

Note that if the data is bounded $(-4, 4)$ (with observations beyond that range censored), any term of the inner product $\mathbf{Z}'\mathbf{Z}$ has sensitivity $\Delta f = 2(16)$, and for an entire row we have $\Delta f = 32\sqrt{k}$.[9][10]

We first set $\sigma_{12} = 0.3$, $\sigma_{13} = 0.15$, and $\sigma_{23} = 0.10$ to have small correlation with both Y and the first X, while leaving all other covariances as 0. In what follows, k is set to 10. We simulate 100 datasets, calculating the distribution of the regression coefficients in the sample data, the Gaussian, and the trimmed mechanisms. The distributions of the estimated coefficients are presented in figure 11. Left to right are the results for the intercept, β_0, the coefficient on the strong relationship β_1, weak relationship β_2, and nonexistent relationship β_3. In the top row in blue are the density plots of the estimates from regression on the private sample data, as the tightest spike, that varies across the simulations because of sample variability. Overlaid are the estimates of the coefficients for the simplest private noise in red, the trimmed private covariances with small sensitivity in green, and the trimmed private covariances with large sensitivity in purple. For each coefficient, the green and purple distributions are noticeably tighter to the private blue distribution.

Figure 12 explores the the standard errors resulting from these regressions. Here the green distribution is generally shifted to the right of the blue distribution, meaning the standard errors are larger from the trimmed private regressions than with the observed data. Given the extra noise and reduction in information, this is reasonable. The red distribution is both much larger and troublingly much smaller than the blue distribution. This means we have simulations where the private regressions convey less uncertainty in the estimated effect than is possible with the observed data.

The bottom row in figure 12 shows the t-test generated from these standard errors. In this model, t-test of magnitude greater than 1.96 reject the null hypothesis of no relationship between x and y at 95 percent confidence. These critical values are drawn as orange lines. The t-test of regressions from the observed data, are plotted against the simple private

[9]Note the peculiarity that zero centering the range of the data to $(-a, a)$ as opposed to bounding $(0, 2a)$ reduces the sensitivity from $4a^2\sqrt{k}$ to $2a^2\sqrt{k}$.

[10]For k variables, $\mathbf{Z}'\mathbf{Z}$ has $(k+2)(k+1)/2$ unique terms.

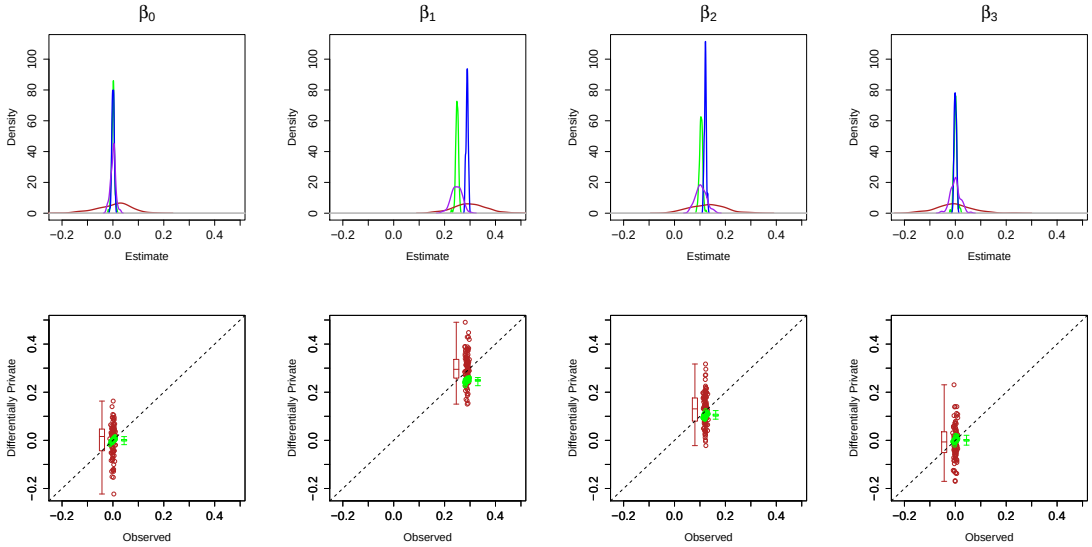

Figure 11: *Estimated coefficients, from private data (blue), trimmed DP statistics with small sensitivity (green), trimmed DP statistics with large sensitivity (purple), and pure differentially private statistics (red).*

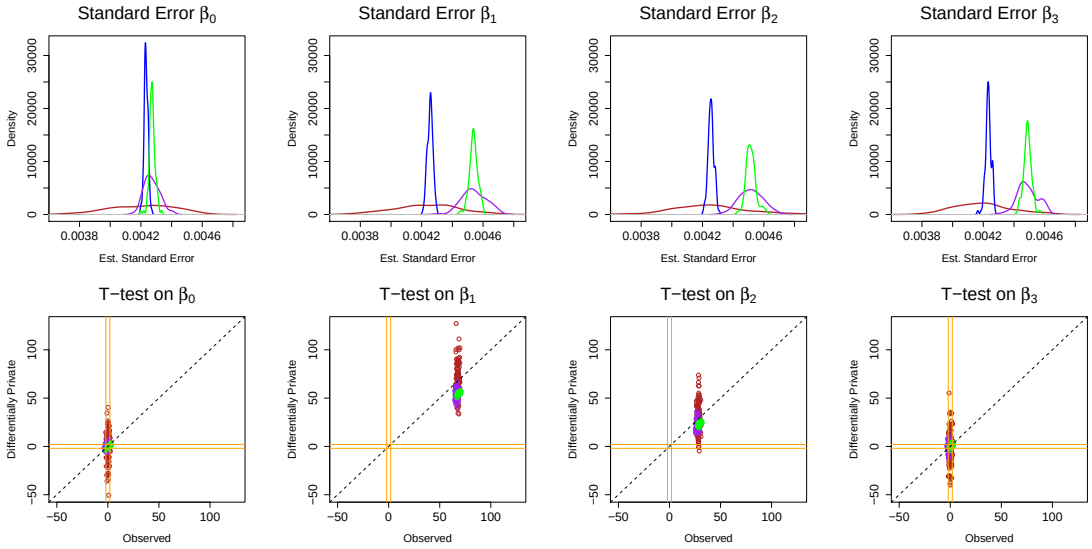

Figure 12: *Estimated standard errors, from private data (blue), trimmed DP statistics with small sensitivity (green), trimmed DP statistics with large sensitivity (purple), and pure differentially private statistics (red).*

regressions (red), those from the trimmed data with small sensitivity (green), and those from the trimmed data with large sensitivity (purple). Recall, β_0 and β_3 both in truth zero, and most of the green simulations fall in the narrow band that correctly fails to reject the null, while the red simulations spread beyond the critical values in both directions. Similarly, β_1 and β_2 are not in truth zero, and while all the simulations reach the correct conclusion for the large effect of β_1, some of the red observations give the wrong answer for weaker true effect of β_2.

Finally, figure 13 shows the Mean Squared Error (MSE) between the observed coefficients and the private versions, as the covariance σ_{12} moves across the range $(-0.95, 0.95)$. For the simple private regressions, the errors are highest when the magnitude of the covariance is greatest. The MSE of the trimmed coefficients might have the same shape, but at this scale they are so much smaller that it is difficult to tell. The ratio of the MSE's between the two estimators, presented in the right graph, shows that the simple private regression has between 50 and 70 times more MSE than the trimmed version.

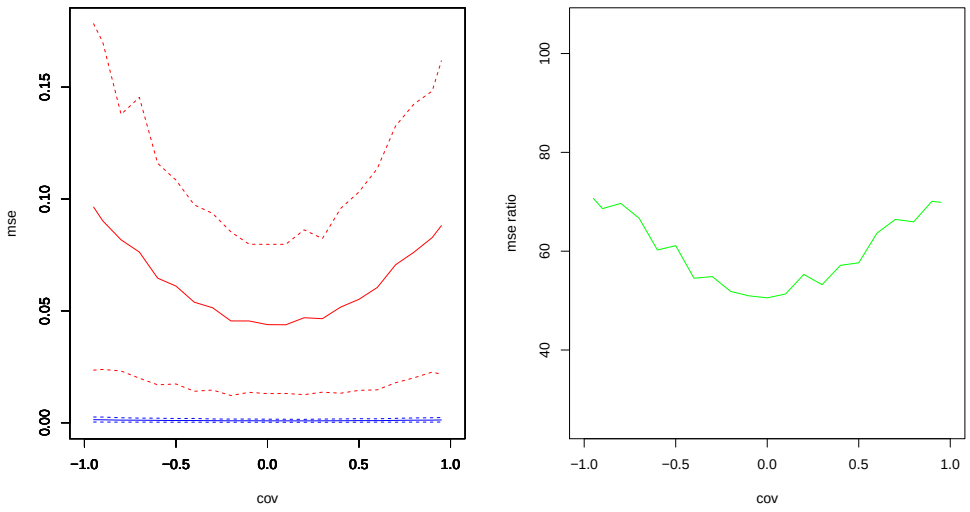

Figure 13: *The left graph show mean squared error of private released statistics for pure private statistics (red), and trimmed private statistics (blue), across various strengths of the relationship. The right graph shows the ratio of these two errors.*

6 A Curator Architecture for Private Data Analysis

Differential privacy gives us mechanisms by which we can release useful statistical estimates in datasets, with guarantees that they can not be combined to violate individual privacy.

A curator model provides an architecture for learning from private data, without access to the raw underlying dataset [13, 5]. The curator acts as an intermediary between data users and private datasets. The data resides in secure storage and is never available to the user. Users submit queries, or perhaps statistical models, to the curator; the curator in turn responds with query answers, or model estimates. In the case of differential privacy, these responses are not exact answers from the data, but differentially private versions of any query or model estimate.

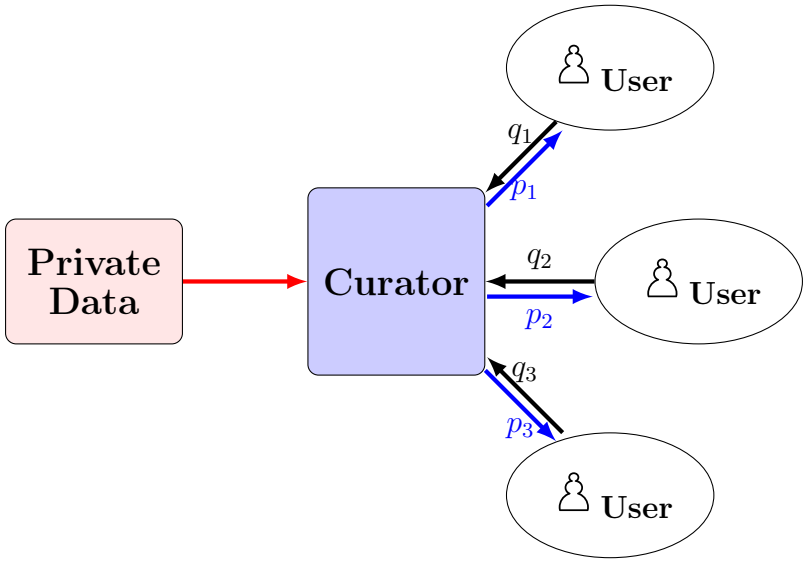

Figure 14: *The curator architecture for data privacy.*

In an interactive setting, the curator has access to the private data and can compute differentially private answers to submitted queries. The data has a *privacy budget*, ϵ, governed by the same expressions we saw previously in section 2.1. This ϵ sets the privacy of the dataset, and as in equation 1 determines the upper bound of the probability difference of any reported result between neighboring datasets. Each query, q_i, to the dataset has its own precision ϵ_i, that determines how much precision, and thus privacy leakage, is permitted for that query. The combination of all ϵ_i can not exceed the total ϵ of the dataset. When the total queries reach ϵ, the privacy budget of the dataset has been exhausted, and no more queries can be answered. If more queries were answered, then some combination of queries might leak information with probability greater than the risk guarantee of differential privacy. A capable curator should record previous questions and answers that have been provided; if

the same question arises the curator can provide the exact same answer without touching the private data, or expending any further ϵ. [11]

In an noninteractive setting, the entire privacy budget is immediately exhausted by a set of questions that are anticipated to be of most use to future users. The dataset is then closed to all future access. The questions might include a variety of summary statistics or measures of relationships in the data. If they include the *sufficient statistics* for a class of statistical model, then all statistical models in that class can be answered, mimicking an interactive setting, even though no new queries are calculated, and all answers are the result of past computation. For example, if for a set of variables, the covariance matrix, vector of means, and sample size have been provided, then any possible linear regression among those variables can be calculated from those sufficient statistics.

6.1 The Depositor Interface

For a curator interface to work, the responsible owner (depositor) of the data must first add the dataset to the curator, and set a global value of ϵ to determine the privacy level of the data. At this point, the depositor might decide to spend some of the privacy budget on statistical answers that the owner anticipates will be of use to future users, and how to parition levels of precision, or ϵ_i, to those answers. If the entire budget is spent on releasing differentially private statistics, then the dataset is closed, and the curator is noninteractive. Or, some portion of ϵ may be reserved for future interactive queries of users that are unanticipated. For example, in a large dataset, a depositor might choose to use the entire budget to release summary statistics for most variables, and a precise version of the covariance matrix of a subset. Or, if they can not anticipate which variables (or interactions) to include in the covariance matrix, they might let any future regression be run, and a Laplace or Gaussian mechanism operating on just those requested coefficients will slowly exhaust the privacy budget by future users until the data set has to be closed.

In our tool, we provide an interface for the depositor, as shown in figure 15. The data depositor, first sets a global ϵ, which may be informed by an interactive interview process. The user then builds up a list of statistics to be released for each variable in the dataset, and can partition the privacy budget between these statistics. For any given ϵ_i given to a particular released statistic, the interface calculates a projected level of accuracy to guide the user in making relative tradeoffs in the privacy budget.

When the data depositor has distributed their privacy budget among the statistics they wish to release, the second portion of our tool system draws differentially private versions of those statistical summaries selected by the data depositor from a library of differentially private routines (which we created in the R statistical language, and also make available for

[11] Similarly, there are circumstances in which certain combinations of questions have less mutual information, and thus their associated ϵ's add up in a more forgiving and less than linear fashion, permitting more questions to the dataset before exhausting the privacy budget. The theoretical results on *composition theorems* that calculate the effect of combinations of queries on the total privacy budget is an active literature.

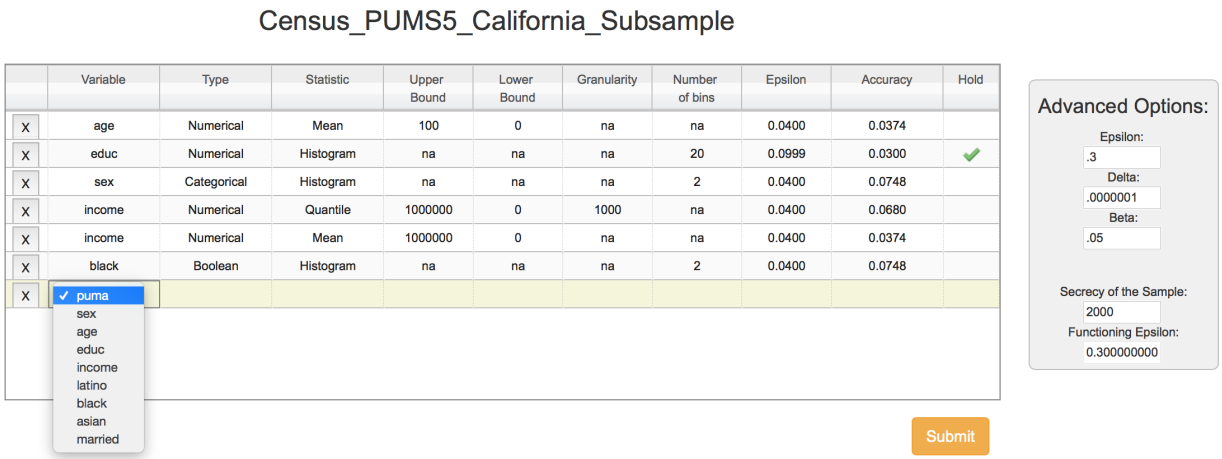

Figure 15: *Example screen from the interactive privacy budget allocation tool for data depositors.*

use by the R community) and stores them in metadata associated with that file on Dataverse. Future researchers who wish to explore restricted social science data can then access these privacy-preserving summary statistics either from the metadata, or through the TwoRavens graphical data exploration tool built for Dataverse, which we have adapted for differentially private statistics, which we describe in the next section.

6.2 The Query Interface

At the other end, users who want to see summary statistics, make queries, or run statistical models on the private data, must work through the curator interface, without access to the raw data. We provide an interface using a branch of the TwoRavens project [33, 34], a thin client, gesture driven, browser based interface for statistical analysis. The architecture of the TwoRavens interface is shown in figure 17. The data remains on a secure server, archived on an instance of Dataverse [35] [36], a repository for social science data. Meta data is available to the TwoRavens interface, which includes any released differentially private statistics, such as means or density plots. Regressions can be constructed by means of directed graph, and sent to run on a remote server, using *R* libraries such as *Zelig*. Importantly, the data itself is never available on the client side, creating a secure architecture for a curator. An example of the user interface for exploring summary statistics and building a model using Census data is shown in figure 16.

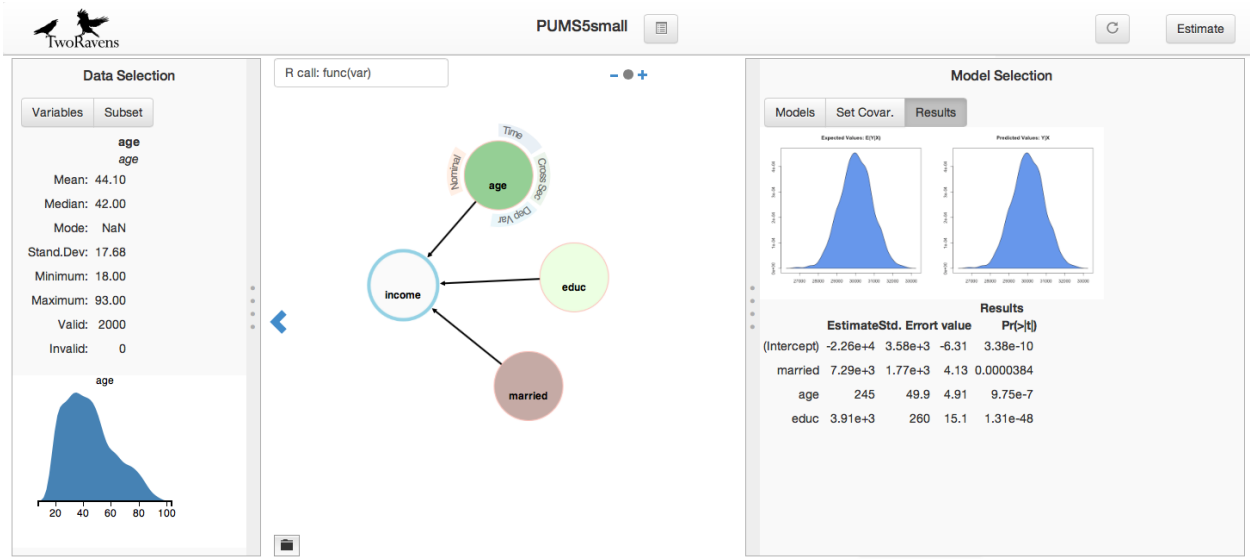

Figure 16: *Example screen from the TwoRavens statistical interface used here as a query interface for exploring differentially private summary statistics and private statistical models.*

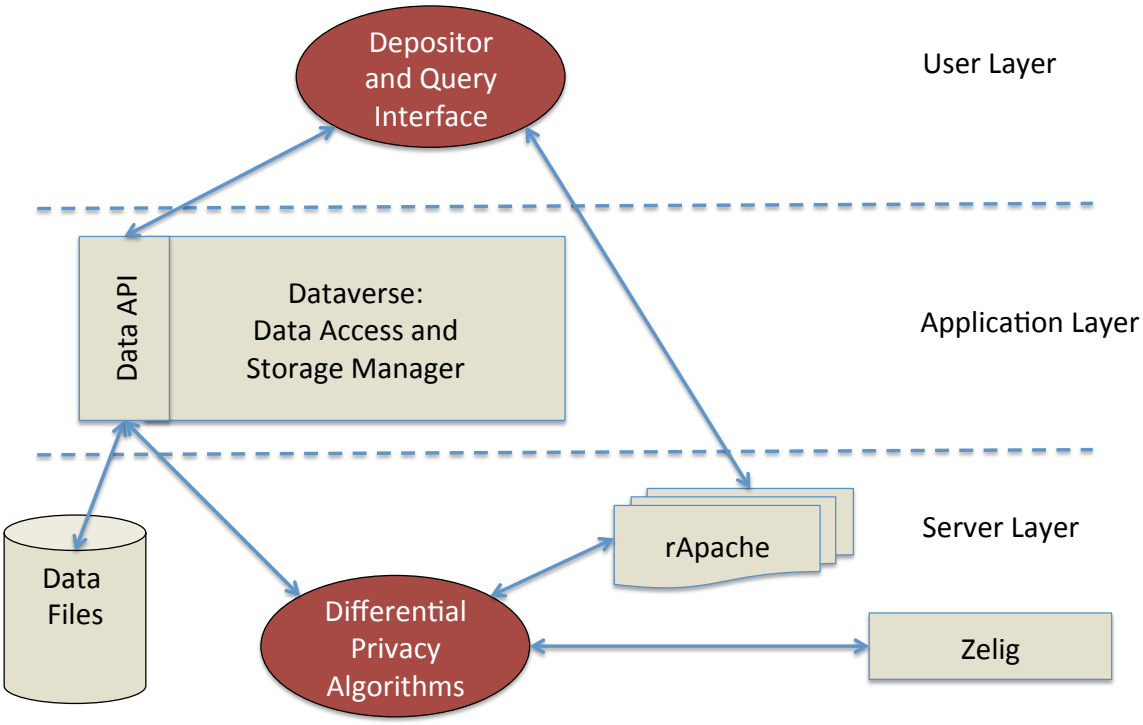

Figure 17: *Privacy architecture for secure curator interfaces.*

7 Conclusion

Social scientists often analyze data that contains information that, for legal, moral, or professional reasons, must guarantee privacy to the research subjects. In tension with this, there are strong scientific motivations to share and distribute data openly.

Differential privacy is one mathematical conception of privacy preservation that allows researchers to release statistical estimates from their dataset with the strong guarantee that no individual's privacy can be comprised, regardless of the combination of queries or estimates computed, and regardless of any auxiliary information that can be combined.

We have introduced differential privacy and basic methods for computing privatized summary statistics. We have derived new results to implement differentially private mechanisms to release causal estimates, and estimates after pairwise matching, as well as a new algorithm for computing a privatized covariance matrix and, therefore, any linear regression from a dataset. These methods enable researchers to compute the results of inferential models in restricted data without any access to the underlying data.

We make these methods available to researchers through a system of privacy preserving tools that implement these methods, as well as many methods for simpler summary statistics. These tools, together with interfaces we have developed for distributing the privacy budget and exploring differentially private statistics, in combination with a secure data repository such as the Dataverse Network, together form a *curator architecture* for data analysis. This architecture allows potential researchers to explore data and calculate statistical estimates, without any access to the underlying raw data, and while provably protecting the privacy of individual-level information.

While ostensibly these methods prevent researchers from accessing data, our goal is to increase the ease with which researchers can gain access to results in restricted access datasets, and increase the facility of researchers to search across restricted sources to identify the ideal studies for their research. The increasing ability of big data collections, sensor data, and social media data to measure individual behavior in nuance, ensures that such privacy concerns will only increase. With this increase in the invasiveness and pervasiveness of data collection methods, and the centrality of such data to modern social science, comes a new need for methodological work on privacy-preservation in quantitative research, so that subjects continue to trust social scientists with their personal information. The strong privacy guarantees of differential privacy provide an important tool in the development of rigorous privacy-preservation in quantitative social science.

References

[1] G. King, "Replication, replication," *PS: Political Science and Politics*, vol. 28, pp. 443–499, 1995.

[2] National Science Foundation, "Award and administration guide. chapter vi: Other post award requirements and considerations. section d.4: Dissemination and sharing of re-

search results," http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4, accessed: 2015-04-14.

[3] W. Jacoby, "The ajps replication policy: Innovations and revisions," http://ajps.org/2015/03/26/the-ajps-replication-policy-innovations-and-revisions/, accessed: 2015-04-14.

[4] N. P. Gleditsch, C. Metelis, and H. Strand, "Posting your data: Will you be scooped or will you be famous?" *International Studies Perspectives*.

[5] M. Crosas, G. King, J. Honaker, and L. Sweeney, "Automating open science for big data," *The ANNALS of the American Academy of Political and Social Science*, vol. 659, no. 1, pp. 260–273, 2015. [Online]. Available: http://ann.sagepub.com/content/659/1/260.abstract

[6] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *The Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, pp. 98–110, 1997.

[7] ——, "Uniqueness of simple demographics in the us population," Technical report, Carnegie Mellon University, Tech. Rep., 2000.

[8] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 2008, pp. 111–125.

[9] T. Steinke and J. Ullman, "Between pure and approximate differential privacy," *arXiv preprint arXiv:1501.06095*, 2015.

[10] J. Ullman, "Answering n {2+ o (1)} counting queries with differential privacy is hard," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 361–370.

[11] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2003, pp. 202–210.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography*. Springer, 2006, pp. 265–284.

[13] C. Dwork and A. Smith, "Differential privacy for statistics: what we know and what we want to learn." *Journal of Privacy and Confidentiality*, vol. 1, no. 2, pp. 135–154, 2009.

[14] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, pp. 1–34, 2000.

[15] N. R. Adam and J. C. Worthmann, "Security-control methods for statistical databases: a comparative study," *ACM Computing Surveys (CSUR)*, vol. 21, no. 4, pp. 515–556, 1989.

[16] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *39th ACM Symposium on Theory of Computing, STOC'07*. ACM, 2011, pp. 75–84.

[17] A. Smith, "Privacy-preserving statistical estimation with optimal convergence rates," in *43rd ACM Symposium on Theory of Computing, STOC'11*. ACM, 2011.

[18] S. S. Lim, L. Dandona, J. A. Hoisington, S. L. James, M. C. Hogan, and E. Gakidou, "India's janani suraksha yojana, a conditional cash transfer programme to increase births in health facilities: an impact evaluation," *Lancet*, vol. 375, no. 9730, pp. 2009–23, 2010.

[19] N. Carvalho and S. Rokicki, "The impact of india's jsy conditional cash transfer programme: A replication study," *3ie Replication Paper*, no. 6, 2015, washington, DC: International Initiative for Impact Evaluation (3ie).

[20] N. Carvalho, S. Kiatpongsan, and S. Rokicki, "Replication data for: A reassessment of india's janani suraksha yojana conditional cash transfer program: State-level effects matter," 2014. [Online]. Available: http://hdl.handle.net/1902.1/15899

[21] International Institute for Population Sciences (IIPS), "District level household and facility survey (dlhs-3), 2007-08," 2010, India. Mumbai: IIPS.

[22] N. Carvalho, N. Thacker, S. S. Gupta, and J. A. Salomon, "More evidence on the impact of india's conditional cash transfer program, janani suraksha yojana: Quasi-experimental evaluation of the effects on childhood immunization and other reproductive and child health outcomes," *PLoS ONE*, vol. 9, no. 10, 2014.

[23] J. H. Goodnight, "A tutorial on the sweep operator," *The American Statistician*, vol. 33, no. 3, pp. 149–158, 1979.

[24] J. L. Schafer, *Analysis of incomplete multivariate data*. London: Chapman & Hall, 1997.

[25] J. Honaker and G. King, "What to do about missing values in time series cross-section data," *American Journal of Political Science*, vol. 54, no. 2, pp. 561–581, April 2010, http://gking.harvard.edu/files/abs/pr-abs.shtml.

[26] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the sulq framework," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2005, pp. 128–138.

[27] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, "Analyze gauss: optimal bounds for privacy-preserving principal component analysis," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM, 2014, pp. 11–20.

[28] C.-K. Chui, B. Kao, and E. Hung, "Mining frequent itemsets from uncertain data," in *Advances in knowledge discovery and data mining*. Springer, 2007, pp. 47–58.

[29] A. Vaona and S. Schiavo, "Nonparametric and semiparametric evidence on the long-run effects of inflation on growth," *Economics Letters*, vol. 94, no. 3, pp. 452–458, 2007.

[30] S.-W. Choi, "The effect of outliers on regression analysis: regime type and foreign direct investment," *Quarterly Journal of Political Science*, vol. 4, no. 2, pp. 153–165, 2009.

[31] D. M. Erceg-Hurn and V. M. Mirosevich, "Modern robust statistical methods: an easy way to maximize the accuracy and power of your research." *American Psychologist*, vol. 63, no. 7, p. 591, 2008.

[32] R. R. Wilcox, "How many discoveries have been lost by ignoring modern statistical methods?" *American Psychologist*, vol. 53, no. 3, p. 300, 1998.

[33] J. Honaker and V. D'Orazio, "Statistical modeling by gesture: A graphical, browser-based statistical interface for data repositories," in *Extended Proceedings of ACM Hypertext 2014*. ACM, 2014.

[34] V. D'Orazio and J. Honaker, *A User Guide to TwoRavens: An overview of features and capabilities*, 2016. [Online]. Available: http://2ra.vn/papers/tworavens-guide.pdf

[35] G. King, "An introduction to the dataverse network as an infrastructure for data sharing," *Sociological Methods and Research*, vol. 36, pp. 173–199, 2007.

[36] M. Crosas, "The dataverse network: An open-source application for sharing, discovering and preserving data," *D-Lib Magazine*, vol. 17, pp. 1–2, 2011, doi:1045/january2011-crosas.

A Proofs

A.1 Notation

Assume every observation i of a random variable Y is observed to have outcome y_i, within bounds $y_{\min} \leq y_i \leq y_{\max}$ after treatment condition $t_i \in \{0, 1\}$. To compute sensitivity of the difference of means test, we wish to determine the largest effect that arbitrarily changing one observation, say the j-th observation, can have on this estimate. Note, we can change one observation by changing both its outcome and its treatment.

Consider the following partial sums over the dataset, ignoring any information in the j-th observation:

$$n_1 = \sum_{i \neq j} t_i \qquad \bar{y}_1 = \frac{\sum_{i \neq j} t_i y_i}{n_1}$$
$$n_0 = \sum_{i \neq j} (1 - t_i) \qquad \bar{y}_0 = \frac{\sum_{i \neq j} (1 - t_i) y_i}{n_0} \tag{26}$$

From this, the difference of means estimate, when the j-th observation is treated can be written as:

$$\bar{y}_{1-0}(y_j|t_j=1) = \frac{y_j(1) + \sum_{i\neq j} t_i y_i}{n_1+1} - \frac{\sum_{i\neq j}(1-t_i)y_i}{n_0} \tag{27}$$

And similarly when the j-th observation is control:

$$\bar{y}_{1-0}(y_j|t_j=0) = \frac{\sum_{i\neq j} t_i y_i}{n_1} - \frac{y_j(0) + \sum_{i\neq j}(1-t_i)y_i}{n_0+1} \tag{28}$$

A.2 Sensitivity of the difference of means estimator

Theorem A.1 *The sensitivity of the difference of means estimator, $\bar{y}_{1-0}$, among n_1 and n_0 treatment and control observations:*

$$\Delta\bar{y}_{1-0} = \frac{y_{max} - y_{min}}{n_1+1} + \frac{y_{max} - y_{min}}{n_0+1} \tag{29}$$

Proof If the j-th observation remains treated, following equation 3, its largest effect on the difference of means estimate is:

$$C = \max_{y_j, y_j'} \bar{y}_{1-0}(y_j, t_j=1) - \bar{y}_{1-0}(y_j', t_j'=1) = \max_{y_j, y_j'} \frac{y_j - y_j'}{n_1+1} = \frac{y_{\max} - y_{\min}}{n_1+1} \tag{30}$$

Similarly, if it remains untreated, its largest effect is:

$$D = \frac{y_{\max} - y_{\min}}{n_0+1} \tag{31}$$

If the j-th observation moves from treatment to control, differencing equations 27 and 28 can be reduced to:

$$A - B = \frac{y_j(1)}{n_1+1} + \frac{y_j(0)}{n_0+1} + \frac{[n_1 - (n_1+1)]\sum_{i\neq j} t_i y_i}{n_1(n_1+1)} - \frac{[(n_0+1) - n_0]\sum_{i\neq j}(1-t_i)y_i}{n_0(n_0+1)} \tag{32}$$

$$= \frac{y_j(1)}{n_1+1} + \frac{y_j(0)}{n_0+1} + \frac{-n_1\bar{y}_1}{n_1(n_1+1)} + \frac{-n_0\bar{y}_0}{n_0(n_0+1)} \tag{33}$$

$$= \frac{y_j(1) - \bar{y}_1}{n_1+1} + \frac{y_j(0) - \bar{y}_0}{n_0+1} \tag{34}$$

Sensitivity of the difference of means test is then:

$$\Delta\bar{y}_{1-0} = \max_{y_j(1),\ y_j(0)} \{|A-B|, C, D\} \tag{35}$$

Examining $|A - B|$, the left and right terms of equation 34 have upper bounds:

$$\max_{y_j(1),\bar{y}_1} \left| \frac{y_j(1) - \bar{y}_1}{n_1 + 1} \right| = \frac{y_{\max} - y_{\min}}{n_1 + 1} = C \tag{36}$$

$$\max_{y_j(0),\bar{y}_0} \left| \frac{y_j(0) - \bar{y}_0}{n_0 + 1} \right| = \frac{y_{\max} - y_{\min}}{n_0 + 1} = D \tag{37}$$

Therefore the sharp bound on the sensitivity is:

$$\Delta\bar{y}_{1-0} \leq C + D \qquad \blacksquare \tag{38}$$

A.3 Sensitivity of the variance and standard errors

Given sample mean and standard errors:

$$\bar{y} = \frac{\sum_i y_i}{n}; \qquad s = \frac{\sum_i (y_i - \bar{y})^2}{n}, \tag{39}$$

the Laguerre-Samuelson inequality gives absolute bounds on the location of any sample observation as:

$$\bar{y} - s\sqrt{n-1} \leq y_j \leq \bar{y} + s\sqrt{n-1} \tag{40}$$

with equality only in the limiting case where all observations but j are identical. We first rewrite this in terms of s as:

$$s \geq \frac{|\bar{y} - y_j|}{\sqrt{n-1}} = \frac{\left| \frac{(n-1)\bar{y}_{-j} + y_j}{n} - y_j \right|}{\sqrt{n-1}} = \frac{\sqrt{n-1}}{n} \left| \bar{y}_{-j} - y_j \right|, \qquad \forall y_j \tag{41}$$

where $\bar{y}_j$ is the mean of all but the jth observation. We might intuitively reason that the maximum effect one observation can have on s is when all the observations are at one bound, and then one observation is moved to the opposite bound. In this case, s goes from zero to the maximum of equation A.3 under equality, and the sensitivity is $\sqrt{n-1}/n$ times the range of the data.

We now formally prove this intuitive justification.

Theorem A.2 *The sensitivity of the sample estimated variance of N observations is:*

$$\Delta s^2 = \frac{n-1}{n^2} \Big(y_{max} - y_{min} \Big)^2 \tag{42}$$

Proof We can rewrite the sample variance as:

$$s^2 = \frac{1}{n}\sum_i (\bar{y} - y_i)^2 \tag{43}$$

$$= \bar{y}^2 - 2\bar{y}\frac{1}{n}\sum_i y_i - \frac{1}{n}\sum_i y_i^2 \tag{44}$$

$$= \left(\frac{1}{n}\sum_i y_i^2\right) - \bar{y}^2 \tag{45}$$

$$= \frac{1}{n}\sum_i y_i^2 - \left(\frac{1}{n}\sum_i y_i\right)^2 \tag{46}$$

$$= \frac{1}{n}\left[y_j^2 + \sum_{i\neq j} y_i^2\right] - \frac{1}{n^2}\left(y_j + \sum_{i\neq j} y_i\right)^2 \tag{47}$$

$$= \frac{1}{n}\left[y_j^2 + \sum_{i\neq j} y_i^2\right] - \frac{1}{n^2}\left(y_j^2 + 2y_j\sum_{i\neq j} y_i + \Big(\sum_{i\neq j} y_i\Big)^2\right) \tag{48}$$

$$= \frac{1}{n}\left[\sum_{i\neq j} y_i^2 - \frac{1}{n}\Big(\sum_{i\neq j} y_i\Big)^2\right] + \frac{1}{n^2}\left((n-1)y_j^2 - 2y_j\sum_{i\neq j} y_i\right) \tag{49}$$

$$= \left[\frac{1}{n}\sum_{i\neq j} y_i^2 - \Big(\frac{1}{n}\sum_{i\neq j} y_i\Big)^2\right] + \frac{n-1}{n^2}\left(y_j^2 - 2y_j\bar{y}_{-j}\right) \tag{50}$$

where $\bar{y}_{-j}$ is the mean after removing the jth observation:

$$\bar{y}_{-j} = \frac{1}{n-1}\sum_{i\neq j} y_i \tag{51}$$

The first term is not a function of y_j (and resembles the variance of the dataset after omitting jth observation, except for the incorrect scaling by n). This gives us partials with respect to the jth observation of:

$$\frac{\delta s^2}{\delta y_j} = \frac{2(n-1)}{n^2}\Big(y_j - \bar{y}_{-j}\Big) \tag{52}$$

$$\frac{\delta^2 s^2}{\delta y_j^2} = \frac{2(n-1)}{n^2} > 0 \tag{53}$$

Therefore:

$$\arg\min_{y_j} s^2 = \bar{y}_{-j} \tag{54}$$

$$\arg\max_{y_j} s^2 = \begin{cases} y_{\min}, & \text{if } |y_{\min} - \bar{y}_{-j}| \geq |y_{\max} - \bar{y}_{-j}| \\ y_{\max}, & \text{if } |y_{\min} - \bar{y}_{-j}| < |y_{\max} - \bar{y}_{-j}| \end{cases} \tag{55}$$

Thus across all possible values of y_j, the variance is minimized when y_j is at the mean of the rest of the data, and the variance is maximized when y_j is at the bound farthest from the mean.

This gives us the difference in the variance from changing the j-th observation from y_j to y'_j as:

$$s^2(y_j) - s^2(y'_j) = \frac{n-1}{n^2}\left(y_j^2 - y_j'^2 - 2(y_j - y'_j)\bar{y}_{-j}\right) \tag{56}$$

And thus sensitivity of the variance as:

$$\begin{aligned}
\Delta s^2 = \max_{y_j, y'_j}\left[s^2(y_j) - s^2(y'_j)\right] &= \max_{y_j, \bar{y}_{-j}} \frac{n-1}{n^2}\left(y_j^2 - \bar{y}_{-j}^2 - 2(y_j - \bar{y}_{-j})\bar{y}_{-j}\right) && (57)\\
&= \max_{y_j, \bar{y}_{-j}} \frac{n-1}{n^2}\left(y_j^2 + \bar{y}_{-j}^2 - 2y_j\bar{y}_{-j}\right) && (58)\\
&= \max_{y_j, \bar{y}_{-j}} \frac{n-1}{n^2}\left(y_j - \bar{y}_{-j}\right)^2 && (59)\\
&= \frac{n-1}{n^2}\left(y_{\max} - y_{\min}\right)^2. \quad \blacksquare && (60)
\end{aligned}$$

It is not generally the case that $\Delta\sqrt{f} = \sqrt{\Delta f}$ for the same reasons that it is not generally true that $\sqrt{a^2 - b^2} = a - b$. Thus sensitivity of the standard deviation need not follow directly as the square-root of the sensitivity of the variance. However, in this special case it does:

Lemma A.3 *The sensitivity of the sample standard deviation, s, is:*

$$\Delta s = \frac{\sqrt{n-1}}{n}\left(y_{max} - y_{min}\right) \tag{61}$$

Proof Since $s \geq 0$, $\forall Y$ then equations 54 and 55 also hold for s. As $\min_{Y_j, Y_{-j}}[s] = 0$ then $\Delta s^2 = \max_{y_j, y'_j}\left[s^2(y_j) - s^2(y'_j)\right] = \max\left[s^2(y_j) - 0\right] \Rightarrow \Delta s = \max\left[s(y_j) - 0\right] = \sqrt{\Delta s^2}$.

Lemma A.4 *The sensitivity of the standard error of the mean, s is:*

$$\Delta s = \sqrt{\frac{n-1}{n^3}}\left(y_{max} - y_{min}\right) \tag{62}$$

Proof The standard error of the mean, $se = s/\sqrt{n}$ is simply postprocessing of s by a known constant, and $\Delta c\ f(x) = c\Delta f(x)$ for constant c.

Lemma A.5 *The sensitivity of the standard error, s_{1-0}, of the difference of means test among n_1 and n_0 treatment and control observations is:*

$$\Delta s_{1-0} = \sqrt{\frac{N^* - 1}{N^{*3}}}\left(y_{max} - y_{min}\right), \qquad \textit{where } N^* = min\left(n_0, n_1\right) \tag{63}$$

Proof If we consider worst-case movement of some observation x_j to x'_j, there are three possible cases for treatment, each of which have their own sensitivity, $(1)\Delta f_1 : t_j = t'_j = 1; (2)\Delta f_2 : t_j = t'_j = 0; (3)\Delta f_3 : t_j = 1 - t'_j$. Lemma A.4, gives Δf_1 and Δf_2, after adjusting n to n_1 or n_0, respectively. Let $s(x_i)$ and $s(x_i, x_j)$ be the standard error of observations with treatment i, with and without x_j in that group. $\Delta f_3 = \max_{x_j,i}[(s^2(x_i, x_j) + s^2(x_{\sim i}))^{1/2} - \min_{x_j}(s^2(x_i) + s^2(x_{\sim i}, x_j))^{1/2}] = \max_{x_j,i}(s^2(x_i, x_j) + s^2(x_{\sim i}))^{1/2} - 0 = \max(\Delta f_1, \Delta f_2)$. So $\max(\Delta f_1, \Delta f_2, \Delta f_3) = \max(\Delta f_1, \Delta f_2)$ where $\Delta f_1 > \Delta f_2 \iff n_1 < n_0$.

Theorem A.6 *If a dataset, D, is constructed by matched pairs, and then a function f with sensitivity Δf is computed on the pair-matched data, then the sensitivity of the entire operation (matching and computing the function) is at most $3\Delta f$.*

Proof We set out a proof by exhaustion. After paired matching, define $n_0(X)$ (and $n_1(X)$) as the number of control (treatment) observations with a matched treatment (control) observation in dataset (X), and $n(X) = n_1(X) + n_0(X) (= 2n_1(X) = 2n_0(X))$. Consider one observation $x_j = (y_j, t_j)$ that can be manipulated to $x'_j = (y'_j, t'_j)$. x_j is either initially matched or unmatched, in which case $n(X_{-j}) \in \{n(X), n(X) - 2\}$. x'_j is either matchable or unmatchable. If matchable it either replaces an observation in an existing match, or forms a match with a previously unmatched observation. thus $n(X, x'_j) \in \{n(X), n(X) + 2\}$, therefore $n(X_{-j}, x'_j) \in \{n(X) - 2, n(X), n(X) + 2\}$. In summary, the exhaustive set of possibilities is:

	x_j		x'_j		
	matched	unmatched	new match	replacement	unmatched
observations added	0	0	2	1	0
observations removed	2	0	0	1	0

Any manipulation of x_j to x'_j must result in one left column outcome and one right column outcome. Global sensitivity upper bounds the effect of arbitrarily manipulating one observation, which is removing one observation and replacing it with another. Two observations removed and two added is the same as manipulating two observations in the dataset, and thus has worst case effect of twice the sensitivity. Three observations removed and one added is less change to the data than manipulating three observations, and thus has an effect bounded by three times the sensitivity of the final function.

B Algorithm for Differentially Private Standard Errors

Following closely Smith [17], but in the context of standard errors, the algorithm is:

Algorithm 4: Differentially Private Standard Errors of Difference of Mean Estimates

1. Divide the dataset into M subsets, $X_1, \ldots, X_M$.
2. **for** i in 1:M Calculate $s_i = \sqrt{\frac{\text{sd}(x_1)^2}{n_1} + \frac{\text{sd}(x_0)}{n_0}}$
3. $\hat{a} \leftarrow \text{PrivateQuantile}(S, \frac{1}{4}, \frac{\epsilon}{4}, \Lambda)$
4. $\hat{b} \leftarrow \text{PrivateQuantile}(S, \frac{3}{4}, \frac{\epsilon}{4}, \Lambda)$
5. $\hat{\mu} = (\hat{a} + \hat{b})/2$
6. $\hat{iqr} = |\hat{a} - \hat{b}|$
7. $u = \hat{\mu} + 2\hat{iqr}$
8. $l = \hat{\mu} - 2\hat{iqr}$
9. Define $\Pi_{[l,u]} = \begin{cases} l & \text{if } x < l \\ x & \text{if } l \leq x \leq u \\ u & \text{if } x > u \end{cases}$
10. Calculate $w = \frac{1}{M}\Sigma_{i=1}^{M}\Pi_{[l,u]}(X_i)$
11. Draw $Y \sim f_{\text{Laplace}}(\mu = 0, b = \frac{|u-l|}{2\epsilon k})$
12. Release $M(X) = w + Y$

Algorithm 5: PrivateQuantile(Z, α, ϵ)

1. Sort Z ascending
2. Replace $Z_i < 0$ with 0, and $Z_i > \Lambda$ with Λ.
3. Define $Z_0 = 0$ and $Z_{k+1} = \Lambda$
4. **For** i in 1:k set $y_i = (Z_{i+1} - Z_i)exp(-\epsilon|i - \alpha k|)$
5. Sample an integer $i \in \{0, \cdots, k\}$ with probability $y_i/\Sigma_{i=0}^{k} y_i$
6. Output a uniform draw $Z_{i+1} - Z_i$.