A Differentially Private Wilcoxon Signed-Rank Test

Abstract

Hypothesis tests are a crucial statistical tool for data mining and are the workhorse of scientific research in many fields. Here we present a differentially private analogue of the classic Wilcoxon signed-rank hypothesis test, which is used when comparing sets of paired (e.g., before-and-after) data values. We present not only a private estimate of the test statistic, but a method to accurately compute a p-value and assess statistical significance. We evaluate our test on both simulated and real data. Compared to the only existing private test for this situation, that of Task and Clifton, we find that our test requires less than half as much data to achieve the same statistical power.

1 Introduction

Consider a scenario where a medical researcher measures a patient's blood pressure before and after a particular intervention. Once the data is obtained, the researcher wants to determine whether the intervention had any effect, or whether the difference between the two samples was just random variation. The data generated in this study is an example of *paired-sample* data, where there exists a natural link between the two measurements of the same individual. This is a common form of data in scientific research, and one of the traditional methods to check for the signficance of an effect is the Wilcoxon signed-rank test [30]. This test does not require any additional assumptions about the data (e.g., that it was normally distributed), and so it can be used on social network data and other data for which it may not be appropriate to assume normality.

Unfortunately, in many of the use cases for the Wilcoxon test, the data in question is private and sensitive (e.g., medical records, financial records or social network data). Ethical and legal concerns prevent the easy sharing of this data. Frequently, research cannot be done because the owners of existing data are not willing to accept the risk inherent in entrusting this data to a third party. Even if the researcher is entrusted with access to the data, the researcher themselves must now decide what they can publish. The release of even just summary statistics has led to privacy violations [11], but without the release of some results the research is pointless.

Differential privacy provides a criterion that is sufficient to provably guarantee that the results of a particular query on a database do not violate individual privacy. Differential privacy generally requires the query be randomized, with the output usually aiming to approximate the result of a non-private query as accurately as possible. Years of research have resulted in a variety of effective algorithms for differentially private queries, allowing the approximation of everything from histograms [8] to regression model coefficients [17, 31, 32] to classifiers from machine learning algorithms [4, 21].

In this paper we provide an algorithm that carries out a private analogue of the Wilcoxon signed-rank test. A private version of the test will always have lower statistical power than the classical version, meaning that more data is needed before a given effect can be distinguished from random noise. The first differentially private analogue of the Wilcoxon test was given by Task and Clifton in 2016 [23]. Here we strive to improve on that result, introducing a new test with higher power.

In a hypothesis test, a test statistic is computed and compared to a reference distribution to see if it is abnormally high. The difficulty of developing a private test comes not just from the need to privately approximate a test statistic, but also from the need to rigorously define "abnormally high." It is not sufficient to treat the approximate test statistic as equivalent to its non-private counterpart.

The key contributions of this work are as follows:

A new private test statistic We estimate not the standard Wilcoxon test statistic but a variant introduced by Pratt in 1959 [14]. We argue that this variant is more amenable to being privatized without losing its utility. We prove a bound on its sensitivity and provide a private algorithm for approximating this statistic.

Accurate p-value calculation We show how to calculate an accurate p-value that takes into account the noise added for privacy. We do this through simulation, numerically approximating (to a high degree of accuracy) the relevant reference distribution. This is a more exact method than the upper bounds given in prior work.

Experimental implementation We implement our algorithm and measure its power. We also implement the test of Task and Clifton and along the way correct a significant error in their work. We then compare the two tests and find that ours can detect the same effect with only 40% of the data needed previously, closing the majority of the gap between the private and public versions of the test. The source code for our implementation is publicly available.[1]

2 Background

We begin by outlining some background on hypothesis testing, followed by some specifics on the Wilcoxon signed-rank test. We then discuss differential privacy and previous applications of differentially private hypothesis testing.

2.1 Hypothesis Testing

Hypothesis testing is a very common form of data analysis, particularly in applied scientific research. The goal of a hypothesis test is to measure whether a particular data set is consistent with a given theory describing how the data is generated. This theory is called the *null hypothesis* (H_0). For example, given blood pressure data before and after some treatment, researchers might want to determine if the change they see could plausibly be consistent with a null hypothesis that the treatment has no effect and the observed effect is due only to sampling variability.

To make this determination, a *test statistic*, t, is computed. Given a function f for computing a test statistic, one can determine analytically, through simulation, or via large sample approximation, the distribution that the statistic $T = f(\mathbf{X})$ would have when the database $\mathbf{X}$ is drawn randomly according to H_0.[2] One can then compute a *p-value*, defined as the probability that a test statistic as or more extreme than that observed would occur under H_0.

Definition 2.1. For an observed database $\mathbf{x}$, the test statistic $t = f(\mathbf{x})$, and a null hypothesis H_0, the *p-value* is defined as

$$\Pr[T \geq t \mid T = f(\mathbf{X}) \text{ and } \mathbf{X} \leftarrow H_0].$$

Low p-values indicate that the test statistic is more extreme than is likely by random chance and may cause a researcher to reject H_0 as a plausible explanation of the data. Normally a threshold value α is chosen and H_0 is rejected if $p < \alpha$. The chosen value for α determines the *type I error rate*, the probability of incorrectly rejecting if the null hypothesis is true. The value of t at which $p = \alpha$, which we will denote t^*, is called the *critical value*.

Hypothesis tests are judged by their *statistical power*. This is a measure of how likely the test is to detect an effect and reject the null hypothesis when it is indeed false. Statistical power is a function of the amount of data and the particular alternate distribution (i.e., the effect size).

Definition 2.2. For a given alternate data distribution H_A, the *statistical power* of a hypothesis test is

$$\Pr[T \geq t^* \mid T = f(\mathbf{X}) \text{ and } \mathbf{X} \leftarrow H_A].$$

A primary goal of hypothesis test design is to find tests with high power. In classical statistics there are many hypothesis tests that have been proven to be optimal for a large set of use cases.

2.2 Wilcoxon Signed-Rank Test

Consider again the example where researchers want to see if a treatment caused a change in blood pressure. A standard hypothesis test for this situation is the Wilcoxon signed-rank test, proposed in 1945 by Frank Wilcoxon [30]. This test evaluates the difference between values obtained from two paired samples (such as subjects before and after treatment, husbands and wives, etc.) to see if they plausibly come from the same distribution. The test assumes that all pairs are independent, random draws from a distribution with an ordinal scale. In the public setting, when the data is known to be normally distributed, the t-test consistently outperforms the Wilcoxon. However, the Wilcoxon test does not assume normality, which is beneficial when the underlying distribution of the sample data is not known.

The function calculating the Wilcoxon test statistic is formalized in Algorithm $\mathcal{W}$. Given a database $\mathbf{x}$

[1] Our code is available at: URL withheld for anonymity.

[2] We use the convention of denoting random variables with a capital roman letter and a particular value taken by that random variable with a lowercase roman letter.

containing sets of pairs (u_i, v_i), the test computes the difference d_i of each pair, drops any with $d_i = 0$, and then ranks them by magnitude. (If magnitudes are equal for several differences, all are given a rank equal to the average rank for that set.) If $s_i = \pm 1$ is the sign of d_i and r_i is its rank, then $w = \sum_i s_i r_i$.

```
Algorithm $\mathcal{W}$ : Wilcoxon Test Statistic
  Calculation
Input: $\mathbf{x}$
begin
    for row $i$ of $\mathbf{x}$ do
        $d_i \longleftarrow |v_i - u_i|$
        $s_i \longleftarrow$ Sign$(v_i - u_i)$
    Order the terms from lowest to highest $d_i$
    Drop any $d_i = 0$
    for row $i$ of $\mathbf{x}$ do
        $r_i \longleftarrow$ rank of row $i$
    $w \longleftarrow \sum_i s_i r_i$
Output: $w$
```

Example 2.3. Table 1 presents an example database containing five data elements. The values u_i and v_i are inputs in the database and d_i and s_i are the results of the initial stage of computation.

Table 1: Initial Table

i	u_i	v_i	d_i	s_i
1	9	18	9	1
2	2	11	9	1
3	3	3	0	-
4	8	10	2	1
5	9	8	1	-1

In table 2 we sort the elements in increasing order by d_i. We drop $i = 3$ because $d_3 = 0$. We then compute ranks. Because there is a tie for ranks 3 and 4, we set the rank of these values to 3.5.

We then compute the test statistic

$$w = \sum_i s_i r_i = 3.5 \cdot 1 + 3.5 \cdot 1 + 2 \cdot 1 + 1 \cdot (-1) = 8.$$

Under the null hypothesis that u_i and v_i are drawn from the same distribution, the distribution of the test statistic W can be calculated exactly using combinatorial techniques. This becomes computationally infeasible for large databases, but an approximation exists in the form of the normal distribution with mean 0 and variance $\frac{n_r(n_r+1)(2n_r+1)}{6}$, where n_r is the number of rows that were not dropped. Knowing this, one can calculate the p-value for any particular value of w.

Table 2: Final Table

i	d	s	r
5	1	-1	1
4	2	1	2
1	9	1	3.5
2	9	1	3.5

2.3 Differential Privacy

When working with data that contains sensitive information about individuals, the need to publicly share the result of the hypothesis test must be balanced with the need to protect privacy. Unfortunately, traditional methods of handling sensitive data provide poor privacy protections. Databases privatized with ad hoc anonymization techniques (replacing names with unique numeric identifiers, for example) have frequently been attacked, resulting in substantial privacy violations [1, 22, 29]. Even summary statistics can allow malicious actors to violate the privacy of people in the database [11].

Differential privacy [8] is a definition that, when satisfied, provably guarantees that the output of a query cannot be used to infer anything about the individuals who contributed to the database. Informally, it requires that for any individual, the query output would be roughly the same for any value of that individual's data. To do this, the query mechanism must be randomized. More formally:

Definition 2.4 (Differential Privacy)**.** A randomized algorithm $\tilde{f}$ on databases is ϵ-differentially private if for all $\mathcal{S} \subseteq$ Range$(\tilde{f})$ and for databases $\mathbf{x}, \mathbf{x}'$ that only differ in one row

$$\Pr[\tilde{f}(\mathbf{x}) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\tilde{f}(\mathbf{x}') \in \mathcal{S}].$$

We say that $\mathbf{x}$ and $\mathbf{x}'$ that differ only in one row are *neighboring databases* and we call ϵ the *privacy parameter*. The user selects ϵ, which determines the strength of the privacy.

One useful property of differential privacy is resistance to post-processing, which ensures that no function of a differentially private algorithm can cause a privacy loss [8].

Theorem 2.5 (Post Processing)**.** *Let $\tilde{f}$ be an ϵ-differentially private randomized algorithm. Let g be an arbitrary randomized algorithm. Then $g \circ \tilde{f}$ is ϵ-differentially private.*

Resistance to post-processing is an important theorem for two reasons. First, any good privacy definition must have this property (or something very similar). If it didn't, then the privacy of the output is only superficial and some sort of analysis can extract the private information. Secondly, this shows that post-processing can be used to design private algorithms. If one step of a computation is differentially private, then as long as the rest of the computation does not directly access the database, the result of the full computation will be private as well. We use this technique later.

One standard technique for creating differentially private algorithms is the *Laplace mechanism*, introduced by Dwork et al. in 2006 [8]. Given an arbitrary (non-private) function f, the Laplace mechanism provides one way to create a private function $\tilde{f}$ that approximates f. To do this, one must first compute the sensitivity of f, meaning the maximum effect a change in a single row can have on the output.

Definition 2.6 (Sensitivity)**.** The sensitivity of a function f is

$$\Delta f = \max_{\mathbf{x},\mathbf{x}'} |f(\mathbf{x}) - f(\mathbf{x}')|,$$

where $\mathbf{x}$ and $\mathbf{x}'$ are neighbouring databases.

The Laplace mechanism adds noise to f, where the noise is sampled from a Laplace distribution, a double-sided exponential distribution.

Definition 2.7 (Laplace Distribution)**.** The Laplace distribution (centered at 0) with scale b is the distribution with probability density function:

$$\text{Lap}(x|b) = \frac{1}{2b}\text{exp}\Big(-\frac{|x|}{b}\Big).$$

We usually write Lap(b) to denote the Laplace distribution with scale b.

The Laplace mechanism simply adds noise produced from the Laplace distribution to the output of a function. The magnitude of the noise is determined by the choice of ϵ and the sensitivity of the function.

Definition 2.8 (Laplace Mechanism)**.** Given any function f, the Laplace mechanism is defined as:

$$\tilde{f}(\mathbf{x}) = f(\mathbf{x}) + Y,$$

where Y is drawn from Lap$(\Delta f/\epsilon)$, and Δf is the sensitivity of f.

The Laplace mechanism is one of the most commonly used methods to achieve differential privacy. Dwork et al. [8] proved the following:

Theorem 2.9. *(Laplace Mechanism) The Laplace mechanism preserves ϵ-differential privacy.*

2.4 Differentially Private Hypothesis Tests

To carry out a differentially private hypothesis test, one must first find a differentially private function $\tilde{f}$ that computes a useful test statistic. This could be either a differentially private estimate of a standard test statistic, or a completely new test statistic.[3]

In a testing framework, a test statistic is *not* a meaningful output on its own; it is an intermediate calculation on the way to a p-value. One cannot simply use the estimated statistic in place of the true one and carry out a p-value computation as in the non-private setting, as resultant p-values can be wildly

[3] We note here that our test statistics are now randomized, rather than deterministic functions of the database. This actually causes very few mathematical issues, but it does mean that the output of a hypothesis test on fixed data will not be the same on each run.

innacurate and artificially inflate the type I error rate [3, 26]. Instead, one must compute a new reference distribution modeling the distribution of $\tilde{f}(\mathbf{X})$ when $\mathbf{X}$ is generated under H_0.

A fully developed differentially private hypothesis test includes not just a method for computing the test statistic, but also for computing the associated p-value. The goal is to develop such a hypothesis test and to achieve power as close as possible to what can be achieved in the classic non-private setting. Just as each type of input data requires a separate hypothesis test in the classical case, each type of input data requires its own differentially private test.

2.5 Related Work

There is a significant body of work on differentially private hypothesis testing, broadly construed, but much of it takes a different approach than this work. For example, some results (e.g., [18, 28, 19]) look at how quickly various private approximations of test statistics converge to their limiting distributions. These are important theoretical results, but they do not bring us to the point of practical, implementable tests. There is often no discussion of a reasonable reference distribution when the noise is not yet negligible, and the results are often purely asymptotic, hiding potentially problematic constants.

The hypothesis test for which private variants have been the most well-studied is the chi-squared test, which tests whether two categorical variables could plausibly be independent of each other.[4] Vu and Slavković [25] give differentially private versions of a single proportion test and a chi-squared test with clinical trial data in mind. They give accurate p-value calculations that adjust for the added Laplace noise. Several studies of private chi-squared tests have been specifically intended for use with genome-wide association study (GWAS) data [9, 24, 12]. These papers make asymptotic arguments for the reliability of their p-value calculations. Other work has used Monte Carlo simulations to achieve higher precision [10, 27]. Rogers and Kifer [15] instead proposes a new statistic with an asymptotic distribution more similar to its non-private analogue. While all of these papers discuss private test statistics and many (but not all) discuss the distribution of that statistic under the null hypothesis, few carefully demonstrate the statistical power of the resulting test. Rogers and Kifer [15] and Gaboardi et al. [10] are notable exceptions, giving power curves for several different approaches.

Methods for numerical (rather than categorical) data are less well-developed. Some work has addressed testing the value of a mean or the difference of means [20, 7, 6]. While there is a large body of work on differentially private linear regression that aims to output an accurate best-fit line, only recently has work been done on using coefficients as test statistics [16, 2]. (Such inference is near-ubiquitous in many academic fields.) Campbell et al. [3] give a private analogue of a one-way analysis of variance (ANOVA) test, and Nguyên and Hui propose a test for surival analysis data [13].

3 Our Algorithm

In this section, we will define our algorithm for the differentially private Wilcoxon test and prove that it is ϵ-differentially private. At a high level, our algorithm is quite straightforward; we compute a test statistic as one might in the public case and add Laplacian noise to make it private. However, there are several important innovations that greatly increase the power of our test.

Our first innovation is to use a different variant of the Wilcoxon test statistic. While the version introduced in Section 2.2 is the one most commonly used, other versions have long existed in the statistics literature. In particular, we look at a variant introduced by Pratt in 1959 [14]. In this variant, rather than dropping rows with $d_i = 0$, those rows are included. When $d_i = 0$ we set $s_i = \text{Sign}(d_i) = 0$, so those rows contribute nothing to the resultant statistic, but they do push up the rank of other, non-tied rows. We define the algorithm for this calculation, which we denote $\mathcal{WP}$, below.

[4] There are variants of this test that can be used in slightly different situations, like seeing of a certain set of categorical data is consistent with a specified distribution.

```
Algorithm $\mathcal{WP}$: Wilcoxon Statistic - Pratt Variant
Input: $\mathbf{x}$
begin
    for row $i$ of $\mathbf{x}$ do
        $d_i \longleftarrow |v_i - u_i|$
        $s_i \longleftarrow \text{Sign}(v_i - u_i)$
    Order the terms from lowest to highest $d_i$
    for row $i$ of $\mathbf{x}$ do
        $r_i \longleftarrow$ rank of row $i$
    $w \longleftarrow \sum_i s_i r_i$
Output: $w$
```

This procedure is almost the same as that for the calculation of the standard Wilcoxon test statistic, but the small difference is crucial. In the sample database given in 1, the Pratt variant produces the values in Table 3.

Table 3: Final Table (Pratt)

i	d	s	r
3	0	0	1
5	1	-1	2
4	2	1	3
1	9	1	4.5
2	9	1	4.5

The resulting statistic is

$$w = \sum_i s_i r_i = 4.5 \cdot 1 + 4.5 \cdot 1 + 3 \cdot 1 + 2 \cdot (-1) = 10.$$

In the public setting, the Pratt variant is not very different from the standard Wilcoxon, being slightly more or less powerful depending on the exact effect one is trying to detect [5]. In the private setting, however, the difference is substantial.

The benefit to the Pratt variant comes from how the test statistics are interpreted. In the standard Wilcoxon, it is known that the test statistic follows an approximately normal distribution, but the variance of that distribution is a function of n_r, the number of non-zero d_i values. In the private setting, this number is not known, and this has caused substantial difficulty in prior work. (See Section 5 for more discussion.) On the other hand, the Pratt variant produces a test statistic that is always compared to the same normal distribution, which depends only on n.

3.1 The Private Test Statistic

Recall that $\mathbf{x}$ is a database and n is the size of the database. The $\mathcal{WP}$ algorithm (above) produces the public version of the Pratt-variant Wilcoxon statistic. The algorithm $\widetilde{\mathcal{WP}}$ that outputs the differentially private analog is shown below.

```
Algorithm $\widetilde{\mathcal{WP}}$: Private Wilcoxon Statistic
Input: $\mathbf{x}$, $\epsilon$
begin
    $n \longleftarrow$ the number of paired samples in $\mathbf{x}$
    $w \longleftarrow \mathcal{WP}(\mathbf{x})$
    $\widetilde{w} \longleftarrow w + \text{Lap}\left(\frac{2n}{\epsilon}\right)$
Output: $\widetilde{w}$
```

3.2 Proof of Differential Privacy

Theorem 3.1. *Algorithm $\widetilde{\mathcal{WP}}$ is ϵ-differentially private.*

Proof. It is sufficient to show that the sensitivity of $\mathcal{WP}$ is bounded by $2n$. The privacy of the algorithm then follows directly from Theorem 2.9.

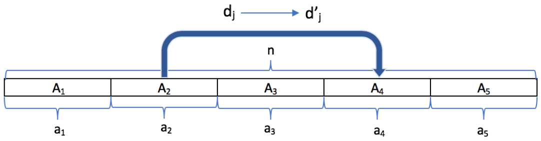

Figure 1: $\mathbf{x}$ and $\mathbf{x}'$ differ only in that d_j is "moved" from region A_2 to region A_4.

Let $\mathbf{x}$ and $\mathbf{x}'$ be neighbouring databases. Let row j be the row that differs between $\mathbf{x}$ and $\mathbf{x}'$. Let d_j and d'_j indicate the difference between the data values in row j in $\mathbf{x}$ and $\mathbf{x}'$, respectively. Similarly, let r_j and r'_j indicate the rank of row j in $\mathbf{x}$ and $\mathbf{x}'$. We assume without loss of generality that $d'_j > d_j$.

We divide the rows of $\mathbf{x}$ (or $\mathbf{x}'$) into five regions, A_1 through A_5. A_1 is the set of rows i with $d_i < d_j$. A_2 is the set with $d_i = d_j$. A_3 has d_i values between d_j and d'_j. A_4 has d_i values equal to d'_j. A_5 has values greater than d'_j (though we don't need to use A_5 in the proof). For convenience, we use a_i to represent the size of the set A_i in $\mathbf{x}$.

As shown in Figure 1, $\mathbf{x}$ and $\mathbf{x}'$ differ only in that d_j is "moved" from region A_2 to region A_4. Let w and w' be test statistics produced from databases $\mathbf{x}$ and $\mathbf{x}'$, respectively.

Recall that $w = \sum_i s_i r_i$. Let $w_i = s_i r_i$ be the contribution of row i to this sum. We then want to bound $\Delta w = |w - w'|$. We similarly set Δw_i to be $|s_i r_i - s'_i r'_i|$. Note that for i in A_1 or A_5 we have $\Delta w_i = 0$.

Next, we consider Δw_j the change in the term corresponding to the row where the data changes. As row j is tied with all other rows in region A_2, $r_j = \frac{(a_1+1)+(a_1+a_2)}{2} = \frac{2a_1+a_2+1}{2}$ and $r'_j = \frac{(a_1+a_2+a_3)+(a_1+a_2+a_3+a_4)}{2} = \frac{2a_1+2a_2+2a_3+a_4}{2}$. In the worst case, s_j changes when the move occurs, so

$$\begin{aligned}\Delta w_j < r_j + r'_j &= \frac{2a_1+a_2+1}{2} + \frac{2a_1+2a_2+2a_3+a_4}{2} \\ &= \frac{4a_1+3a_2+2a_3+a_4+1}{2}.\end{aligned}$$

Now we consider Δw_i where $i \in A_2$ (excluding row j). As there is one fewer row tied, each of their ranks (excluding row j) is changed from $\frac{2a_1+a_2+1}{2}$ to $\frac{2a_1+a_2}{2}$, which means that each row is changed by $\frac{1}{2}$. As there are $a_2 - 1$ rows in this region, the contribution to Δw from this set of rows is bounded by $\frac{a_2-1}{2}$.

Similarly, each of the row ranks (excluding row j) in region A_4 is changed from $\frac{2a_1+2a_2+2a_3+a_4+1}{2}$ to $\frac{2a_1+2a_2+2a_3+a_4}{2}$ as one more row is tied with rows in A_4. As there are a_4 rows, Δw from the rows in region A_4 is bounded by $\frac{a_4}{2}$.

Now we consider region A_3. As d_j has moved from before this region to after this region, each of the row ranks is changed by 1. Therefore the contribution to Δw from the rows in region A_3 is bounded by a_3.

It follows that

$$\begin{aligned}\Delta w &= |w - w'| \\ &\leq \frac{4a_1+3a_2+2a_3+a_4+1}{2} + \frac{a_2-1}{2} + \frac{a_4}{2} + a_3 \\ &= 2a_1 + 2a_2 + 2a_3 + a_4 \\ &\leq 2(a_1+a_2+a_3+a_4) \\ &\leq 2n\end{aligned}$$

Having bounded the sensitivity, the proof is complete.

□

As we discussed previously, a test statistic on its own is not useful for statistical inference. Given the test statistic, we must calculate a p-value. In the public case, we compare to a simple normal distribution, but in the private case the distribution of $\widetilde{w}$ under the null distribution is equal to the sum of a normal and a Laplacian. This is difficult to compute analytically, so we simply simulate a large number c of draws from this distribution. (In our experiments, c is set to 10 million. This gives a very accurate estimate of p.) We combine everything into a final test algorithm in Algorithm $\widetilde{\mathcal{C}}$.

```
Algorithm $\widetilde{\mathcal{C}}$ : Complete Wilcoxon Test
Input: $\mathbf{x}$ , $\epsilon$, $c$
begin
    $\widetilde{w} := \widetilde{\mathcal{WP}}(\mathbf{x}, \epsilon)$
    for $k := 1$ to $c$ do
        $W_k \longleftarrow$ Normal$(0, n(n+1)(2n+1)/6)$ + Lap$(2n/\epsilon)$;
    $p \longleftarrow$ fraction of $W_k$ more extreme than $\widetilde{w}$
Output: $\widetilde{w}, p$
```

In this algorithm we draw our reference distribution (the W_k values) assuming there are no $d_i = 0$ rows. The distribution will technically differ slightly when there are many rows with $d_i = 0$, but the difference is inconsequential in all but very extreme circumstances, and even then the calculations are overly conservative. See Appendix A for more details.

Theorem 3.2. *Algorithm $\widetilde{\mathcal{C}}$ is ϵ-differentially private.*

Proof. The computation of $\widetilde{w}$ was already shown to be private. The remaining computation needed to find the p-value does not need access to the database—it is simply post-processing. By Theorem 2.5, it follows that the complete algorithm is also private. □

4 Experimental Results

We assess the power of our differentially-private Wilcoxon signed-rank test first on synthetic data. (For tests with real data, see Section 6.) In order to measure power, we must first fix an effect size. We chose to have the u_i and v_i values both generated according to normal distributions with means one standard deviation apart. We then measure the statistical power of Algorithm $\widetilde{\mathcal{C}}$ (for a given choice of n and ϵ) by repeatedly randomly sampling a database $\mathbf{x}$ from that distribution and then running $\widetilde{\mathcal{C}}$ on that database.[5] The power is the percentage of the time $\widetilde{\mathcal{C}}$ returns a p-value less than α. (We use $\alpha = .05$ in all our experiments.)

We consider several values of ϵ. The lowest, .01 is an extremely conservative privacy parameter and allows for safe composition with many other queries of comparable ϵ value. In Figure 2, we also use ϵs of $.1$ and 1, which, while higher, still provide very meaningful privacy protection. Ultimately, the choice of ϵ is a question of policy and depends on the relative weight placed on privacy and utility. We also measure for comparison the power of the standard, public Wilcoxon test.[6] As one might expect, the amount of data needed to detect a given effect scales roughly proportionately with $1/\epsilon$.

An important part of the algorithm design is the treatment of rows with $d_i = 0$. In Figure 3, 30% of rows have $d_i = 0$ and the other 70% are distributed as before, with the two data points sampled from normal distributions with means one standard deviation apart. We find that while (as expected) all tests lose power, but they lose very little and the relative relationship between our tests and the non-private Wilcoxon test is unchanged.

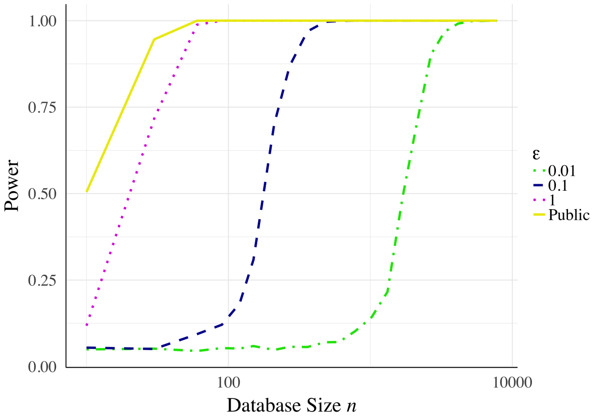

Figure 2: Power at various values of ϵ and sample size n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\alpha = .05$)

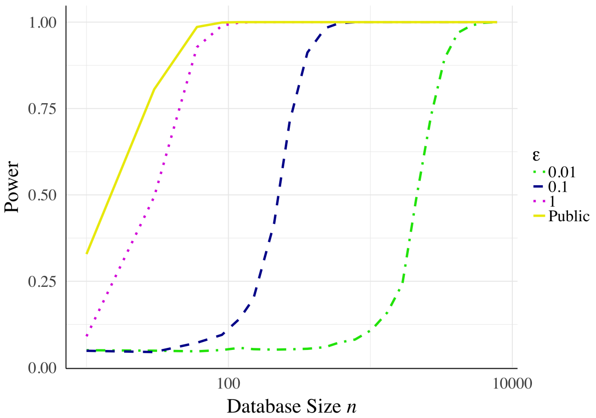

Figure 3: Power at various values of ϵ and sample size n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\alpha = .05$; proportion of zeroes in d_i is $.3$)

In a properly calibrated hypothesis test, the distribution of the p-values under H_0 will be uniform, indi-

[5] Our actual implementation differs slightly from this. To save time when running a huge number of tests with identical n and ϵ, we first generate the reference distribution W_k values, which can be reused across runs.

[6] We use the standard version of the test because it is the one usually used. The Pratt variant has very similar power.

cating that the Type I error rate is exactly α. We verified that Algorithm $\widetilde{\mathcal{C}}$ yields uniform p-values under H_0 in realistic settings. See Appendix A for further discussion.

We also examine the power of our test at varying effect sizes. As shown in Figure 4, at large sample sizes n, in this case 2500, there is essentially no difference between the minimum effect detectable in the private and public setting. This is because the random variation in the sample overwhelms the relatively small random noise being added for privacy. We also note, while not shown in the figure, that for small enough choices of n and ϵ no effect, no matter how large, can be detected. (Once all d_i values are positive, increasing the effect size further has no effect.)

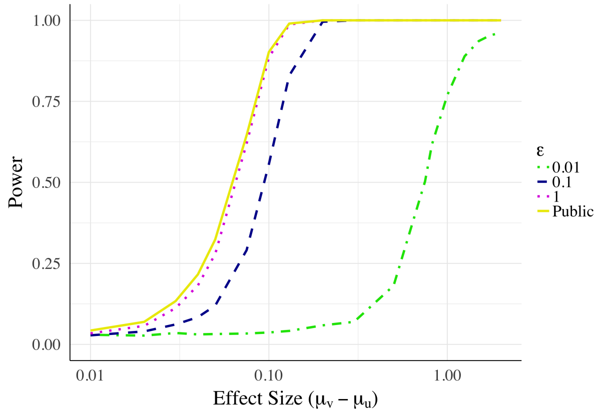

Figure 4: Power at various values of ϵ and effect size. ($\alpha = .05$, $n = 2500$)

5 Comparison to Previous Work

In 2016, Task and Clifton [23] introduced the first differentially private version of the Wilcoxon signed-rank test, from here on referred to as the TC test. Our work improves upon their test in two ways. We describe the two key differences below, and then compare the power of our test to theirs. We also found a significant error in their work.[7] All comparisons are made to our implementation of the TC test with the relevant error corrected.

Computing critical values Task and Clifton compute an analytic upper bound on the critical value t^*. For a given n and ϵ, the private test statistic $\widetilde{W}$ under H_0 is equal to a sum $W + \Lambda$, where W is a random draw from a normal distribution (scaled according to n) and Λ is a Laplace random variable (scaled according to n and ϵ). In particular, say that b is a value such that $\Pr[W > b] < \beta$ and g is a value such that $\Pr[\Lambda > g] < \gamma$. Then we can compute the following bound. (The last line follows from the fact that the two events are independent.)

$$\begin{aligned}\Pr[\widetilde{W} > b + g] &< \Pr[W > b \text{ or } \Lambda > g] \\ &= \Pr[W > b] + \Pr[\Lambda > g] \\ &\quad - \Pr[W > b \text{ and } \Lambda > g] \\ &= \beta + \gamma - \beta\gamma\end{aligned}$$

Task and Clifton always set $\gamma = .01$ and then vary the choice of β such that they have $\alpha = \beta + \gamma - \beta\gamma$ for whatever α is intended as the significance threshold. (This is where Task and Clifton make an error. This formula is correct, but they used an incorrect density function for the Laplace distribution and as a result calculated incorrect values of g.)

The bound described above is correct but is very loose. We instead compute the critical value by simulation. We use 10 million draws from the relevant distribution and compute the $1 - \alpha$ quantile of their absolute values. This gives drastically lower critical values. Table 4 contains examples of the critical values achieved by each method for several parameter choices. More values can be found in Appendix B.

Handling zero values Our second key change from the TC test is that we handle rows with $d_i = 0$ according to the Pratt variant of the Wilcoxon, rather than dropping them completely as is more traditional. The reason the traditional method is so dif-

[7] This error has been confirmed by Task and Clifton in personal correspondence.

Table 4: Critical Value Comparison for $n = 100$

ϵ	α	Public	New	TC
1	0.1	1.282	1.417	2.680
	0.05	1.645	1.826	3.091
	0.025	1.960	2.186	3.511
0.1	0.1	1.282	5.684	14.786
	0.05	1.645	8.063	15.197
	0.025	1.960	10.438	15.617
0.01	0.1	1.282	55.350	135.843
	0.05	1.645	79.233	136.254
	0.025	1.960	103.116	136.674

Critical values for *n* = 100 and several values of ϵ and α. To allow easy comparison, these values are for a normalized W statistic, i.e., W has been divided by the relevant constant so that it is (before the addition of Laplacian noise) distributed according to a standard normal. See Appendix B for the equivalent table at n = 1000.

ficult in the private setting is that the reference distribution one must compare to depends on the number of rows that were dropped. If n_r is the number of non-zero rows (i.e., rows that weren't dropped), one is supposed to look up the critical value associated with n_r, rather than the original size n of the database.

Unfortunately, n_r is a sensitive value and cannot be released privately.[8] Task and Clifton show that it is *acceptable* (in that it does not result in type I error greater than α) to compare to a critical value for a value of n_r that is lower than the actual value. This allows them to give two options for how one might deal with the lack of knowledge about n_r.

High Utility This version of the TC test simply assumes $n_r \geq .3n$ and uses the critical value that would be correct for $n_r = .3n$. We stress that this algorithm is *not* actually differentially private, though it could easily be captured by a sufficiently weakened definition that limited the universe of allowable databases. Another problem is that for most realistic data, n_r is much greater than $.3n$ and using this loose lower bound still results in a large loss of power.

High Privacy This version adds k dummy values to the database with $d_i = \infty$ and k with $d_i = -\infty$.[9] Then one can be certain of the bound $n_r \geq 2k$. This is a guaranteed bound so this variant truly satisfies differential privacy. On the other hand it is a very loose lower bound in most cases, leading to a large loss of power.

Experimental results We compare our test to the TC test by measuring statistical power, just as we compared it to the public test in Section 4. We begin by again measuring the power when detecting the difference between two normal distributions with means one standard deviation apart. The results can be seen in Figure 5. If we look at the database size needed to achieve 80% power, we find that the 32 data points we need, while more than the public test (14), are many fewer than the TC high utility variant (80) or the TC high privacy variant (122). Appendix B includes a figure for $\epsilon = .01$ as well. What we see is that, while all private tests require more data, our test (requiring $n \approx 236$) still requires about 40% as much data as the TC high utility variant (588). The TC high privacy variant, however, scales much less well to low ϵ and requires roughly 2974 data points.

The results in Figure 5 use a continuous distribution for the real data, so there are no data points with $d_i = 0$. Because one of the crucial differences between our algorithms is the method for handling these zero values, we also consider the effect when there are a large number of zeros. In particular, Figures 6 and 7 compare our algorithm to the high utility and high privacy variants of the TC test, respectively. Here we first choose the number of rows with $d_i = 0$ and then sample the remaining data points as before. Of course, as the number of rows showing

[8] A private estimate could be released, but one would have to devote a significant portion of the privacy budget for the hypothesis test to this estimate, greatly decreasing the accuracy/power of $\widetilde{W}$.

[9] Task and Clifton do not discuss how to choose k, and in our experimental comparisons we set $k = 15$, the same value they use.

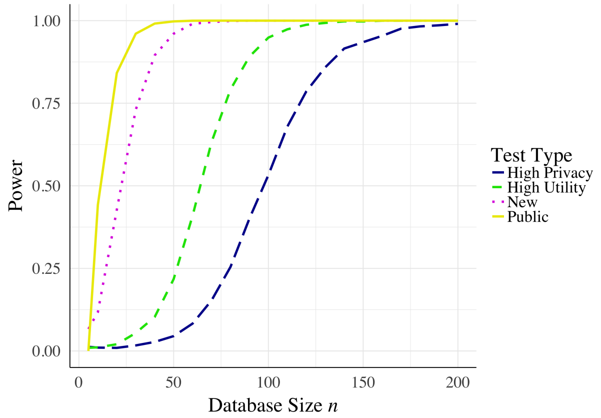

Figure 5: Power comparison of Task and Clifton's [23] algorithms, our new algorithm, and the public algorithm at various n. (Effect size: $\mu_u - \mu_v = 1\sigma$, $\epsilon = 1$; $\alpha = .05$)

no difference increases, the power of all tests decreases, but we see that in all cases our test retains power longer.

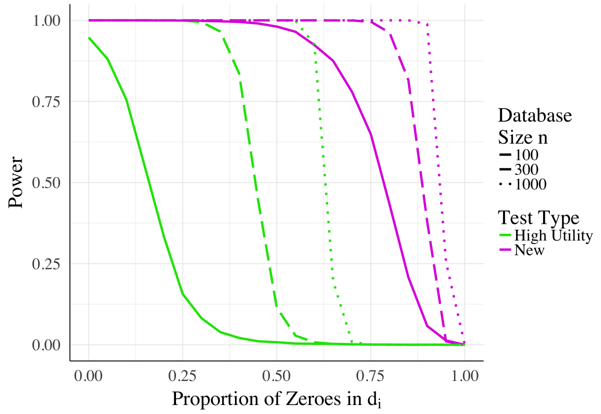

Figure 6: Power comparison of the TC [23] *High Utility* algorithm and our algorithm at various proportions of tied values and sample sizes n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\epsilon = 1$; $\alpha = .05$)

Overall, we see that both in situations with no zero values and situations with many, our test achieves the rigorous privacy guarantees of the TC high privacy test while achieving greater utility than the TC high utility test.

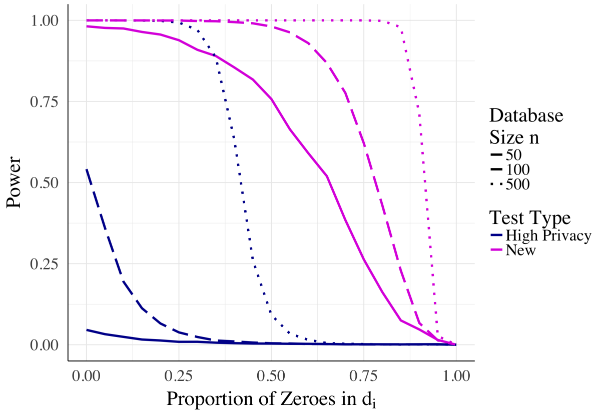

Figure 7: Power comparison of the TC *High Privacy* algorithm ($k = 15$) and our algorithm at various proportions of tied values and sample sizes n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\epsilon = 1$; $\alpha = .05$)

Relative contribution of improvements Given that we make two meaningful changes to the TC test, one might naturally wonder whether both are truly useful or whether the vast majority of the improvement comes from one of the two changes. To test this, we compare to an updated variant of the TC test where we calculate critical values exactly through simulation, as we do in our algorithm, but otherwise run the TC test unchanged. The result is presented in Figure 8. "High Privacy +" and "High Utility +", in Figure 8, refer to algorithms using our critical value computation and their methods of handling data as stated before.

We find the resulting algorithm to rest comfortably between the original TC test and our proposed test. This means that both the change to the critical value calculation and the switch to the Pratt method of handling $d_i = 0$ rows are important contributions to achieving the power of our test.

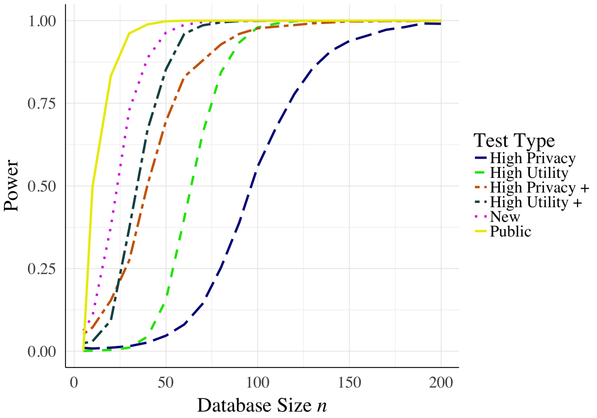

Figure 8: Power comparison of the TC algorithms, the TC algorithms with our critical values (denoted with a +), our new algorithm, and the public algorithm at various sample sizes n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\epsilon = 1$; $\alpha = .05$)

6 Application to Real-World Data

We now demonstrate the use of our algorithm on real-world data. We use a database of New York City tax ride information released in 2014. The database contains information on every Yellow Taxi ride in New York City in 2013, and its release resulted in high-profile de-anonymization attacks [29]. Our main finding is that (at least for some natural analyses that we attempt) a differentially private query interface would have been sufficient, and the release of the data set was unnecessary. We also again compare our statistical power to that of the TC tests and find it superior.[10]

This dataset contains several variables of interest for every Yellow Taxi ride in New York City in 2013, of which the following will be most useful:

- Hack License: a unique identifier for every taxi driver in the city

- Number of Passengers: how many people rode together in the taxi

- Trip Time: the duration of the ride in seconds

- Trip Distance: the total distance traveled during the ride

We subsetted this dataset, initially, to include rides occuring on January 1st and 2nd of 2013. Then, we only kept rides given by a driver who drove on both the 1st and the 2nd. In one data set, u_i and v_i were the average trip time on the 1st and 2nd, respectively. The Wilcoxon test can then be used to test whether trip time varied between the two days. We calculated two other data sets similarly for trip distance and number of passengers.

The resulting data sets have 17,066 entries. We sampled (with replacement) 10^5 different smaller datasets of size $n = 400$. Finally, we ran each algorithm on each of these resulting data sets and report the proportion in which a significant result (at $\alpha = .05$) was found. The results are summarized in Figure 9.

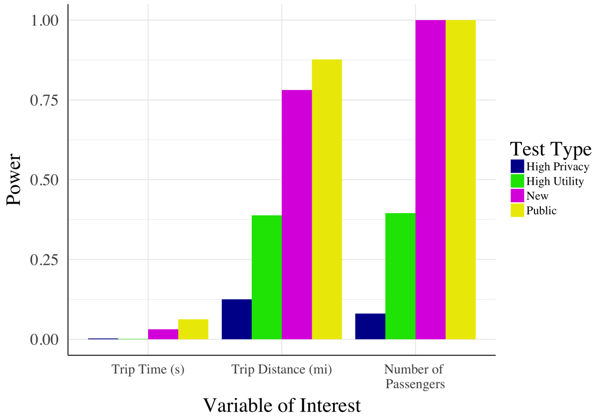

Figure 9: Power of the TC tests, our test, and the public test at sample sizes of 400, sampled with replacement from the NYC taxi data. ($\epsilon = 1$; $\alpha = .05$)

The power of our test follows the public test closely, and we see that for a data set as small as 400, we can achieve results with a private query that are almost as useful as a full public release of the data. The TC algorithms do not achieve this goal, though the high utility variant is still meaningfully

[10] This data set is the same one Task and Clifton originally use for evaluating their test.

useful. Of course, there will be choices of n and ϵ for which the gap between our power and the power of the public test is quite large, but our point here is to argue that for even reasonably small data sets a private query interface can be sufficient for many important tasks.

7 Conclusion and Open Questions

We have shown a new private variant of the Wilcoxon hypothesis test complete with an accurate method for computing p-values. It contains two significant technical innovations over prior work. We show that this test has much higher power than the previously available tests, and (at least for $\epsilon = 1$) its power is actually not very far from that of the public test. We also show that it is a realistic algorithm to run on at least some real-world data sets at reasonable database sizes for many applications. We think this is a major step forward, but several questions remain open.

The most obvious goal is simply to continue improving the power of the test. The gap between our test and the public test is still significant when ϵ is meaningfully less than 1.

It would also be useful to compare this test to other tests in the statistics literature. For example, if one knows that the underlying data is drawn from a normal distribution, classical statistics would use a paired-sample t-test, which is known to have higher power in this case than the Wilcoxon. We suspect that the Wilcoxon test holds up better under the constraints of differential privacy and that it might actually be better in the private setting to use the private Wilcoxon *even when* the data is known to be normally distributed. Unfortunately, to our knowledge no one has yet developed a private analogue of the t-test, so we are unable to make the comparison.[11]

[11] This material is based upon work supported by (funding sources withheld for anonymity).

References

[1] Michael Barbaro and Tom Zeller Jr. A face is exposed for aol searcher no. 4417749. *The New York Times*, page A1, Aug 2006.

[2] Andrés F Barrientos, Jerome P Reiter, Ashwin Machanavajjhala, and Yan Chen. Differentially private significance tests for regression coefficients. *arXiv preprint arXiv:1705.09561*, 2017.

[3] Zachary Campbell, Andrew Bray, Anna M. Ritz, and Adam Groce. Differentially private anova testing. *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 281–285, 2018.

[4] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[5] William Jay Conover. On methods of handling ties in the wilcoxon signed-rank test. *Journal of the American Statistical Association*, 68(344):985–988, 1973.

[6] Bolin Ding, Harsha Nori, Paul Li, and Joshua Allen. Comparing population means under local differential privacy: With significance and power, 2018.

[7] Vito D'Orazio, James Honaker, and Gary King. Differential privacy for social science inference. 2015.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[9] Stephen E Fienberg, Aleksandra Slavkovic, and Caroline Uhler. Privacy preserving gwas data sharing. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 628–635. IEEE, 2011.

[10] Marco Gaboardi, Hyun-Woo Lim, Ryan M Rogers, and Salil P Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR, 2016.

[11] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[12] Aaron Johnson and Vitaly 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*, pages 1079–1087. ACM, 2013.

[13] Thông T Nguyên and Siu Cheung Hui. Differentially private regression for discrete-time survival analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1199–1208. ACM, 2017.

[14] John W Pratt. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):655–667, 1959.

[15] Ryan Rogers and Daniel Kifer. A new class of private chi-square hypothesis tests. In *Artificial Intelligence and Statistics*, pages 991–1000, 2017.

[16] Or Sheffet. Differentially private ordinary least squares. *arXiv preprint arXiv:1507.02482*, 2015.

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

[18] Adam Smith. Efficient, differentially private point estimators. *arXiv preprint arXiv:0809.4794*, 2008.

[19] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

[20] Eftychia Solea. Differentially private hypothesis testing for normal random variables. 2014.

[21] Ben Stoddard, Yan Chen, and Ashwin Machanavajjhala. Differentially private algorithms for empirical machine learning. *arXiv preprint arXiv:1411.5428*, 2014.

[22] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[23] Christine Task and Chris Clifton. Differentially private significance testing on paired-sample data. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 153–161. SIAM, 2016.

[24] Caroline Uhlerop, Aleksandra Slavković, and Stephen E Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of privacy and confidentiality*, 5(1):137, 2013.

[25] Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 138–143. IEEE, 2009.

[26] Yue Wang, Jaewoo Lee, and Daniel Kifer. Differentially private hypothesis testing, revisited. *CoRR*, abs/1511.03376, 2015.

[27] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

[28] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[29] Chris Whong. FOILing NYC's taxi trip data. 2014.

[30] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[31] Fei Yu, Michal Rybar, Caroline Uhler, and Stephen E Fienberg. Differentially private logistic regression for detecting multiple-snp association in gwas databases. In *International Conference on Privacy in Statistical Databases*, pages 170–184. Springer, 2014.

[32] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: Regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 5(11):1364–1375, 2012.

A Uniformity of p-values

The uniformity of p-values for a hypothesis test can be assessd by simulating many p-values and evaluating how well they follow the uniform distribution on the unit interval. A common tool for this purpose is the quantile-quantile (or Q-Q) plot, which plots the quantiles of the theoretical distribution (the uniform in this case) against the empirical quantiles of the p-values. A sample that perfectly follows the theoretical distribution will coincide at all quantiles and be represented on the Q-Q plot as the identity line.

Figure 10 shows a Q-Q plot of three sets of p-values, all generated under H_0, with $\epsilon = 1$, $n = 500$. When there are no ties in the original data (0% of $d_i = 0$), the Q-Q plot line is indistinguishable from the identity line, indicating that the test is properly calibrated. Encouragingly, introducing a substantial number of ties into the data (30% of $d_i = 0$) has no noticible effect.

In order to induce non-uniformity in the p-values, one needs an extremely high proportion of rows with $d_i = 0$. The curve with 90% zero values is shown as an illustration. When the proportion of zeros is very high, the variance of the p-values will be narrower than the reference distribution, resulting in a lower critical value. Since the value we are using is higher, our test is overly conservative,[12] but this is acceptable as type I error is still below α.

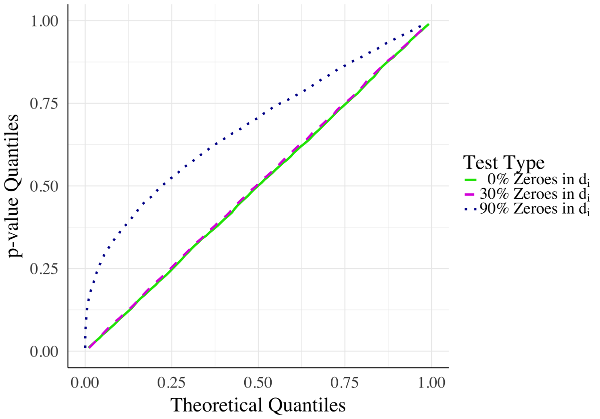

Figure 10: A quantile-quantile plot comparing the distribution of simulated p-values to the uniform distribution ($\epsilon = 1$, $n = 500$).

[12] One could try to estimate the number of zeros to be less conservative, but that would require allocating some of the privacy budget towards that estimate, which is not worth it in most circumstances.

B Additional Figures

B.1 Power Comparison

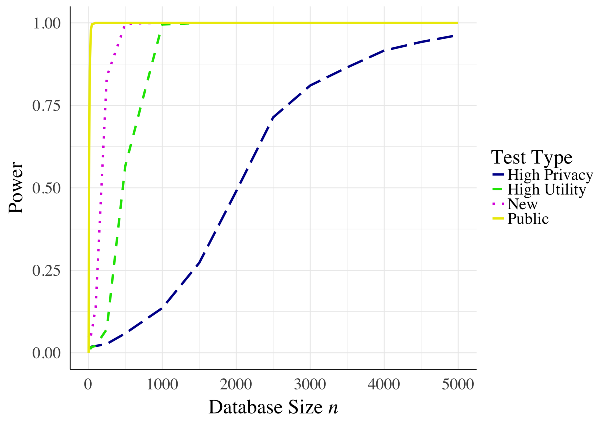

Figure 11: Power comparison of Task and Clifton's algorithms, our new algorithm, and the public algorithm at various database sizes n. (Effect size: $\mu_u - \mu_v = 1\sigma$; $\epsilon = .1$; $\alpha = .05$)

B.2 Critical Value Tables

Table 5: Critical Value Comparison for $n = 1000$

ϵ	α	Public	New	TC
1	0.1	1.282	1.296	1.763
	0.05	1.645	1.665	2.174
	0.025	1.960	1.984	2.594
0.1	0.1	1.282	2.203	5.617
	0.05	1.645	2.975	6.028
	0.025	1.960	3.740	6.448
0.01	0.1	1.282	17.681	44.157
	0.05	1.645	25.234	44.568
	0.025	1.960	32.844	44.988

Critical values for *n* = 1000 and several values of ϵ and α. To allow easy comparison, these values are for a normalized W statistic, i.e., W has been divided by the relevant constant so that it is (before the addition of Laplacian noise) distributed according to a standard normal.

Table 6: New Critical Value Table for ϵ = 1.0

n	0.05	0.025	0.01	0.005
10	70	83	102	116
20	155	183	220	248
30	256	299	355	397
40	369	429	506	562
50	494	572	670	742
75	854	984	1143	1257
100	1271	1460	1690	1853
200	3402	3895	4486	4900
300	6127	7012	8069	8798
400	9335	10679	12276	13382
500	12978	14845	17061	18592
1000	36235	41443	47637	51906

Critical values at several sample sizes n and two-sided significance levels α. To calcuate these values, we run 10 million simulations for each parameter combination and compute the $1 - \alpha$th percentile of the absolute value of the distribution.

Table 7: New Critical Value Table for ϵ = 0.1

n	0.05	0.025	0.01	0.005
10	600	739	922	1061
20	1202	1479	1846	2123
30	1806	2220	2770	3185
40	2413	2968	3704	4261
50	3018	3713	4628	5324
75	4541	5577	6954	7989
100	6073	7461	9294	10677
200	12328	15098	18767	21531
300	18733	22892	28391	32519
400	25296	30837	38193	43736
500	32054	38979	48128	55083
1000	68258	82120	100408	114230

Critical values at several sample sizes n and two-sided significance levels α. To calcuate these values, we run 10 million simulations for each parameter combination and compute the $1-\alpha$th percentile of the absolute value of the distribution.

Table 8: New Critical Value Table for ϵ = 0.01

n	0.05	0.025	0.01	0.005
10	5992	7377	9209	10596
20	11971	14742	18416	21196
30	17976	22137	27644	31774
40	23974	29516	36877	42425
50	29964	36905	46081	53034
75	44933	55371	69105	79513
100	59921	73792	92066	106005
200	119902	147619	184222	212010
300	179942	221477	276678	317895
400	239695	295106	368374	423528
500	299627	368763	460256	529522
1000	600096	738071	921529	1061150

Critical values at several sample sizes n and two-sided significance levels α. To calcuate these values, we run 10 million simulations for each parameter combination and compute the $1-\alpha$th percentile of the absolute value of the distribution.