Private False Discovery Rate Control

Cynthia Dwork* Weijie Su† Li Zhang‡

April 2, 2015

Abstract

Hypothesis testing refers to the use of a statistical procedure to assess the validity of a putative scientific discovery, for example, that a given treatment for a condition is more effective than a placebo. The difficulty in hypothesis testing is to distinguish random effects due to sampling from true nulls. The problem is exacerbated when the number of hypotheses is large, and the focus of a large subfield in statistics is the relaxation of the goal of preventing all false discoveries to instead controlling the *rate* of false discovery.

We provide the first differentially private algorithms for controlling the false discovery rate (FDR) in multiple hypothesis testing, with essentially no loss in power. Our general approach is to adapt a well-known variant of the famous Benjamini-Hochberg "BHq" method, making each step differentially private. This destroys the classical proof of FDR control. To prove FDR control of our method we

1. Develop a new proof of the original (non-private) BHq algorithm and its robust variants – a proof requiring fewer independence assumptions than previously shown in the vast literature on this topic – and
2. Relate the FDR control properties of the differentially private version to the control properties of the non-private version.

We also present a low-distortion "one-shot" differentially private primitive for "top k" problems, *e.g.*, "Which are the k most popular hobbies?" (which we apply to: "Which hypotheses have the k most significant p-values?"), and use it to get a faster privacy-preserving instantiation of our general approach at little cost in accuracy. The proof of privacy for the one-shot top k algorithm introduces a new technique of independent interest.

*Microsoft Research. dwork@microsoft.com

†Department of Statistics, Stanford University. wjsu@stanford.edu

‡Google Inc. liqzhang@google.com

1 Introduction

Hypothesis testing is the use of a statistical procedure to assess the validity of a given hypothesis, for example, that a given treatment for a condition is more effective than a placebo. The traditional approach to hypothesis testing is to (1) formulate a pair of *null* (the treatment and the placebo are equally effective) and *alternative* (the treatment is more effective than the placebo) hypotheses; (2) devise a test statistic for the null hypothesis, (3) collect data; (4) compute the *p-value* (the probability of observing an effect as or more extreme than the observed effect, were the null hypothesis true; smaller p-values are "better evidence" for rejecting the null); and (5) compare the p-value to a standard threshold α to determine whether to *accept* the null hypothesis (conclude there is no interesting effect) or to *reject* the null hypothesis in favor of the alternative hypothesis.

In the *multiple hypothesis testing* problem, also known as the *multiple comparisons* problem, p-values are computed for multiple hypotheses, leading to the problem of false discovery: since the p-values are typically defined to be uniform in $(0, 1)$ for true nulls, by definition we expect an α fraction of the true nulls to have p-values bounded above by α.

Multiple hypothesis testing is an enormous problem in practice; for example, in a single genome-wide association study a million SNPs[1] may be tested for an association with a given condition. Accordingly, there is a vast literature on the problem of controlling the *false discovery rate* (FDR), which, roughly speaking, is the expected fraction of erroneously rejected hypotheses among all the rejected hypotheses, where the expectation is over the choice of the data and any randomness in the algorithm (see below for a formal definition).

The seminal work of Benjamini and Hochberg [2] and their beautiful "BHq" procedure (Algorithm 1 below) is our starting point. The procedure takes as input a parameter q and, assuming certain independence conditions together with a mild condition on the distribution of p-values for true nulls, "controls", or bounds, the false discovery rate by q. An extensive literature explores the control capabilities and power of this procedure and its many variants[2].

Differential privacy [8] is a definition of privacy tailored to statistical data analysis. The goal in a differentially private algorithm is to hide the presence or absence of any individual or small group of individuals, the intuition being that an adversary unable to tell whether or not a given individual is even a member of the dataset surely cannot glean information specific to this individual. To this end, a pair of databases x, y are said to be *adjacent* (or *neighbors*) if they differ in the data of just one individual. Then a differentially private algorithm (Definition 2.4) ensures that the behavior on adjacent databases is statistically close, so it is difficult to infer if any particular individual is in the database from the output of the algorithm and any other side information.

Contributions. We provide the first differentially private algorithms for controlling the FDR in multiple hypothesis testing. The problem is difficult because the data of a single individual can affect the p-values of all hypotheses simultaneously. Our general approach is to adapt a well-known variant (Algorithm 2 below) of the BHq procedure, making each step differentially private. This destroys the classical proof of FDR control. To prove FDR control of our method we

1. Develop a new proof of the original (non-private) BHq algorithm and its robust variants – a proof requiring fewer independence assumptions than previously shown in the vast literature on this topic – and

2. Relate the FDR control properties of the differentially private version to the control properties of the non-private version. Power is also argued in relation to the non-private version.

[1] A *single nucleotide polymorphism*, or SNP, is a location in the DNA in which there is variation among individuals.

[2] The power of a procedure is its ability to recognize "false nulls" (what a lay person may describe as "true positives").

Central to BHq and its variants is the procedure for reporting the experiments with k most significant p-values, or known as the top-k problem, *e.g.*, "Which are the k most popular hobbies?" To our knowledge, the most accurate approximately private top-k algorithm is a "peeling" procedure, in which one runs a differentially private maximum procedure k times, tuning the "inaccuracy" to roughly $\sqrt{k}/\varepsilon$. Here we present a low-distortion "one-shot" differentially private primitive for "top k" with the same dependence on k, and use it to get a faster privacy-preserving instantiation of our general approach[3]. The proof of privacy for our one-shot top k algorithm introduces a new technique of independent interest.

1.1 Description of Our Approach

The BHq procedure. Denote by R the number of rejections made by any procedure and V ($\leq R$) the number of true null hypotheses that are falsely rejected (false discoveries). The FDR is defined as FDR $\doteq \mathbb{E}(\frac{V}{R}; V \geq 1)$, where the semi-colon notation amounts to always interpreting V/R as zero if $V = 0$ (note that $R = 0$ implies $V = 0$).

Algorithm 1 presents the original Benjamini-Hochberg "BHq" procedure for controlling false discovery rate at q. The thresholds $\alpha_j = jq/m$ for $1 \leq j \leq m$, are known as the *BHq cutoffs*. The following intuition may demystify the BHq algorithm. In the case that all m null hypotheses are true and their p-values are i. i. d. uniform on $(0, 1)$, then we expect, approximately, a iq/m fraction of the p-values to lie in the interval $[0, jq/m]$. If instead there are j many p-values in this interval (this is precisely what the condition $p_{(j)} \leq iq/m$ says), then there are "too many" p-values in $[0, p_{(j)}]$ for all of them to correspond to null hypotheses: We have j p-values, and should attribute no more than jq of these to the true nulls; so if we reject all these j hypotheses we would expect that at most a q fraction correspond to true nulls.

Algorithm 1 The Original BHq procedure

Input: $0 < q < 1$, and m p-values $p_1, \cdots, p_m$
Output: a set of rejected hypotheses

```
sort the p-values in increasing order, obtaining p_(1) ≤ p_(2) ≤ ··· ≤ p_(m);
for j = m to 1 do
    if p_(j) ≤ α_j then
        reject all remaining hypotheses (include itself) and halt
    else
        remove (j) from the set of hypotheses under consideration
    end if
end for
```

Making BHq Private: There is an extensive literature on, and burgeoning practice of, differential privacy. For now, we need only two facts: (1) Differential privacy is closed under composition, permitting us to bound the cumulative privacy loss over multiple differentially private computations. This permits us to build complex differentially private algorithms from simple differentially private primitives; and (2) we will make use of the well known Report Noisy Max (respectively, Report Noisy Min) primitive, in which appropriately distributed fresh random noise is added to the result of each of m computations, and the index of the computation yielding the maximum (respectively, minimum) noisy value is returned. By returning only one index the procedure allows us to pay an

[3]For pure privacy, previous work e.g. [4] showed a way to avoid peeling with $O(k)$ Laplacian noise. But it is much more challenging to get to the dependence on k down to$\sqrt{k}$ for approximate privacy, which is what we show in this paper.

accuracy price for a single computation, rather than for all m.[4]

A natural approach to making the BHq procedure differentially private is by repeated use of the Report Noisy Max primitive: Starting with $i = m$ and decreasing: use Report Noisy Max to find the hypothesis H_i with the (approximately) largest p-value; estimate that p-value and, if the estimate is above (an appropriately more conservative cutoff) $\alpha'_i < \alpha_i$, accept H_i as null, remove H_i from consideration, and repeat. Once an H_i is found whose p-value is below the threshold, reject all the remaining hypotheses. The principal difficulty with this approach is that every iteration of the algorithm incurs a privacy loss which can only be mitigated by increasing the magnitude of the noise used by Report Noisy Max. Since each iteration corresponds to the acceptance of a null hypothesis, the procedure is paying in privacy/accuracy precisely for the null hypotheses accepted, which are by definition not the "interesting" ones.

If, instead of starting with the largest p-value and considering the values in decreasing order, we were to start with the smallest p-value and consider the values in increasing order, rejecting hypotheses one by one until we find a $j \in [m]$ such that $p_{(j)} > \alpha'_j$, the "demystification" intuition still applies, This widely-studied variation is called the Benjamini-Hochberg *step-down* procedure (see Algorithm 2 in Section 2.1), although we find the nomenclature counterintuitive and so call it BH_{small} to remind the reader that we start with the smallest p-value[5].

If we make the natural modifications to BH_{small} (using Report Noisy Min now, instead of Report Noisy Max), then we pay a privacy cost only for nulls rejected in favor of the thecorresponding alternative hypotheses, which by definition are the "interesting" ones. Since the driving application of BHq is to select promising directions for future investigation that have a decent chance of panning out, we can view its outcome as advice for allocating resources. Thus, a procedure that finds a relatively small number k of high-quality hypotheses, still enjoying false discovery rate control, may be as useful as a procedure that finds a much larger set.

Proving FDR Control. The BHq and BH_{small} algorithms have the property that the rejected hypotheses are contiguous in sorted order; for example, it is not possible that the hypotheses with the smallest and third smallest p-values are rejected while the hypothesis with the second smallest p-value is accepted. Because of noise introduced for protecting privacy, this may not be the case in our differentially private algorithms. Our investigation of the FDR control of these more relaxed algorithms shows that the key condition is as follows: if R hypotheses are rejected, the maximum p-value of any rejected hypothesis is bounded by Rq/m. Note that this perfectly matches our "demystification" intuition. To this end, we introduce the notion of *adaptive* procedures.

Definition 1.1. A multiple testing procedure is said to be *adaptive to cutoffs* $\{\alpha_j\}_{j=1}^m$, if it either rejects none or the rejected p-values are upper-bounded by α_R, where the number of rejections R can be arbitrary.

Clearly, both BHq and BH_{small} procedures are adaptive to BH cutoffs. But an adaptive procedure can be more broad as it does not require to maintain the exact order of the p-values and rejects some consecutive minimum p-values. Yet, we are still able to show an adaptive procedure controls the FDR. In addition, we are interested in studying k-FDR [12, 13] of such procedures, where FDR_k is defined as $\text{FDR}_k \triangleq \mathbb{E}(V/R; V \geq k)$.

[4]The variance of the noise distribution depends on the maximum amount that any single datum can (additively) change the outcome of a computation and the inverse of the privacy price one is willing to pay.

[5]Theoretical results show that the power of BH_{small} is similar to that of BHq:[1] shows that the optimal thresholding for Gaussian sequence denoising is between the step-down and step-up BHq. The difference between these two BHq's is asymptotically negligible. Also, by [11] we can choose less stringent cutoffs, which often give even more discoveries than the step-up BHq.

Theorem 1.2. *Assume that the test statistics corresponding to the true null hypotheses are jointly independent. Then any algorithm adaptive to the BHq cutoffs $\alpha_j = jq/m$ ensures* $\mathrm{FDR} \le q\log(1/q) + Cq$, $\mathrm{FDR}_2 \le Cq$, *and* $\mathrm{FDR}_k \le (1 + 2/\sqrt{qk})q$, *where $C < 3$ is a universal constant.*

One novelty of Theorem 1.2 lies in the absence of any assumptions about the relationship between the true null test statistics and false null test statistics. In the literature, independence, or some (very stringent) sort of positive dependence [3] between these two groups of test statistics, is necessary in order to have provable FDR control. In a different line, Theorem 1.3 of [3] controls FDR without any assumptions by using the (very stringent) cutoffs $\tilde{\alpha}_j = jq/(m\sum_{i=1}^{m}\frac{1}{i}) \approx jq/(m\log m)$, effectively paying a factor of $\log m$, whereas we pay only a constant factor. This simple independence assumption within the true nulls shall also capture more real life scenarios. As we will see, the additive $q\log(1/q)$ term for the original FDR, *i.e.*, FDR_1, is unavoidable given so few assumptions. Surprisingly, FDR_2 no longer has this dependency, and FDR_k even approaches q as k grows.

Theorem 1.2 is the key to proving FDR control even if the procedure is given "noisy" versions of the p-values, as happens in our differentially private algorithms. To determine how to add noise to ensure privacy, we study the *sensitivity* of a p-value, a measure of how much the p-value can change between adjacent datasets (Section 4.1)[6]. For standard statistical tests this change is best measured multiplicatively, rather than additively, as is typically studied in differential privacy. Exploiting multiplicative sensitivity is helpful when the values involved are very small. Since we are interested in the regime where m, the number of hypotheses, is much larger than the number of discoveries, the p-values we are interested in are quite small: the kth BH cutoff is only kq/m. We remark that adapting algorithms such as Report Noisy Min and the one-shot top k to incorporate multiplicative sensitivity can be achieved by working with the logarithms of the values involved.

Informally, we say a p-value is η multiplicatively sensitive if for any two neighboring databases D and D', the p-value computed on them are within multiplicative factor of e^{η} of each other, unless they are very small. (See Definition 4.1) The justification of the multiplicative sensitivity is provided by the definition of some standard p-values under independent bounded statistics. Indeed as we show in Section 4.1, such p-values are $\widetilde{O}(1/\sqrt{n})$ multiplicatively sensitive[7]. With these results, we can show that our algorithm controls FDR under the bound given in Theorem 1.2 while having the comparable power to the $\mathrm{BH}_{\mathrm{small}}$ (q).

Theorem 1.3 (Private FDR (informal))**.** *For η-multiplicatively sensitive p-values, the FDR control properties of our differentially private algorithms are no worse than what is proven for the non-private case if we replace q with $\mathrm{e}^{\widetilde{O}(\eta\sqrt{k})}q + o(1)$.*

Applying to the standard p-values for independent bounded statistics, we have that

Theorem 1.4 (Power of Private FDR on Independent Bounded Statistics (informal))**.** *For independent bounded statistics, if $k \ll n/\log^{O(1)} m$, then with probability $1 - o(1)$, our differentially private algorithms reject at least the same number of hypotheses as BH_{small} (q) while maintaining FDR control with the bound given in Theorem 1.2 provided q is replaced by $(1 + o(1))q$.*

The One-Shot Top k Mechanism. Consider a database of n binary strings $d_1, d_2, \ldots, d_n$. Each d_i corresponds to an individual user and has length m. Let $x_j = \sum_i d_{ij}$ be the total sum of the the j-th column. In the reporting top-k problem[8], we wish to report, privately, some locations $j_1, \ldots, j_k$ such that x_{j_ℓ} is close to the ℓ-th smallest element as possible. The peeling mechanism reports and removes the minimum noisy count and repeats on the remaining elements, each time

[6] Recall that adjacent datasets differ in the data of just one person.

[7] We use $\widetilde{O}$ to hide polynomial factors in $\log(m/\delta)/\varepsilon$.

[8] In our paper, it is more convenient to consider the minimum-k (or bottom-k) elements. But we still call it the top-k problem following the convention.

adding fresh noise. Such a mechanism is (ε, δ)-differentially private if we add $\mathrm{Lap}(\sqrt{k \log(1/\delta)}/\varepsilon)$ noise each time.

In contrast, in the one-shot top k mechanism, we add $\widetilde{O}(\sqrt{k})$ noise to each value and then report the k locations with the minimum noisy values.[9] Compared to the peeling mechanism, the one-shot mechanism is appealingly simple and much more efficient. But it is surprisingly challenging to prove its privacy. Here we give some intuition.

When there are large gaps between x_i's, the change of one individual can only change the value of each x_i by at most 1, so result of the one-shot algorithm is stable. Hence the privacy is easily guaranteed. In the more difficult case, when there are many similar values (think of the case when all the values are equal), the true top k set can be quite sensitive to the change of input values. But in the one-shot algorithm we add independent symmetric noise, centered at zero, to these values. Speaking intuitively, this yields an (almost) equal chance that on two adjacent input values a noisy value will "go up" or "go down," leading to cancellation of certain first order terms in the (logarithms of) the probabilities of events and hence a tight control between their ratio.

To capture this intuition, we consider the "bad" events, which have large probability bias between two neighboring inputs. Those bad events can be shown to happen when the sum of some dependent random variables deviates from its mean. The dependencies among the random variables prevents us from applying a concentration bound directly. To deal with this difficulty, we first partition the event spaces to remove the dependence. Then we apply a coupling technique to pair up the partitions for the two neighboring inputs. For each pair we apply a concentration bound to bound the probability of bad events. The technique appears to be quite general and might be useful in other settings.

2 Preliminaries

2.1 Hypothesis testing

We revisit some basic concepts in multiple hypothesis testing. For two random variables X and Y, X is said to be stochastically larger than or equal to Y, if $\mathbb{P}(X > a) \geq \mathbb{P}(Y > a)$ for all a.

Assumption 2.1. *A p-value is a random variable distributed in* $[0, 1]$*, computed from a test statistic, such that under the null hypothesis it is stochastically larger than or equal to uniform on* $(0, 1)$*.*

In testing hypotheses $H_1, \ldots, H_m$, recall that we write R for the number of rejected hypotheses, among which V are erroneously rejected.

Definition 2.2. Define the false discovery rate $\mathrm{FDR} = \mathbb{E}[V/(R \vee 1)]$, where $R \vee 1 = \max\{R, 1\}$. Following [12, 13], define $\mathrm{FDR}_k = \mathbb{E}[V/R; V \geq k]$, recalling that the semi-colon notation says to interpret V/R as zero if $V < k$.

2.2 Step-down BHq

For any sequence $p_1, \ldots, p_m$, let $p_{(1)}, \ldots, p_{(m)}$ be increasingly sorted, and (j) be the subscript of the jth smallest value. Set $\alpha_j = jq/m$, the BHq cutoffs. The step-down BHq procedure [11], which is denoted by $\mathrm{BH}_{\mathrm{small}}$, is summarized in Algorithm 2[10].

Under the assumptions that the true null test statistics are jointly independent, and are further independent from those of false nulls. Then [11] shows that Algorithm 2 controls the FDR at level

[9] For technical reasons, if we want the values of the computations in addition to their indices we only know how to prove privacy if we add fresh random noise before releasing these values.

[10] In [11], the BHq cutoffs are replaced with $\tilde{\alpha}_j = jq/(m + 1 - j + jq)$ with improved power. We will use the original BHq cutoffs for the ease of presentation.

Algorithm 2 $\mathrm{BH}_{\mathrm{small}}$: Adaptive step-down FDR procedure

Input: $0 < q < 1$, and m p-values $p_1, \cdots, p_m$

Output: a set of rejected hypotheses

```
sort the p-values in increasing order, obtaining p_(1) ≤ p_(2) ≤ ... ≤ p_(m);
for j = 1 to m do
    if p_(j) ≤ α_j then
        reject (j);
    else
        return
    end if
end for
```

q. Translated to p-value domain, this condition is equivalent to that the true null p-values are independent and stochastically larger than or equal to $U(0,1)$, and the p-values corresponding to the false nulls can have arbitrary distribution, as long as they are independent from the true null p-values.

2.3 Differential privacy

Definition 2.3. Data sets D, D' are said to be *neighbors*, or *adjacent*, if one is obtained by removing or adding a single data item.

Differential privacy, now sometimes called *pure differential privacy*, was defined and first constructed in [9]. The relaxation defined next is sometimes referred to as *approximate differential privacy*.

Definition 2.4 (Differential privacy [9, 7])**.** A randomized mechanism M is (ε, δ)-differentially private if for all adjacent x, y, and for any event S: $\mathbb{P}_D[S] \leq \mathrm{e}^{\varepsilon}\mathbb{P}_{D'}[S] + \delta$. Pure differential privacy is the special case of approximate differential privacy in which $\delta = 0$.

Denote by $\mathrm{Lap}(\lambda)$ the (symmetric) Laplacian distribution with standard deviation λ, whose cumulative distribution function (CDF) reads $G_\lambda(z) = \frac{1}{2}\mathrm{e}^{z/\lambda}$ for $z < 0$ and $G_\lambda(z) = 1 - \frac{1}{2}\mathrm{e}^{-z\lambda}$ for $z \geq 0$. Let $G(z)$ denote $G_1(z)$ so $G_\lambda(z) = G(z/\lambda)$.

Definition 2.5 (Sensitivity of a function)**.** Let f be a function mapping databases to $\mathbb{R}^k$. The *sensitivity of* f, denoted Δf, is the maximum over all pairs D, D' of adjacent datasets of $\|f(D) - f(D')\|$.

We will make heavy use of the following two lemmas on differential privacy.

Lemma 2.6 (Laplace Mechanism [9])**.** *Let f be a function mapping databases to $\mathbb{R}^k$. The mechanism that, on database D, adds independent random draws from $\mathrm{Lap}((\Delta f)/\varepsilon)$ to each of the k components of $f(D)$, is $(\varepsilon, 0)$-differentially private.*

Lemma 2.7 (Advanced Composition [10])**.** *For all $\varepsilon, \delta, \delta' \geq 0$, the class of (ε, δ')-differentially private mechanisms satisfies $(\sqrt{2k\ln(1/\delta)}\varepsilon + k\varepsilon(e^{\varepsilon} - 1)/2, k\delta' + \delta)$-differential privacy under k-fold adaptive composition.*

3 Robust $\mathrm{BH}_{\mathrm{small}}$

We now study the FDR control of the *adaptive* procedures defined in the introduction and then prove Theorem 1.2. The class of adaptive procedures includes BHq and $\mathrm{BH}_{\mathrm{small}}$. This serves as a first step in proving FDR control properties of our differentially private algorithm.

We consider the bounds on FDR, FDR_2 and on FDR_k separately. Here we outline the proof for FDR and FDR_2 and leave FDR_k case in the supplemented full version.

3.1 Controlling FDR and FDR_2

We will prove the result by constructing the most "adversarial" set of p-values. Imagine the following game for a powerful adversary A who are informed of all the m_1 false null hypotheses and can even set p-values for them. The remaining p-values, which are all from the true nulls, are then drawn from i. i. d. $U(0,1)$. Then A can pick out a subset S of p-values with the only requirement that those p-values are upper bounded by $\alpha_{|S|}$ for the cutoffs $\alpha_j = jq/m$. A's payoff is then the ratio of the true nulls (i.e. FDR) in S. The expected payoff of A would be the upper bound on FDR for any adaptive procedure. If we require A only receive payoff when he includes at least k true nulls in S, then the corresponding payoff is an upper bound of FDR_k.

First how should A set the p-values of alternatives? 0! because this way A can include any number of alternatives in S to push up the size of S so raise up the cutoffs but without wasting any "space" for including more true nulls. With this we can reduce bounding FDR to bounding the expected value of a random variable.

We now present the rigorous argument below. Denote by $\mathscr{N}_0$, with cardinality m_0, the set of true null hypotheses, and $\mathscr{N}_1$, with cardinality m_1, the set of false nulls. Define $\mathrm{FDP}_k = V/(R \vee 1)$ for $V \geq k$ and 0 otherwise. Given a realization of $p_1, \ldots, p_m$, we would like to obtain a tight upper bound on $V/(R \vee 1)$ with the constraint that the maximum of the rejected p-values is no larger than α_R. With this in mind, call $(p_{i_1}, p_{i_2}, \ldots, p_{i_R})$ the rejected p-values, among which V of them are from the m_0 many true null p-values. Hence, denoting by $p^0_1, \ldots, p^0_{m_0}$ the true null p-values, we see $p^0_{(V)} \leq \max_{1 \leq j \leq R} p_{i_j} \leq \alpha_R$. Taking the BHq cutoff $\alpha_R = qR/m$ and rearranging this inequality yield $R \geq \lceil mp^0_{(V)}/q \rceil$, which also makes use of the additional information that R is an integer. As a consequence, we get $V/(R \vee 1) \leq V/\lceil mp^0_{(V)}/q \rceil$. Hence, it follows that

$$\mathrm{FDP}_k \leq \max_{k \leq j \leq m_0} \frac{j}{\lceil mp^0_{(j)}/q \rceil}\,. \tag{3.1}$$

Recognizing that the true null p-values are independent and stochastically no smaller than uniform on $(0,1)$, we may assume the true null p-values are i. i. d. uniform on $(0,1)$ in taking expectations of both sides of (3.1). In addition, it is easy to see that increasing m_0 to m only makes this inequality more likely to hold. Hence, we have

$$\mathrm{FDR}_k \leq \mathbb{E}\left(\max_{k \leq j \leq m} \frac{j}{\lceil mU_{(j)}/q \rceil} \right), \tag{3.2}$$

where, as earlier, $U_{(1)} \leq U_{(2)} \cdots \leq U_{(m)}$ are the order statistics of i. i. d. uniform random variables $U_1, \ldots, U_m$ on $(0,1)$.

Taking $k = 2$ in (3.2), we proceed to obtain the bound on FDR_2 in the following lemma. We note that because of some technical issues, the following analysis applies to the case of $k = 2$. For general k, we use a different technique to bound FDR_k. See the full version.

Lemma 3.1. *There exists an absolute constant C such that*

$$\mathbb{E}\left(\max_{2 \leq j \leq m} \frac{qj}{mU_{(j)}} \right) \leq Cq\,.$$

To bound the above expectation, we use a well-known representation for the uniform order statistics (e.g. see [5]): $(U_{(1)}, \ldots, U_{(m)}) \overset{d}{=} (T_1/T_{m+1}, \ldots, T_m/T_{m+1})$, where $T_j = \xi_1 + \cdots + \xi_j$, and

$\xi_1, \ldots, \xi_{m+1}$ are i. i. d. exponential random variables. Denote by $W_j = jT_{m+1}/T_j$. Then Lemma 3.1 is equivalent to bounding $\mathbb{E}\left(\max_{2\le j\le m} W_j/m\right) \le C$.

Intuitively the maximum is more likely to be realized by smaller j's as increasing j would increase the concentration of $T_j/j = \sum_{i=1}^{j} \xi_i/j$. Then above expectation can be bounded by considering a few terms of W_j for small j's. Indeed this intuition can be made rigorous by observing that $W_1, \ldots, W_{m+1}$ is a *backward submartingale*, and hence we can use the well known technique to bound $\mathbb{E}\max_{2\le j\le m} W_j$ by some statistics of W_2, which we can then estimate.

Lemma 3.2. *With T_j, W_j defined as above, we have that $W_1, \ldots, W_{m+1}$ is a backward submartingale with respect to the filtration (or conditional on the "history") $\mathcal{F}_j = \sigma(T_j, T_{j+1}, \ldots, T_{m+1})$ for $j = 1, \ldots, m+1$, i.e. $\mathbb{E}(W_j|\mathcal{F}_{j+1}) \ge W_{j+1}$, for $j = 1, \ldots, m$.*

Now we apply Lemma 3.2 to prove Lemma 3.1.

Proof of Lemma 3.1. Given that W_j/m is a backward submartingale by Lemma 3.2, we can apply Theorem 5.4.4 in [6], which concludes

$$\begin{aligned}\mathbb{E}\Big(\max_{2\le j\le m} W_j/m\Big) &\le (1-\mathrm{e}^{-1})^{-1}\Big[1+\mathbb{E}\Big(\frac{W_2}{m}\log\frac{W_2}{m}; \frac{W_2}{m}\ge 1\Big)\Big] \\ &= (1-\mathrm{e}^{-1})^{-1}\Big[1+\mathbb{E}\Big(\frac{2}{mU_{(2)}}\log\frac{2}{mU_{(2)}}; \frac{2}{mU_{(2)}}\ge 1\Big)\Big].\end{aligned}$$

A bit of analysis reveals that the R.H.S. of the last display is bounded. □

Provided the bound on FDR_2, we can easily obtain the bound on FDR. Taking $k = 1$ in (3.2), we get

$$\mathrm{FDR} \le \mathbb{E}\Big(\max_{2\le j\le m} \frac{j}{\lceil mU_{(j)}/q\rceil}\Big) + \mathbb{E}\Big(\min\Big\{\frac{1}{\lceil mU_{(1)}/q\rceil}, 1\Big\}\Big),$$

It is easy to show the second term is bounded by $q\log(1/q)$. This proves the FDR bound in Theorem 1.2. We note that if we do not put any lower bound on the number of discoveries, we suffer an extra additive term of $q\log(1/q)$. This can be shown to be necessary. See the full version for the example.

4 Private FDR algorithm

One alternative interpretation of Theorem 1.2 is that $\mathrm{BH}_{\mathrm{small}}$ is robust with respect to small perturbation of p-values. That is if we add small enough noise to the p-values and then apply $\mathrm{BH}_{\mathrm{small}}$ procedure, the FDR can still be controlled within the bound given in the theorem. It turns out the additive sensitivity of p-values might be large, but the relative change is much smaller. This motivates us to consider multiplicative sensitivity. Hence we will add multiplicative noise (or additive noise to the logarithm of p-values). Together with the private top-k algorithm, this gives us the private FDR algorithm.

For the ease of presentation, we assume that we are given an upper bound k of the number of rejections. The parameter k should be comparable to the number of true rejections and small compared to m. It is easy to adapt our algorithm to the case when k is not given, for example, by the standard doubling trick. This only incurs an extra logarithmic factor in our bound.

4.1 Sensitivity of p-values

Assume the input to our private FDR algorithm is an m-tuple of p-values $p = (p_1, \ldots, p_m)$ obtained by running m statistical tests on a dataset D. Motivated by the normal approximation (or related χ^2 approximation) frequently used in computing p-values, the change in a p-value caused by the addition or deletion of the data of one individual is best measured multiplicatively; however, when the p-value is very small the relative change can be very large. This gives rise to the following definition of (truncated) multiplicative neighborhood.

Definition 4.1 ((η, ν)-neighbors)**.** Tuples $p = (p_1, \ldots, p_m), p' = (p'_1, \ldots, p'_m)$ are (η, ν)-*neighbors* if, for all $1 \le i \le m$, either $p_i, p'_i < \nu$, or $\mathrm{e}^{-\eta} p_i \le p'_i \le \mathrm{e}^{\eta} p_i$.

The privacy of our FDR algorithm will be defined with respect to such neighborhood. Next we will explain that some standard p-values computed on neighboring databases are indeed (η, ν)-neighbors for small η and ν. We will give an intuitive explanation using Gaussian approximation but omit the proof details. Consider a database consisting of the records of n people and a hypothesis H to be tested, where each person contributes a statistic $t_i, i = 1, \ldots, n$ with $|t_i| \le B$ (for example, t_i is the number of minor alleles for a given SNP.) In many interesting cases, the sufficient statistic for testing the hypothesis is $T = t_1 + \ldots + t_n$, and under the null hypothesis H, each t_i has mean μ and variance σ^2. Then $(T - n\mu)/(\sqrt{n}\sigma)$ is asymptotically distributed as standard normal variable. Assuming T tends to be larger under the alternative hypothesis, we can approximately compute p-value $p(T) = \Phi(-(T - n\mu)/(\sqrt{n}\sigma))$, where Φ is the CDF of standard Gaussian. Consider a neighboring database where person i is replaced, so $T' - T = t'_i - t_i$. Writing $p' = p(T')$ and invoking that $\Phi(-x) \approx \frac{1}{x}\phi(x) = \frac{1}{x\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$ for large $x > 0$, we can show that $p(T)$ is (η, ν) multiplicative sensitive for $\eta \approx B\sqrt{2\log(1/\nu)/n}/\sigma$. More generally, we have

Lemma 4.2. *For independent bounded statistics of n people, the p-value is (η, ν) multiplicatively sensitive with respect to any individual change for $\eta = O(\sqrt{\log(1/\nu)/n})$.*

4.2 The private FDR algorithm

To achieve privacy with respect o (η, μ)-neighborhood, we apply $\mathcal{M}^{\mathrm{Peel}}$ with properly scaled noise to the logarithm of the input p-values (see Step 1 of Algorithm 3). Define $\Delta_k = O(\sqrt{k\log(1/\delta)}\log m/\varepsilon)$ to be the accuracy bound of $\mathcal{M}^{\mathrm{Peel}}$. We run $\mathrm{BH}_{\mathrm{small}}$ with cutoff values set as $\alpha'_j = \log(\alpha_j + \nu) + \eta\Delta_k$. The details are described in Algorithm 3.

Algorithm 3 Private FDR

Input: m values $p_1, \cdots, p_m$ and $k \ge 1$ and $\varepsilon, \delta, \eta, \nu$

Output: a set of up to k rejected hypotheses

1: for each $1 \le i \le m$, set $x_i = \log(\max(p_i, \nu))$;
2: apply $\mathcal{M}^{\mathrm{Peel}}$ to $x_1, \ldots, x_m$ with noise $\mathrm{Lap}(\eta\sqrt{k\log(1/\delta)}/\varepsilon)$ to obtain $(i_1, y_1), \ldots, (i_k, y_k)$;
3: apply $\mathrm{BH}_{\mathrm{small}}$ to $y_1, \ldots, y_k$ with cut-off values α'_j for $1 \le j \le k$ and reject the corresponding hypotheses.

Theorem 4.3. *Algorithm 3 satisfies the following*

1. *it is (ε, δ) private with respect to (η, ν)-neighborhood;*
2. *if BH_{small} rejects $k' \le k$ hypotheses, Algorithm 3 rejects at least k' hypotheses with probability $1 - o(1)$;*

3. *suppose $\nu = o(1/m)$, then the FDR of Algorithm 3 satisfies the bounds in Theorem 1.2 with q replaced by $\mathrm{e}^{\eta\Delta_k}(1+o(1))q$.*

The proof follows from Theorems 1.2 and the bound on the peeling mechanism. See the supplemented full version for details.

By Lemma 4.2, for the binary database with independent statistics, $\eta = O(\sqrt{\log m/n})$ and $\nu = 1/m^2$, so we have

Corollary 4.4. *Algorithm 3 is (ε,δ)-private. In addition, if $k \ll n/\log^{O(1)}(m)$, with probability $1-o(1)$, it rejects at least the same amount of hypotheses as in Algorithm 2 and controls FDR under $q\log(1/q) + C(1+o(1))q$, FDR_2 under $C(1+o(1))q$, and FDR_k under $(1+o(1))q$.*

One drawback of the above algorithm is that it needs to run the peeling algorithm which takes $\widetilde{O}(km)$ time. This can be expensive for large k. In the next section, we show the privacy of one-shot algorithm $\mathcal{M}^{\mathrm{os}}$. By replacing $\mathcal{M}^{\mathrm{Peel}}$ with $\mathcal{M}^{\mathrm{os}}$ in Algorithm 3, we can obtain the *Fast Private FDR algorithm* which has essentially the same quality bound but with running time $\widetilde{O}(k+m)$.

5 One-shot mechanism for reporting top-k

In the one-shot mechanism $\mathcal{M}^{\mathrm{os}}$, we add noise $\mathrm{Lap}(\lambda)$ once to each value and report the indices of the minimum-k noisy values.

Using a coupling argument, it is easy to show by setting $\lambda = O(k/\varepsilon)$, the one-shot mechanism is $(\varepsilon,0)$-private. Our goal is to reduce the dependence on k to $\sqrt{k}$ for approximate privacy. Here we will only outline the main claims and leave the details to the supplemented full version.

Theorem 5.1. *Assume $\varepsilon \le \log(m/\delta)$. There exists universal constant $C>0$ such that if we set $\lambda = C\sqrt{k\log(m/\delta)}/\varepsilon$, then $\mathcal{M}^{\mathrm{os}}$ is (ε,δ)-private. In addition, with probability $1-o(1)$, for every $1 \le j \le k$, $x_{i_j} - x_{(j)} = O(\sqrt{k\log(m/\delta)}\log m/\varepsilon)$.*

Unlike in the peeling approach, in which an ordered set is returned, in $\mathcal{M}^{\mathrm{os}}$, only a subset of k elements, but not their ordering, is returned. The privacy proof crucially depends on this. If we would also like to report the values, we can report the noisy values by adding random noise *freshly* sampled from $\mathrm{Lap}(O(\sqrt{k\log(1/\delta)}/\varepsilon))$ to each value of those k elements.

The quality bound of Theorem 5.1 is immediate from the properties of exponential random variables. We will need to prove the privacy property.

We divide the event space by fixing the k-th smallest noisy element, say j, together with the noise value, say g_j. For each partition, whether an element $i \neq j$ is selected by $\mathcal{M}^{\mathrm{os}}$ only depends on whether $x_i + g_i \le x_j + g_j$, which happens with probability $q_i = G((x_j+g_j-x_i)/\lambda)$. Here G denotes the CDF of standard Laplacian distribution. Hence, we consider the following mechanism $\mathcal{M}$ instead: given $(q_1,\dots,q_m)$ where $0<q_i<1$, output a subset of indices where each i is included with probability q_i. It turns out for two adjacent databases, their respective probability vectors q and q' are multiplicatively close, defined as follows. As a reminder, the notation q is slightly abused to represent a vector, rather than the nominal level in multiple testing.

Definition 5.2 (c-closeness for vectors)**.** For two vectors $q=(q_1,\dots,q_m)$ and $q'=(q'_1,\dots,q'_m)$, we say q,q' are *c-close* if for each $1\le i\le m$, $|q_i - q'_i| \le cq_i(1-q_i)$.

The following is the crucial lemma whose proof requires much technical work.

Lemma 5.3. *Assume $\varepsilon \le \log(1/\delta)$ and $k \ge \log(1/\delta)$. There exists constant $C_1>0$ such that if q and q' are c-close with $c \le \frac{\varepsilon}{C_1\sqrt{k\log(1/\delta)}}$, then for any set S of k-subsets of $\{0,1\}^m$, $\mathbb{P}(\mathcal{M}(q)\in S) \le \mathrm{e}^{\varepsilon}\mathbb{P}(\mathcal{M}(q')\in S)+\delta$.*

Theorem 5.1 then follows from the above partitioning and Lemma 5.3. The details can be found in the full version.

References

[1] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, pages 584–653, 2006.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistics Society: Series B (Statistical Methodology)*, 57:289–300, 1995.

[3] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[4] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. Discovering frequent patterns in sensitive data. In *KDD*, 2010.

[5] H. David and H. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[6] R. Durrett. *Probability: theory and examples*. Cambridge University Press, 2010.

[7] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of EUROCRYPT*, pages 486–503, 2006.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer, 2006.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer, 2006.

[10] C. Dwork, G. Rothblum, and S. Vadhan. Boosting and differential privacy. In *Proceedings of FOCS*, 2010.

[11] Y. Gavrilov, Y. Benjamini, and S. K. Sarkar. An adaptive step-down procedure with proven FDR control under independence. *The Annals of Statistics*, pages 619–629, 2009.

[12] S. K. Sarkar. Stepup procedures controlling generalized FWER and generalized FDR. *The Annals of Statistics*, pages 2405–2420, 2007.

[13] S. K. Sarkar and W. Guo. On a generalized false discovery rate. *The Annals of Statistics*, pages 1545–1565, 2009.