2016 IEEE 57th Annual Symposium on Foundations of Computer Science

Max-Information, Differential Privacy, and Post-Selection Hypothesis Testing

Ryan Rogers*, Aaron Roth†, Adam Smith‡ and Om Thakkar‡
*Applied Math and Computational Sciences
University of Pennsylvania, Philadelphia, PA USA
E-mail: ryrogers@sas.upenn.edul
†Computer and Information Science
University of Pennsylvania, Philadelphia, PA USA
E-mail: aaroth@cis.upenn.edu
‡Computer Science and Engineering Department
The Pennsylvania State University, University Park, PA USA
E-mail: {asmith,omthkkr}@cse.psu.edu

***Abstract*—In this paper, we initiate a principled study of how the generalization properties of approximate differential privacy can be used to perform adaptive hypothesis testing, while giving statistically valid p-value corrections. We do this by observing that the guarantees of algorithms with bounded approximate *max-information* are sufficient to correct the p-values of adaptively chosen hypotheses, and then by proving that algorithms that satisfy (ϵ, δ)-differential privacy have bounded approximate max-information when their inputs are drawn from a product distribution. This substantially extends the known connection between differential privacy and max-information, which previously was only known to hold for (pure) $(\epsilon, 0)$-differential privacy. It also extends our understanding of max-information as a partially unifying measure controlling the generalization properties of adaptive data analyses. We also show a lower bound, proving that (despite the strong composition properties of max-information), when data is drawn from a product distribution, (ϵ, δ)-differentially private algorithms can come *first* in a composition with other algorithms satisfying max-information bounds, but not necessarily second if the composition is required to itself satisfy a nontrivial max-information bound. This, in particular, implies that the connection between (ϵ, δ)-differential privacy and max-information holds only for inputs drawn from product distributions, unlike the connection between $(\epsilon, 0)$-differential privacy and max-information.**

I. Introduction

Adaptive Data Analysis refers to the reuse of data to perform analyses suggested by the outcomes of previously computed statistics on the same data. It is the common case when *exploratory data analysis* and *confirmatory data analysis* are mixed together, and both conducted on the same dataset. It models both well-defined, self-contained tasks, like selecting a subset of variables using the LASSO and then fitting a model to the selected variables, and also much harder-to-specify sequences of analyses, such as those that occur when the same dataset is shared and reused by multiple researchers.

Recently two lines of work have arisen, in statistics and computer science respectively, aimed at rigorous statistical understanding of adaptive data analysis. By and large, the goal in the statistical literature (often called "selective" or "post-selection" inference [1]) is to derive valid hypothesis tests and tight confidence intervals around parameter values that arise from very specific analyses, such as LASSO model selection followed by least squares regression (see e.g. [2, 3]). In contrast, the second line of work has aimed for generality (at the possible expense of giving tight application-specific bounds). This second literature imposes conditions on the algorithms performing each stage of the analysis, and makes no other assumptions on how, or in what sequence, the results are used by the data analyst. Two algorithmic constraints that have recently been shown to guarantee that future analyses will be statistically valid are differential privacy [4, 5] and bounded output description length, which are partially unified by a measure of information called *max-information* [6]. This paper falls into the second line of research—specifically, we extend the connection made in [4, 5] between *differential privacy* and the adaptive estimation of *low-sensitivity queries* to a more general setting that includes adaptive hypothesis testing with statistically valid p-values.

Our main technical contribution is a quantitatively tight connection between differential privacy and a *max-information*. Max-information is a measure of correlation, similar to Shannon's mutual information, which allows bounding the change in the conditional probability of events relative to their a priori probability. Specifically, we extend a bound on the max-information of $(\epsilon, 0)$-differentially private algorithms, due to [6], to the much larger class of (ϵ, δ)-differentially private algorithms.

A. Post-Selection Hypothesis Testing

To illustrate an application of our results, we consider a simple model of one-sided hypothesis tests on real valued test statistics. Let $\mathcal{X}$ denote a data domain. A *dataset* $\mathbf{x}$ consists of n elements in $\mathcal{X}$: $\mathbf{x} \in \mathcal{X}^n$. A hypothesis test is defined by a *test statistic* $\phi_i : \mathcal{X}^n \to \mathbb{R}$, where we use i to index different test statistics. Given an output $a = \phi_i(\mathbf{x})$,

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.59

together with a distribution $\mathcal{P}$ over the data domain, the p-value associated with a and $\mathcal{P}$ is simply the probability of observing a value of the test statistic that is at least as extreme as a, assuming the data was drawn independently from $\mathcal{P}$: $p_i^{\mathcal{P}}(a) = \Pr_{\mathbf{X}\sim\mathcal{P}^n}[\phi_i(\mathbf{X}) \geq a]$. Note that there may be multiple distributions $\mathcal{P}$ over the data that induce the same distribution over the test statistic. With each test statistic ϕ_i, we associate a *null hypothesis* $H_0^{(i)} \subseteq \Delta(\mathcal{X})$,[1] which is simply a collection of such distributions. The p-values are always computed with respect to a distribution $\mathcal{P} \in H_0^{(i)}$, and hence from now on, we elide the dependence on $\mathcal{P}$ and simply write $p_i(a)$ to denote the p-value of a test statistic ϕ_i evaluated at a.

The goal of a hypothesis test is to *reject the null hypothesis* if the data is not likely to have been generated from the proposed model, that is, if the underlying distribution from which the data were drawn was not in $H_0^{(i)}$. By definition, if $\mathbf{X}$ truly is drawn from $\mathcal{P}^n$ for some $\mathcal{P} \in H_0^{(i)}$, then $p_i(\phi_i(\mathbf{X}))$ is uniformly distributed over $[0, 1]$. A standard approach to hypothesis testing is to pick a *significance level* $\alpha \in [0, 1]$ (often $\alpha = 0.05$), compute the value of the test statistic $a = \phi_i(\mathbf{X})$, and then *reject* the null hypothesis if $p_i(a) \leq \alpha$. Under this procedure, the probability of incorrectly rejecting the null hypothesis—i.e., of rejecting the null hypothesis when $\mathbf{X} \sim \mathcal{P}^n$ for some $\mathcal{P} \in H_0^{(i)}$—is at most α. An incorrect rejection of the null hypothesis is called a *false discovery*.

The discussion so far presupposes that ϕ_i, the test statistic in question, was chosen independently of the dataset $\mathbf{X}$. Let $\mathcal{T}$ denote a collection of test statistics, and suppose that we select a test statistic using a data-dependent selection procedure $\mathcal{A} : \mathcal{X}^n \to \mathcal{T}$. If $\phi_i = \mathcal{A}(\mathbf{X})$, then rejecting the null hypothesis when $p_i(\phi_i(\mathbf{X})) \leq \alpha$ may result in a false discovery with probability much larger than α (indeed, this kind of naive approach to *post-selection* inference is suspected to be a primary culprit behind the prevalence of false discovery in empirical science [7, 8, 9]). This is because even if the null hypothesis is true ($\mathbf{X} \sim \mathcal{P}^n$ for some $\mathcal{P} \in H_0^{(i)}$), the distribution on $\mathbf{X}$ *conditioned on* $\phi_i = \mathcal{A}(\mathbf{X})$ *having been selected* need not be $\mathcal{P}^n$. Our goal in studying valid post-selection hypothesis testing is to find a *valid* p-value correction function $\gamma : [0, 1] \to [0, 1]$, which we define as follows:

Definition I.1 (Valid p-value Correction Function)**.** *A function $\gamma : [0, 1] \to [0, 1]$ is a* valid p-value correction function *for a selection procedure $\mathcal{A} : \mathcal{X}^n \to \mathcal{T}$ if for every significance level $\alpha \in [0, 1]$, the procedure:*

1) *Select a test statistic $\phi_i = \mathcal{A}(\mathbf{X})$ using selection procedure $\mathcal{A}$.*
2) *Reject the null hypothesis $H_0^{(i)}$ if $p_i(\phi_i(\mathbf{X})) \leq \gamma(\alpha)$.*

has probability at most α of resulting in a false discovery.

[1] $\Delta(\mathcal{X})$ denotes the set of probability distributions over $\mathcal{X}$.

Necessarily, to give a nontrivial correction function γ, we will need to assume that the selection procedure $\mathcal{A}$ satisfies some useful property. In this paper, we focus on *differential privacy*, which is a measure of algorithmic stability, and more generally, max-information, which is defined in the next subsection. Differential privacy is of particular interest because it is closed under post-processing and satisfies strong composition properties. This means that, if the test statistics in $\mathcal{T}$ are themselves differentially private, then the selection procedure $\mathcal{A}$ can represent the decisions of a worst-case data analyst, who chooses which hypothesis tests to run in arbitrary ways as a function of the outcomes of previously selected tests.

Finally, we note that, despite the fact that previous works [4, 5] are explicitly motivated by the problem of false discovery in empirical science, most of the technical results to date have been about estimating the means of adaptively chosen predicates on the data (i.e., answering *statistical queries*) [4], and more generally, estimating the values of low-sensitivity (i.e., Lipschitz continuous) functions on the dataset [5, 10, 11]. These kinds of results do not apply to the problem of adaptively performing hypothesis tests while generating statistically valid p-values, because p-values are by definition not low-sensitivity statistics. See the full version for a detailed discussion.

There is one constraint on the selection procedure $\mathcal{A}$ that does allow us to give nontrivial p-value corrections—that $\mathcal{A}$ should have bounded max-information. A condition of bounded mutual information has also been considered [10] to give p-value corrections - but as we discuss in the full version, it is possible to obtain a strictly stronger guarantee by instead reasoning via max-information. Max-information is a measure introduced by [6], which we discuss next.

B. Max-Information (and p-values)

Given two (arbitrarily correlated) random variables X, Z, we let $X \otimes Z$ denote a random variable (in a different probability space) obtained by drawing independent copies of X and Z from their respective marginal distributions. We write $\log$ to denote logarithms base 2.

Definition I.2 (Max-Information [6])**.** *Let X and Z be jointly distributed random variables over the domain $(\mathcal{X}, \mathcal{Z})$. The max-information between X and Z, denoted by $I_\infty(X; Z)$, is the minimal value of k such that for every x in the support of X and z in the support of Z, we have $\Pr[X = x|Z = z] \leq 2^k \Pr[X = x]$. Alternatively,*

$$I_\infty(X; Z) = \log \sup_{(x,z)\in(\mathcal{X},\mathcal{Z})} \frac{\Pr[(X, Z) = (x, z)]}{\Pr[X \otimes Z = (x, z)]}.$$

The β-approximate max-information between X and Z is defined as

$$I_\infty^\beta(X; Z) = \log \sup_{\substack{\mathcal{O}\subseteq(\mathcal{X}\times\mathcal{Z}),\\ \Pr[(X,Z)\in\mathcal{O}]>\beta}} \frac{\Pr[(X, Z) \in \mathcal{O}] - \beta}{\Pr[X \otimes Z \in \mathcal{O}]}.$$

We say that an algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ has β-approximate max-information of k, denoted as $I^{\beta}_{\infty}(\mathcal{A}, n) \leq k$, if for every distribution $\mathcal{S}$ over elements of $\mathcal{X}^n$, we have $I^{\beta}_{\infty}(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq k$ when $\mathbf{X} \sim \mathcal{S}$. We say that an algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ has β-approximate max-information of k over product distributions*, written $I^{\beta}_{\infty,P}(\mathcal{A}, n) \leq k$, if for every distribution $\mathcal{P}$ over $\mathcal{X}$, we have $I^{\beta}_{\infty}(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq k$ when $\mathbf{X} \sim \mathcal{P}^n$.*

It follows immediately from the definition that if an algorithm has bounded max-information, then we can control the probability of "bad events" that arise as a result of the dependence of $\mathcal{A}(\mathbf{X})$ on $\mathbf{X}$: for every event $\mathcal{O}$, we have $\Pr[(\mathbf{X}, \mathcal{A}(\mathbf{X})) \in \mathcal{O}] \leq 2^k \Pr[\mathbf{X} \otimes \mathcal{A}(\mathbf{X}) \in \mathcal{O}] + \beta$. For example, if $\mathcal{A}$ is a data-dependent selection procedure for selecting a test statistic, we can derive a valid p-value correction function γ as a function of a max-information bound on $\mathcal{A}$:

Theorem I.3. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{T}$ be a data-dependent algorithm for selecting a test statistic such that $I^{\beta}_{\infty,P}(\mathcal{A}, n) \leq k$. Then the following function γ is a valid p-value correction function for $\mathcal{A}$:*

$$\gamma(\alpha) = \max\left(\frac{\alpha - \beta}{2^k}, 0\right).$$

Proof: Fix a distribution $\mathcal{P}^n$ from which the dataset $\mathbf{X}$ is drawn. If $\frac{\alpha-\beta}{2^k} \leq 0$, then the theorem is trivial, so assume otherwise. Define $\mathcal{O} \subset \mathcal{X}^n \times \mathcal{T}$ to be the event that $\mathcal{A}$ selects a test statistic for which the null hypothesis is true, but its p-value is at most $\gamma(\alpha)$:

$$\mathcal{O} = \{(\mathbf{x}, \phi_i) : \mathcal{P} \in H_0^{(i)} \text{ and } p_i(\phi_i(\mathbf{x})) \leq \gamma(\alpha)\}$$

Note that the event $\mathcal{O}$ represents exactly those outcomes for which using γ as a p-value correction function results in a false discovery. Note also that, by definition of the null hypothesis, $\Pr[\mathbf{X} \otimes \mathcal{A}(\mathbf{X}) \in \mathcal{O}] \leq \gamma(\alpha) = \frac{\alpha-\beta}{2^k}$. Hence, by the guarantee that $I^{\beta}_{\infty,P}(\mathcal{A}, n) \leq k$, we have that $\Pr[(\mathbf{X}, \mathcal{A}(\mathbf{X}) \in \mathcal{O})]$ is at most $2^k \cdot \left(\frac{\alpha-\beta}{2^k}\right) + \beta = \alpha$. ■

Because of Theorem I.3, we are interested in methods for usefully selecting test statistics using data dependent algorithms $\mathcal{A}$ for which we can bound their max-information. It was shown in [6] that algorithms which satisfy *pure* differential privacy also have a guarantee of bounded max-information:

Theorem I.4 (Pure Differential Privacy and Max-Information [6])**.** *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an $(\epsilon, 0)$-differentially private algorithm. Then for every $\beta \geq 0$:*

$$I_{\infty}(\mathcal{A}, n) \leq \log(e) \cdot \epsilon n, \text{ and}$$

$$I^{\beta}_{\infty,P}(\mathcal{A}, n) \leq \log(e) \cdot \left(\epsilon^2 n/2 + \epsilon\sqrt{n \ln(2/\beta)/2}\right)$$

This connection is powerful, because there are a vast collection of data analyses for which we have differentially private algorithms— including a growing literature on differentially private hypothesis tests [12, 13, 14, 15, 16, 17, 18, 19]. However, there is an important gap: Theorem I.4 holds only for *pure* $(\epsilon, 0)$-differential privacy, and not for approximate (ϵ, δ)-differential privacy. Many statistical analyses can be performed much more accurately subject to approximate differential privacy, and it can be easier to analyze private hypothesis tests that satisfy approximate differential privacy, because the approximate privacy constraint is amenable to perturbations using Gaussian noise (rather than Laplace noise) [19]. Most importantly, for pure differential privacy, the privacy parameter ϵ degrades *linearly* with the number of analyses performed, whereas for approximate differential privacy, ϵ need only degrade with the *square root* of the number of analyses performed [20]. Hence, if the connection between max-information and differential privacy held also for approximate differential privacy, it would be possible to perform quadratically more adaptively chosen statistical tests without requiring a larger p-value correction factor.

C. Our Results

In addition to the framework just described for reasoning about adaptive hypothesis testing, our main technical contribution is to extend the connection between differential privacy and max-information to approximate differential privacy. We show the following (see Section III for a complete statement):

Theorem III.1 (Informal). *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an (ϵ, δ)-differentially private algorithm. Then,*

$$I^{\beta}_{\infty,P}(\mathcal{A}, n) = O\left(n\epsilon^2 + n\sqrt{\tfrac{\delta}{\epsilon}}\right) \quad \text{for } \beta = O\left(n\sqrt{\tfrac{\delta}{\epsilon}}\right).$$

It is worth noting several things. First, this bound nearly matches the bound for max-information over product distributions from Theorem I.4, except Theorem III.1 extends the connection to the substantially more powerful class of (ϵ, δ)-differentially private algorithms. The bound is qualitatively tight in the sense that despite its generality, it can be used to nearly recover the tight bound on the generalization properties of differentially private mechanisms for answering low-sensitivity queries that was proven using a specialized analysis in [5] (see the full version for a comparison).

We also only prove a bound on the max-information for product distributions on the input, and not for all distributions (that is, we bound $I^{\beta}_{\infty,P}(\mathcal{A}, n)$ and not $I^{\beta}_{\infty}(\mathcal{A}, n)$). A bound for general distributions would be desirable, since such bounds compose gracefully [6]. Unfortunately, a bound for general distributions based solely on (ϵ, δ)-differential privacy is impossible: a construction of De [21] implies the existence of (ϵ, δ)-differentially private algorithms for which

the max-information between input and output on arbitrary distributions is much larger than the bound in Theorem III.1.

One might nevertheless hope that bounds on the max-information under product distributions can be meaningfully composed. Our second main contribution is a negative result, showing that such bounds do not compose when algorithms are selected adaptively. Specifically, we analyze the adaptive composition of two algorithms, the first of which has a small finite range (and hence, by [6], small bounded max-information), and the second of which is (ϵ, δ)-differentially private. We show that the composition of the two algorithms can be used to exactly recover the input dataset, and hence, the composition does not satisfy any nontrivial max-information bound.

1) Further Interpretation: Although our presentation thus far has been motivated by p-values, an algorithm $\mathcal{A}$ with bounded max-information allows a data analyst to treat *any event* $\mathcal{A}(\mathbf{x})$ that is a function of the output of the algorithm "as if" it is independent of the dataset $\mathbf{x}$, up to a correction factor determined by the max-information bound. Our results thus substantially broaden the class of analyses for which approximate differential privacy promises generalization guarantees—this class was previously limited to estimating the values of low-sensitivity numeric valued queries (and more generally, the outcomes of low-sensitivity optimization problems) [5].

Our result also further develops the extent to which max-information can be viewed as a unifying information theoretic measure controlling the generalization properties of adaptive data analysis. Dwork et al. [6] previously showed that algorithms that satisfy bounded output description length, and algorithms that satisfy pure differential privacy (two constraints known individually to imply adaptive generalization guarantees), both have bounded max-information. Because bounded max-information satisfies strong composition properties, this connection implies that algorithms with bounded output description length and pure differentially private algorithms can be composed in arbitrary order and the resulting composition will still have strong generalization properties. Our result brings approximate differential privacy partially into this unifying framework. In particular, *when the data is drawn from a product distribution*, if an analysis that starts with an (arbitrary) approximate differentially private computation is followed by an arbitrary composition of algorithms with bounded max-information, then the resulting composition will satisfy a max-information bound. However, unlike with compositions consisting solely of bounded description length mechanisms and pure differentially private mechanisms, which can be composed in arbitrary order, in this case *it is important that the approximate differentially private computation come first*. This is because, even if the dataset $\mathbf{x}$ is initially drawn from a product distribution, the conditional distribution on the data that results after observing the outcome of an initial computation need not be a product distribution any longer. In fact, the lower bound we prove in Section IV is an explicit construction in which the composition of a bounded description length algorithm, followed by an approximate differentially private algorithm can be used to exactly reconstruct a dataset drawn from a product distribution (which can in turn be used to arbitrarily overfit that dataset).

Finally, we draw a connection between max-information and mutual information that allows us to improve on several prior results that dealt with mutual information [10, 22] . We present the proof in the full version.

Lemma I.5. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{T}$ be a selection rule.*

- *If $I(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq m$ and $\mathbf{X} \sim \mathcal{S}$ for any distribution over $\mathcal{X}^n$, then for any $k > 0$, $I_\infty^{\beta(k)}(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq k$ for $\beta(k) \leq \frac{m+0.54}{k}$.*
- *If $I_\infty^{\beta}(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq k$ for $\beta \in [0, 0.3]$, then $I(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq 2k\ln(2) + \beta n \log_2 |\mathcal{X}| + \beta \ln(1/\beta)$.*

We are able to improve on the p-value correction function implicitely given in [10] given a mutual information bound, by first converting mutual information to a max-information bound and applying the p-value correction function from this paper. Our main theorem Theorem III.1 combined with Lemma I.5 also obtains an improved bound on the mutual information of approximate differentially private mechanisms from Proposition 4.4 in [22]. The following corollary improves the bound from [22] in its dependence on $|\mathcal{X}|$ from $|\mathcal{X}|^2 \cdot \log(1/|\mathcal{X}|)$ to $\log |\mathcal{X}|$.

Corollary I.6. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{T}$ and $\mathbf{X} \sim \mathcal{P}^n$. If $\epsilon \in (0, 1/2]$, $\delta = \Omega\left(\frac{\epsilon e^{-2n\epsilon^2}}{n^2}\right)$ and $\delta = O(\frac{\epsilon}{n^2})$, we then have*

$$I(\mathbf{X}; \mathcal{A}(\mathbf{X})) = O\left(n\epsilon^2 + n\sqrt{\tfrac{\delta}{\epsilon}}\left(1 + \ln\left(\tfrac{1}{n}\sqrt{\tfrac{\epsilon}{\delta}}\right) + n\log|\mathcal{X}|\right)\right).$$

D. Other Related Work

Differential Privacy is an algorithmic stability condition introduced by Dwork et al. [23]. Its connection to adaptive data analysis was made by Dwork et al. [4] and both strengthened and generalized by Bassily et al. [5]. Dwork et al. [6] showed that algorithms with bounded description length outputs have similar guarantees for adaptive data analysis, and introduced the notion of max-information. Cummings et al. [24] give a third method—compression schemes—which can also guarantee validity in adaptive data analysis in the context of learning. Computational and information theoretic lower bounds for adaptively estimating means in this framework were proven by Hardt and Ullman [25], and Steinke and Ullman [26].

Russo and Zou [10] show how to bound the *bias* of sub-gaussian statistics selected in a data-dependent manner, in terms of the mutual information between the selection procedure and the value of the statistics. In particular (using

our terminology), they show how to give a valid p-value correction function in terms of this mutual information. In the full version , we demonstrate that if a bound on the mutual information between the dataset and the output of the selection procedure is known, then it is possible to substantially improve on the p-value correction function given by [10] by instead using the mutual information bound to prove a max-information bound on the selection procedure. [11] study adaptive data analysis in a similar framework to [10], and give a minimax analysis in a restricted setting.

McGregor et al. [22], and De [21] also study (among other things) information theoretic bounds satisfied by differentially private algorithms. Together, they prove a result that is analogous to ours, for *mutual information*—that while pure differentially private algorithms have bounded mutual information between their inputs and their outputs, a similar bound holds for (approximate) (ϵ, δ)-differentially private algorithms only if the data is drawn from a product distribution.

II. Preliminaries

We will use the following vector notation throughout: $\mathbf{x} = (x_1, \cdots, x_n)$, $\mathbf{x}_a^b = (x_a, x_{a+1}, \cdots, x_b)$, $(\mathbf{x}_{-i}, t) = (x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n)$. We denote the distribution of a random variable X as $p(X)$. In our analysis, jointly distributed random variables (X, Z) will typically be of the form $(\mathbf{X}, \mathcal{A}(\mathbf{X}))$ where $\mathbf{X} \sim \mathcal{P}^n$ is a dataset of n elements sampled from domain $\mathcal{X}$, and $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ is a (randomized) algorithm that maps a dataset to some range $\mathcal{Y}$. We denote by $\mathcal{A}(\mathbf{X})$ the random variable that results when $\mathcal{A}$ is applied to a dataset $\mathbf{X} \sim \mathcal{P}^n$ (note that here, the randomness is both over the choice of dataset, and the internal coins of the algorithm). When the input variable is understood, we will sometimes simply write $\mathcal{A}$.

It will be useful in our analysis to compare the distributions of two random variables. In the introduction, we define (approximate-) max-information, and we now give some other measures between distributions. We first define indistinguishability, and then differential privacy.

Definition II.1 (Indistinguishability [27])**.** *Two random variables X, Y taking values in a set $\mathcal{D}$ are (ϵ, δ)-indistinguishable, denoted $X \approx_{\epsilon,\delta} Y$, if for all $\mathcal{O} \subseteq \mathcal{D}$,*

$$\Pr[X \in \mathcal{O}] \leq e^{\epsilon} \cdot \Pr[Y \in \mathcal{O}] + \delta \text{ and}$$

$$\Pr[Y \in \mathcal{O}] \leq e^{\epsilon} \cdot \Pr[X \in \mathcal{O}] + \delta.$$

Definition II.2 (Point-wise indistinguishibility [27])**.** *Two random variables X, Z taking values in a set $\mathcal{D}$ are point-wise (ϵ, δ)-indistinguishable if with probability $1 - \delta$ over $a \sim p(X)$:*

$$e^{-\epsilon} \Pr[Z = a] \leq \Pr[X = a] \leq e^{\epsilon} \Pr[Z = a].$$

Before we define differential privacy, we say that two databases $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ are *neighboring* if they differ in at most one entry. We now define differential privacy in terms of indistinguishability:

Definition II.3 (Differential Privacy [23, 28])**.** *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ is (ϵ, δ)-differentially private if for all neighboring datasets $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$, we have $\mathcal{A}(\mathbf{x}) \approx_{\epsilon,\delta} \mathcal{A}(\mathbf{x}')$.*

In the appendix, we give several useful connections between these definitions along with other more widely known measures between distributions, e.g., KL-divergence, and total-variation distance.

III. Max-Information for (ϵ, δ)-Differentially Private Algorithms

In this section, we prove a bound on approximate max-information for (ϵ, δ)-differentially private algorithms over product distributions.

Theorem III.1. *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an (ϵ, δ)-differentially private algorithm for $\epsilon \in (0, 1/2]$ and $\delta \in (0, \epsilon)$. For $\beta = e^{-\epsilon^2 n} + O\left(n\sqrt{\frac{\delta}{\epsilon}}\right)$, we have*

$$I_{\infty,P}^{\beta}(\mathcal{A}, n) = O\left(\epsilon^2 n + n\sqrt{\tfrac{\delta}{\epsilon}}\right).$$

We will prove Theorem III.1 over the course of this section, using a number of lemmas. We first set up some notation. We will sometimes abbreviate conditional probabilities of the form $\Pr[\mathbf{X} = \mathbf{x} | \mathcal{A} = a]$ as $\Pr[\mathbf{X} = \mathbf{x} | a]$ when the random variables are clear from context. Further, for any $\mathbf{x} \in \mathcal{X}^n$ and $a \in \mathcal{Y}$, we define

$$\begin{aligned} Z(a, \mathbf{x}) &\stackrel{\text{def}}{=} \log\left(\frac{\Pr[\mathcal{A} = a, \mathbf{X} = \mathbf{x}]}{\Pr[\mathcal{A} = a] \cdot \Pr[\mathbf{X} = \mathbf{x}]}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_1^{i-1}\right]}{\Pr[X_i = x_i]}\right) \end{aligned} \tag{1}$$

If we can bound $Z(a, \mathbf{x})$ with high probability over $(a, \mathbf{x}) \sim p(\mathcal{A}(\mathbf{X}), \mathbf{X})$, then we can bound the approximate max-information by using the following lemma:

Lemma III.2 ([6, Lemma 18])**.** *If $\Pr[Z(\mathcal{A}(\mathbf{X}), \mathbf{X}) \geq k] \leq \beta$, then $I_{\infty}^{\beta}(\mathcal{A}(\mathbf{X}); \mathbf{X}) \leq k$.*

We next define each term in the sum of $Z(a, \mathbf{x})$ as

$$Z_i(a, \mathbf{x}_1^i) \stackrel{\text{def}}{=} \log \frac{\Pr\left[X_i = x_i | a, \mathbf{x}_1^{i-1}\right]}{\Pr[X_i = x_i]}. \tag{2}$$

The plan of the proof is simple: our goal is to apply Azuma's inequality to the sum of the Z_i's to achieve a bound on Z with high probability. Applying Azuma's inequality requires both understanding the expectation of each term $Z_i(a, \mathbf{x}_1^i)$, and being able to argue that each term is bounded. Unfortunately, in our case, the terms are not always bounded – however, we will be able to show that they are bounded

with high probability. This plan is somewhat complicated by the conditioning in the definition of $Z_i(a, \mathbf{x}_1^i)$.

First, we argue that we can bound each Z_i with high probability. This argument takes place over the course of Claims III.3, III.4, III.5 and III.6.

Claim III.3. *If $\mathcal{A}$ is (ϵ, δ)-differentially private and $\mathbf{X} \sim \mathcal{P}^n$, then for each $i \in [n]$ and each prefix $\mathbf{x}_1^{i-1} \in \mathcal{X}^{i-1}$, we have:*

$$(\mathcal{A}, X_i)|_{\mathbf{x}_1^{i-1}} \approx_{\epsilon,\delta} \mathcal{A}|_{\mathbf{x}_1^{i-1}} \otimes X_i.$$

We now define the following set of "good outcomes and prefixes" for any $\hat{\delta} > 0$:

$$\mathcal{E}_i(\hat{\delta}) = \left\{(a, \mathbf{x}_1^{i-1}) : X_i \approx_{3\epsilon,\hat{\delta}} X_i|_{a,\mathbf{x}_1^{i-1}}\right\} \tag{3}$$

We use a technical lemma from [27] (stated in the full version), and Claim III.3 to derive the following result:

Claim III.4. *If $\mathcal{A}$ is (ϵ, δ)-differentially private and $\mathbf{X} \sim \mathcal{P}^n$, then for each $i \in [n]$ and each prefix $\mathbf{x}_1^{i-1} \in \mathcal{X}^{i-1}$ we have for $\hat{\delta} > 0$ and $\delta' \stackrel{def}{=} \frac{2\delta}{\hat{\delta}} + \frac{2\delta}{1-e^{-\epsilon}}$:*

$$\Pr\left[(\mathcal{A}, \mathbf{X}_1^{i-1}) \in \mathcal{E}_i(\hat{\delta})|\mathbf{x}_1^{i-1}\right] \geq 1 - \delta'.$$

We now define the set of outcome/dataset prefix pairs for which the quantities Z_i are not large:

$$\mathcal{F}_i = \left\{(a, \mathbf{x}_1^i) : |Z_i(a, \mathbf{x}_1^i)| \leq 6\epsilon\right\}. \tag{4}$$

Using another technical lemma from [27] (which we state in the full version), we prove:

Claim III.5. *Given $(a, \mathbf{x}_1^{i-1}) \in \mathcal{E}_i(\hat{\delta})$ and $\delta'' \stackrel{def}{=} \frac{2\hat{\delta}}{1-e^{-3\epsilon}}$ we have:*

$$\Pr\left[(\mathcal{A}, \mathbf{X}_1^i) \in \mathcal{F}_i|a, \mathbf{x}_1^{i-1}\right] \geq 1 - \delta''.$$

We now define the "good" tuples of outcomes and databases as

$$\mathcal{G}_i(\hat{\delta}) = \left\{(a, \mathbf{x}_1^i) : (a, \mathbf{x}_1^{i-1}) \in \mathcal{E}_i(\hat{\delta}) \quad \& \quad (a, \mathbf{x}_1^i) \in \mathcal{F}_i\right\}, \tag{5}$$

$$\mathcal{G}_{\leq i}(\hat{\delta}) = \left\{(a, \mathbf{x}_1^i) : (a, x_1) \in \mathcal{G}_1(\hat{\delta}), \cdots, (a, \mathbf{x}_1^i) \in \mathcal{G}_i(\hat{\delta})\right\} \tag{6}$$

Claim III.6. *If $\mathcal{A}$ is (ϵ, δ)-differentially private and $\mathbf{X} \sim \mathcal{P}^n$, then*

$$\Pr\left[(\mathcal{A}, \mathbf{X}_1^i) \in \mathcal{G}_i(\hat{\delta})\right] \geq 1 - \delta' - \delta''$$

for δ' and δ'' given in Claim III.4 and Claim III.5, respectively.

Having shown a high probability bound on the terms Z_i, our next step is to bound their expectation so that we can continue towards our goal of applying Azuma's inequality. Note a complicating factor – throughout the argument, we need to condition on the event $(\mathcal{A}, \mathbf{X}_1^i) \in \mathcal{F}_i$ to ensure that Z_i has bounded expectation.

We will use the following shorthand notation for conditional expectation:

$$\begin{aligned}&\mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_1^i)|a, \mathbf{x}_1^{i-1}, \mathcal{F}_i\right] \\ &\quad \stackrel{def}{=} \mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_1^i)|\mathcal{A} = a, \mathbf{X}_1^{i-1} = \mathbf{x}_1^{i-1}, (\mathcal{A}, \mathbf{X}_1^i) \in \mathcal{F}_i\right],\end{aligned}$$

with similar notation for sets $\mathcal{G}_i(\hat{\delta}), \mathcal{G}_{\leq i}(\hat{\delta})$.

Lemma III.7. *Let $\mathcal{A}$ be (ϵ, δ)-differentially private and $\mathbf{X} \sim \mathcal{P}^n$. Given $(a, \mathbf{x}_1^{i-1}) \in \mathcal{E}_i(\hat{\delta})$, for all $\epsilon \in (0, 1/2]$ and $\hat{\delta} \in (0, \epsilon/15]$,*

$$\mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_1^i)|a, \mathbf{x}_1^{i-1}, \mathcal{F}_i\right] = O(\epsilon^2 + \hat{\delta}).$$

Finally, we need to apply Azuma's inequality to a set of variables that are bounded with probability 1, not just with high probability. Towards this end, we define variables T_i that will match Z_i for "good events", and will be zero otherwise—and hence, are always bounded:

$$T_i(a, \mathbf{x}_1^i) = \begin{cases} Z_i(a, \mathbf{x}_1^i) & \text{if } (a, \mathbf{x}_1^i) \in \mathcal{G}_{\leq i}(\hat{\delta}) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The next lemma verifies that the variables T_i indeed satisfy the requirements of Azuma's inequality:

Lemma III.8. *Let $\mathcal{A}$ be (ϵ, δ)-differentially private and $\mathbf{X} \sim \mathcal{P}^n$. The variables T_i defined in (7) are bounded by 6ϵ with probability 1, and for any $(a, \mathbf{x}_1^{i-1}) \in \mathcal{Y} \times \mathcal{X}^{i-1}$ and $\hat{\delta} \in [0, \epsilon/15]$,*

$$\mathbb{E}\left[T_i(\mathcal{A}, \mathbf{X}_1^i)|a, \mathbf{x}_1^{i-1}\right] = O(\epsilon^2 + \hat{\delta}/\epsilon), \tag{8}$$

where the bound does not depend on n or i.

We can then apply Azuma's inequality to the sum of $T_i(a, \mathbf{x}_1^i)$, where each term will match $Z_i(a, \mathbf{x}_1^i)$ for most $(a, \mathbf{x}_1^i)$ coming from $(\mathcal{A}(\mathbf{X}), \mathbf{X}_1^i)$ for each $i \in [n]$. Note that, from Lemma III.2, we know that a bound on $\sum_{i=1}^n Z_i(a, \mathbf{x}_1^i)$ with high probability will give us a bound on approximate max-information. See the full version for a formal analysis.

IV. A Counterexample to Nontrivial Composition and a Lower Bound for Non-Product Distributions

It is known that algorithms with bounded description length have bounded approximate max-information [6]. In section III, we showed that (ϵ, δ)-differentially private algorithms have bounded approximate max-information when the dataset is drawn from a product distribution. In this section, we show that although approximate max-information composes adaptively [6], one cannot always run a bounded description length algorithm, followed by a differentially private algorithm, and expect the resulting composition to

have strong generalization guarantees. In particular, this implies that (ϵ, δ)-differentially private algorithms cannot have any nontrivial bounded max-information guarantee over non-product distributions.

Specifically, we give an example of a pair of algorithms $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{A}$ has output description length $o(n)$ for inputs of length n, and $\mathcal{B}$ is (ϵ, δ)-differentially private, but the adaptive composition of $\mathcal{A}$ followed by $\mathcal{B}$ can be used to exactly reconstruct the input database with high probability. In particular, it is easy to overfit to the input $\mathbf{X}$ given $\mathcal{B}(\mathbf{X}; \mathcal{A}(\mathbf{X}))$, and hence, no nontrivial generalization guarantees are possible. Note that this does not contradict our results on the max-information of differentially private algorithms for *product distributions*: even if the database used as input to $\mathcal{A}$ is drawn from a product distribution, the distribution on the database is no longer a product distribution *once conditioned on the output of* $\mathcal{A}$. The distribution of $\mathcal{B}$'s input violates the hypothesis that is used to prove a bound on the max-information of $\mathcal{B}$.

Theorem IV.1. *Let $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{\mathcal{X}^n \cup \{\perp\}\}$. Let $\mathbf{X}$ be a uniformly distributed random variable over $\mathcal{X}^n$. For $n > 64e$, for every $\epsilon \in \left(0, \frac{1}{2}\right], \delta \in \left(0, \frac{1}{4}\right]$, there exists an integer $r > 0$ and randomized algorithms $\mathcal{A} : \mathcal{X}^n \to \{0,1\}^r$, and $\mathcal{B} : \mathcal{X}^n \times \{0,1\}^r \to \mathcal{Y}$, such that:*

1) *$r = O\left(\frac{\log(1/\delta)\log n}{\epsilon}\right)$ and $I_\infty^\beta(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq r + \log(\frac{1}{\beta})$ for all $\beta > 0$;*
2) *for every $\mathbf{a} \in \{0,1\}^r$, $\mathcal{B}(\mathbf{X}, \mathbf{a})$ is (ϵ, δ)-differentially private and $I_\infty^\beta(\mathbf{X}; \mathcal{B}(\mathbf{X}, \mathbf{a})) \leq 1$ for all $\beta \geq 2\delta$;*
3) *for every $\mathbf{x} \in \mathcal{X}^n$, with probability at least $1 - \delta$, we have that $\mathcal{B}(\mathbf{x}; \mathcal{A}(\mathbf{x})) = \mathbf{x}$. In particular, $I_\infty^\beta(\mathbf{X}, \mathcal{B}(\mathbf{X}; \mathcal{A}(\mathbf{X}))) \geq n - 1$ for all $0 < \beta \leq \frac{1}{2} - \delta$.*

De [21] showed that the *mutual information* of (ϵ, δ)-differentially private protocols can be large: if $\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right) = O(n)$, then there exists an (ϵ, δ)-differentially private algorithm $\mathcal{B}$ and a distribution $\mathcal{S}$ such that for $\mathbf{X} \sim \mathcal{S}$, $I(\mathbf{X}; \mathcal{B}(\mathbf{X})) = \Omega(n)$, where I denotes mutual information. De's construction also has large approximate max-information.

By the composition theorem for approximate max-information (given in the full version), our construction implies a similar bound:

Corollary IV.2. *There exists an (ϵ, δ)-differentially private mechanism $\mathcal{C} : \mathcal{X}^n \to \mathcal{Y}$ such that $I_\infty^{\beta_2}(\mathcal{C}, n) \geq n - 1 - r - \log(1/\beta_1)$ for all $\beta_1 \in (0, 1/2 - \delta)$ and $\beta_2 \in (0, 1/2 - \delta - \beta_1)$, where $r = O\left(\frac{\log(1/\delta)\log(n)}{\epsilon}\right)$.*

We adapt ideas from De's construction in order to prove Theorem IV.1. In De's construction, the input is not drawn from a product distribution—instead, the support of the input distribution is an error-correcting code, meaning that all points in the support are far from each other in Hamming distance. For such a distribution, De showed that adding the level of noise required for differential privacy does not add enough distortion to prevent decoding of the dataset.

Our construction adapts De's idea. Given as input a *uniformly random* dataset $\mathbf{x}$, we show a mechanism $\mathcal{A}$ which outputs a short description of a code that contains $\mathbf{x}$. Because this description is short, $\mathcal{A}$ has small max-information. The mechanism $\mathcal{B}$ is then parameterized by this short description of a code. Given the description of a code and the dataset $\mathbf{x}$, $\mathcal{B}$ approximates (privately) the distance from $\mathbf{x}$ to the nearest *codeword*, and outputs that codeword when the distance is small. When $\mathcal{B}$ is composed with $\mathcal{A}$, we show that it outputs the dataset $\mathbf{x}$ with high probability.

We define the mechanisms $\mathcal{A}$ and $\mathcal{B}$ from the theorem statement in Algorithm 1 and Algorithm 2, respectively.

Brief description of $\mathcal{A}$: For any input $\mathbf{x} \in \mathcal{X}^n$, mechanism $\mathcal{A}$ returns a vector $\mathbf{a}_\mathbf{x} \in \{0,1\}^r$ such that $\mathbf{x} \in C_{\mathbf{a}_\mathbf{x}}$, where $C_{\mathbf{a}_\mathbf{x}} = \{\mathbf{c} \in \mathcal{X}^n : H\mathbf{c} = \mathbf{a}_\mathbf{x}\}$ is an affine code with minimum distance t. We give further details in the full version of the paper.

Input: $\mathbf{x} \in \{0,1\}^n$
Output: $\mathbf{a}_\mathbf{x} \in \{0,1\}^r$
Return $H\mathbf{x}$ (multiplication in $\mathbb{F}_2$).

Algorithm 1: $\mathcal{A}$

Brief description of $\mathcal{B}_{\epsilon,\delta}$: For any input $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{a} \in \{0,1\}^r$, mechanism $\mathcal{B}_{\epsilon,\delta}$ first computes $d_\mathbf{x}$, which is the distance of $\mathbf{x}$ from $f(\mathbf{x})$, i.e., the nearest codeword to $\mathbf{x}$ in code $C_\mathbf{a}$. Next, it sets $\hat{d}_\mathbf{x}$ to be $d_\mathbf{x}$ perturbed with Laplace noise $L \sim \text{Lap}(1/\epsilon)$. It returns $f(\mathbf{x})$ if $\hat{d}_\mathbf{x}$ is below a threshold $w \stackrel{def}{=} \left(\frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right)$, and $\perp$ otherwise.

Input: $\mathbf{x} \in \{0,1\}^n$ (private) and $\mathbf{a} \in \{0,1\}^r$(public)
Output: $\mathbf{b} \in \mathcal{Y}$
Compute the distance of $\mathbf{x}$ to the nearest codeword in code $C_\mathbf{a}$. Let $d_\mathbf{x} = \min_{\mathbf{c} \in C_\mathbf{a}}(dist_{Hamm}(\mathbf{x}, \mathbf{c}))$ and $f(\mathbf{x}) = \arg\min_{\mathbf{c} \in C_\mathbf{a}}(dist_{Hamm}(\mathbf{x}, \mathbf{c}))$ (breaking ties arbitrarily).
Let $\hat{d}_\mathbf{x} = d_\mathbf{x} + L$, where $L \sim \text{Lap}(1/\epsilon)$, and Lap($c$) denotes a random variable having Laplace(0,c) distribution.
if $\hat{d}_\mathbf{x} < \left(\frac{t-1}{4} - \frac{\log(1/\delta)}{\epsilon}\right)$ **then**
  Return $f(\mathbf{x})$.
else
  Return $\perp$.
end

Algorithm 2: $\mathcal{B}_{\epsilon,\delta}$

We prove Theorem IV.1 in the full version.

ACKNOWLEGMENTS

R.R. acknowledges support in part by a grant from the Sloan foundation, and NSF grant CNS-1253345. A.R. acknowledges support in part by a grant from the Sloan foundation, a Google Faculty Research Award, and NSF grants CNS-1513694 and CNS-1253345. O.T. and A.S. acknowledge support in part by a grant from the Sloan foundation, a Google Faculty Research Award, and NSF grant IIS-1447700. We thank Salil Vadhan for pointing out that our max-information bound can be used to bound mutual information, thus improving on a result in [22].

REFERENCES

[1] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao, "Valid post-selection inference," *The Annals of Statistics*, vol. 41, no. 2, pp. 802–837, 2013.

[2] W. Fithian, D. Sun, and J. Taylor, "Optimal inference after model selection," *arXiv preprint arXiv:1410.2597*, 2014.

[3] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor, "Exact post-selection inference, with application to the lasso," *arXiv preprint arXiv:1311.6238*, 2013.

[4] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, "Preserving statistical validity in adaptive data analysis," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, ser. STOC '15. New York, NY, USA: ACM, 2015, pp. 117–126. [Online]. Available: http://doi.acm.org/10.1145/2746539.2746580

[5] R. Bassily, K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of the 48th Annual ACM on Symposium on Theory of Computing, STOC*, 2016.

[6] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2341–2349. [Online]. Available: http://papers.nips.cc/paper/5993-generalization-in-adaptive-data-analysis-and-holdout-reuse.pdf

[7] A. Gelman and E. Loken, "The statistical crisis in science," *American Scientist*, vol. 102, no. 6, p. 460, 2014.

[8] R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: context, process, and purpose," *The American Statistician*, vol. 0, no. ja, pp. 00–00, 2016. [Online]. Available: http://dx.doi.org/10.1080/00031305.2016.1154108

[9] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, Oct. 2011. [Online]. Available: http://pss.sagepub.com/lookup/doi/10.1177/0956797611417632

[10] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2016.

[11] Y. Wang, J. Lei, and S. E. Fienberg, "A minimax theory for adaptive data analysis," *CoRR*, vol. abs/1602.04287, 2016. [Online]. Available: http://arxiv.org/abs/1602.04287

[12] A. Johnson and V. Shmatikov, "Privacy-preserving data exploration in genome-wide association studies," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 1079–1087.

[13] C. Uhler, A. Slavkovic, and S. E. Fienberg, "Privacy-preserving data sharing for genome-wide association studies," *Journal of Privacy and Confidentiality*, vol. 5, no. 1, 2013.

[14] F. Yu, S. E. Fienberg, A. B. Slavkovic, and C. Uhler, "Scalable privacy-preserving data sharing methodology for genome-wide association studies," *Journal of Biomedical Informatics*, vol. 50, pp. 133–141, 2014.

[15] V. Karwa and A. Slavković, "Inference using noisy degrees: Differentially private beta-model and synthetic graphs," *The Annals of Statistics*, vol. 44, no. 1, pp. 87–112, 2016.

[16] C. Dwork, W. Su, and L. Zhang, "Private false discovery rate control," *arXiv preprint arXiv:1511.03803*, 2015.

[17] O. Sheffet, "Differentially private least squares: Estimation, confidence and rejecting the null hypothesis," *arXiv preprint arXiv:1507.02482*, 2015.

[18] Y. Wang, J. Lee, and D. Kifer, "Differentially private hypothesis testing, revisited," *arXiv preprint arXiv:1511.03376*, 2015.

[19] M. Gaboardi, H. Lim, R. Rogers, and S. Vadhan, "Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing," *arXiv preprint arXiv:1602.03090*, 2016.

[20] C. Dwork, G. N. Rothblum, and S. P. Vadhan, "Boosting and differential privacy," in *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, 2010, pp. 51–60. [Online]. Available: http://dx.doi.org/10.1109/FOCS.2010.12

[21] A. De, "Lower bounds in differential privacy," in *Proceedings of the 9th International Conference on Theory of Cryptography*, ser. TCC'12, 2012, pp. 321–338.

[22] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. P. Vadhan, "The limits of two-party differential privacy," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 18, p. 106, 2011. [Online]. Available: http://eccc.hpi-web.de/report/2011/106

[23] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *In Proceedings of the 3rd Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[24] R. Cummings, K. Ligett, K. Nissim, A. Roth, and Z. S. Wu, "Adaptive learning with robust generalization guarantees," *arXiv preprint arXiv:1602.07726*, 2016.

[25] M. Hardt and J. Ullman, "Preventing false discovery in interactive data analysis is hard," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 454–463.

[26] T. Steinke and J. Ullman, "Interactive fingerprinting codes and the hardness of preventing false discovery," in *Proceedings of The 28th Conference on Learning Theory*, 2015, pp. 1588–1628.

[27] S. Kasiviswanathan and A. Smith, "On the 'Semantics' of Differential Privacy: A Bayesian Formulation," *Journal of Privacy and Confidentiality*, vol. Vol. 6: Iss. 1, Article 1, 2014, available at http://repository.cmu.edu/jpc/vol6/iss1/1. The theorem numbers and exact statements refer to the Arxiv version (v3).

[28] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, pp. 486–503.