A Quantitative Approach for Evaluating the Utility of a Differentially Private Behavioral Science Dataset

Article · March 2015

DOI: 10.1109/ICHI.2014.45

CITATION
1

READS
89

6 authors, including:

Raquel Hill
Grand Canyon University
17 PUBLICATIONS **146** CITATIONS
SEE PROFILE

Michael Hansen
United States Air Force Research Laboratory
12 PUBLICATIONS **384** CITATIONS
SEE PROFILE

Erick Janssen
Institute for Family and Sexuality Studies
201 PUBLICATIONS **8,506** CITATIONS
SEE PROFILE

Stephanie Sanders
Indiana University Bloomington
228 PUBLICATIONS **11,624** CITATIONS
SEE PROFILE

A Quantitative Approach for Evaluating the Utility of a Differentially Private Behavioral Science Dataset

Raquel Hill
and Michael Hansen
School of Informatics
Indiana University
Bloomington, IN 47405
Email: ralhill,mihansen@indiana.edu

Erick Janssen
and Stephanie A. Sanders
and Julia R. Heiman
Kinsey Institute
Indiana University
Bloomington, IN 47405
Email: ejanssen, sanders, jheiman@indiana.edu

Li Xiong
Emory University
Atlanta, Georgia 30322
Email: lxiong@mathcs.emory.edu

***Abstract*—Objective: Social scientists who collect large amounts of medical data value the privacy of their survey participants. As they follow participants through longitudinal studies, they develop unique profiles of these individuals. A growing challenge for these researchers is to maintain the privacy of their study participants, while sharing their data to facilitate research. Differential privacy is a new mechanism which promises improved privacy guarantees for statistical databases. We evaluate the utility of a differentially private dataset. Our results align with the theory of differential privacy and show when the number of records in the database is sufficiently larger than the number of cells covered by a database query, the number of statistical tests with results close to those performed on original data increases.**

I. Introduction

As part of human subjects' protections, researchers are required to analyze and minimize all potential risks to research participants connected with the study procedures. These include physical, psychological, social, legal, loss of confidentiality or other potential risks. The researcher needs to consider all of the potential risks including whether "any disclosure of the subjects' responses could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, insurability, or reputation." [1] Therefore, for research that results in datasets containing sensitive personal information, loss of confidentiality/privacy is often the primary risk that must be managed.

In sexuality research for example, information about a person's marital infidelity or whether they have had a sexually transmitted infection could easily affect their reputation or be used in legal cases such as divorce proceedings. Therefore, protection of the identities of research participants is critical to the ongoing process of scientific investigations that deal with sensitive information. As researchers may desire to share datasets with other scientists, or be legally required to provide access to their data, finding ways to assure that identities are protected in shared or archival datasets that may be subject to disclosure is of significant importance.

Disclosure of research data may be compelled by subpoena or by institutional interpretations of "open-records laws [2]." Although some research data involving sensitive information may be protected by Certificates of Confidentiality, most is not. Certificates of Confidentiality, authorized by the U.S. Department of Health and Human Services, are meant to protect sensitive data from subpoena, but the degree of protection may be limited [2]. Therefore, researchers may desire to de-identify data as soon as possible and guard against possible statistical re-identification in the interests of protecting human subjects and the research enterprise. Exploring ways in which data may be perturbed while the scientific usefulness of the data is preserved may lead to new techniques for researcher to share datasets with other researchers and archive datasets for future analyses in ways that minimize potential risks to subjects.

The dataset that we study contains medical, sexual, demographic, and psychological information, all of which are, or can be, directly relevant to health of subsamples and individuals. For example, a given disease status may interact with sexual activity and economic variables, and collecting these data carries critical security and privacy concerns–similar to specific biologically derived data. So the data here overlaps with biomedical data in its power to predict health and illness and its sensitivity to persons.

A study's participants can be vulnerable to privacy violations if, for example, an adversary already has some information about the person he or she is targeting [3]. The study participant's privacy may be at risk even if the dataset contains no explicitly identifying information. It is therefore important to protect these data beyond any standard removal of clearly identifying information. However, the goal of protecting privacy must be balanced with the need to conserve the data's research utility. Data protection techniques should optimize privacy protection but not at the cost of data utility.

In this work we evaluate Differential Privacy (hereafter referred to as DP) as a technique to protect social science datasets while preserving their research utility. DP is a data perturbation framework that provides strong and formal privacy guarantees [4]. The essential goal is to prevent a possible adversary from discovering whether or not some specific individual's data is present in a differentially private dataset, given some risk threshold. While there are many theoretical results ('in vitro') for DP and utility outside of actual data cases [5], [4], [6], [7], very little has been done to evaluate its effect on data utility in a real-world research setting ('in vivo'). Since it offers such promising privacy guarantees, we examine

DP as a possible mechanism for protecting large social science datasets.

To evaluate whether or not the DP algorithms preserve utility, we measure the mean absolute deviation between results obtained from the original and differentially private data. These analyses fall into two different categories: (1) multivariate logistic regression, (2) feature importance. The uses cases for the multivariate logistic regressions are derived from analysis presented in [8]. To assess utility, we compare the distances between the odds ratios for regressions performed on the original and DP data. Distances that approach zero would tend to indicate a utility-preserving result. We also compare the p-values derived from analysis on the original and DP data to determine whether decisions to accept or reject the null hypothesis change for DP data. For the latter analysis, we compare the relative importances of features (i.e., dimensions, variables) in the DP histograms to those in the original data. Our results confirm that dimensionality is a major challenge for DP algorithms, especially when the number of records in the database is sufficiently less than the number of cells covered by the query. One very interesting finding is that all algorithms produced noisy histograms that had strong results for Feature Importance.

The remainder of this paper is organized as follows. In the following section, we discuss background and related work. Next, we describe the differential privacy algorithm and the methodology that we use to evaluate it. Afterwards, we present the results of our evaluation. Finally, we conclude and discuss the limitations of our empirical analysis.

II. Background

Differential privacy (DP) has recently emerged as one of the strongest privacy guarantees for statistical data release. A statistical aggregation or computation is differentially private, or satisfies differential privacy, if the outcome is formally indistinguishable when run with and without any particular record in the dataset. The level of indistinguishability is quantified as a privacy parameter. A common differential privacy mechanism is to add calibrated noise to a statistical aggregation or computation result determined by the privacy parameter and the sensitivity of the computation to the inclusion or exclusion of any individual record. While traditional de-identification methods [9] or perturbation methods that add noise to individual values of data records [10] are subject to re-identification or data reconstruction attacks depending on the background knowledge of an adversary, a differential privacy mechanism adds noise to statistical aggregation or computation outputs and provides a strong and provable privacy guarantee with little assumptions on the background knowledge of an adversary.

A. Differential Privacy- Definition

Differential privacy guarantees that if an adversary knows complete information of all the tuples in D except one, the output of a differentially private randomized algorithm should not give the adversary too much additional information about the remaining tuples. We say datasets D and D' differing in only one tuple if we can obtain D' by removing or adding only one tuple from D. A formal definition of differential privacy is given as follows:

Definition II.1 (ϵ-differential privacy [5])**.** *Let $\mathcal{A}$ be a randomized algorithm over two datasets D and D' differing in only one tuple, and let $\mathcal{O}$ be any arbitrary set of possible outputs of $\mathcal{A}$. Algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if and only if the following holds:*

$$Pr[\mathcal{A}(D) \in \mathcal{O}] \leq e^{\epsilon} Pr[\mathcal{A}(D') \in \mathcal{O}]$$

Intuitively, differential privacy ensures that the released output distribution of $\mathcal{A}$ remains nearly the same whether or not an individual tuple is in the dataset.

The most common mechanism to achieve differential privacy is the Laplace mechanism [5] that adds a small amount of independent noise to the output of a numeric function f to fulfill ϵ-differential privacy of releasing f, where the noise is drawn from *Laplace distribution* with a probability density function $Pr[\eta = x] = \frac{1}{2b} e^{-\frac{|x|}{b}}$. A Laplace noise has a variance $2b^2$ with a magnitude of b. The magnitude b of the noise depends on the concept of *sensitivity* which is defined as follows.

Definition II.2 (Sensitivity [5])**.** *Let f denote a numeric function and the sensitivity of f is defined as the maximal L_1-norm distance between the outputs of f over the two datasets D and D' which differs in only one tuple. Formally,*

$$\Delta_f = max_{D,D'} ||f(D) - f(D')||_1.$$

With the concept of sensitivity, the noise follows a zero-mean Laplace distribution with the magnitude $b = \frac{\Delta_f}{\epsilon}$. To fulfill ϵ-differential privacy for a numeric function f over D, it is sufficient to publish $f(D) + X$, where X is drawn from $Lap(\frac{\Delta_f}{\epsilon})$.

For a sequence of differentially private mechanisms, the composability [11] theorems guarantee the overall privacy.

Theorem II.1 (Sequential Composition [11])**.** *For a sequence of n mechanisms $M_1, \ldots, M_n$ and each M_i provides ϵ_i-differential privacy, the sequence of M_i provides $(\sum_{i=1}^{n} \epsilon_i)$-differential privacy.*

Theorem II.2 (Parallel Composition [11])**.** *If D_i are disjoint subsets of the original database and M_i provides α-differential privacy for each D_i, then the sequence of M_i provides α-differential privacy.*

B. Related Work

Dankar et al. [12] provide a thorough treatment of the state of the art in differential privacy. They also outline some of the limitations of the model and the various mechanisms that have been proposed to implement it. In addition, this work discusses several recent applications of differential privacy [13], [14], [15], [16], [17], [18], [19]. In this paper, we evalute the output of several DP processes, including cell-based [13], range query [19], and space partitioning [13]. The space partitioning approach differs from the basic cell-based and range query approaches in that it attempts to preserve the characteristics within the original data by adding noise uniformly to cells that belongs to a partitioned group. Xiao et al. [13], [20] uses a kd-tree (k-dimensional tree) to partition the data. A kd-tree is a space partitioning data structure for organizing data points in a k-dimensional space. First, the DP algorithm partitions the data

D based on the domain and adds noise to each cell to create a synthetic dataset D'. D' is then partitioned using a kd-tree algorithm. The resulting keys from the kd-tree partitioning are then used to subdivide the original dataset. Finally, Laplace noise is added to each partition's count. Each cell within a partition is assigned the value of its partition's $noisy\_count/\beta$, where β is the number of cells within the partition. The perturbed dataset is used in the kd-tree phase of the algorithm so as not to waste the privacy budget on accessing the original dataset multiple times during the partitioning phase.

Xiao et al. [13] evaluate the utility of their DP mechanism by comparing query counts of the original data to that of the differentially private data. We build upon this work by performing predictive analysis on a large social science dataset and the corresponding differentially private data. We derive our use cases from the actual analyses that were previously performed by the researchers who collected and evaluated the original data [8].

Researchers have proposed a variety of approaches for managing biomedical data and protecting patient information. El Emam et al. [21] consider extensions to k-anonymity in the context of two attack scenarios: one in which the attacker wants to re-identify a specific individual that he/she knows in the anonymized dataset (called the prosecutor scenario), and one in which the attacker simply wants to demonstrate that an arbitrary individual from some population could be re-identified in the dataset (called the journalist scenario). The best k-anonymity extension selects an appropriate k using hypothesis testing and a truncated-at-zero Poisson distribution. While this method out-performs standard k-anonymity in terms of information loss (computed using the discernability metric) on their sample datasets, the authors acknowledge that increasing the number of quasi-identifiers may lead to unacceptable amounts of information loss, even for small values of k. Additionally, there was no discussion about how different values of the discernability metric may impact the results of common statistical analyses performed by consumers of an anonymized dataset (i.e., how much information loss will a logistic regression tolerate?).

Brown et al. [22] present a "distributed" query system designed to allow data holds to maintain physical control of their data. This system is contrasted with a centralized database, where users submit queries outside of their local firewalls (and also receive results from a remote server). Data privacy is maintained by physically co-locating the query system software with the data. The system does not attempt to anonymize data for access by a third party.

Murphy et al. [23] present the i2b2 system (integrating biology and the bedside), which provides graded access to patient data depending on the privacy level of the user. At the lowest level, only aggregate counts with Gaussian noise added is available. The problem of multiple queries allowing for convergence on the true count is discussed, but only solved for single user accounts (a user with multiple accounts could still discover the true count). For all other privacy levels, some form of anonymized patient data is available to the user, with the highest level having access to the original data. The de-identification methods for this data are not discussed (they are only listed as HIPAA compliant), and neither is the link between dataset dimensionality and re-identification.

Kushida et al. [24] perform a literature survey on de-identification and anonymization techniques. They focus on three main scenarios: free-text fields, images, and biological samples. For free-text fields, statistical learning-based systems provide the best performance (at or above manual de-identification). Anonymization techniques for images and biological samples are briefly discussed as well. Unfortunately, there is no mention of data privacy methods for datasets with coded fields (e.g., surveys) or dimensionality – i.e., how much easier is re-identification with high-dimensional data?

Bredfeldt et al [25] develop a set of templates and a common zip-file directory structure for multi-site research collaboration with sensitive data. Additionally, some best practices are identified, such as not transferring the zip-file over e-mail or any unencrypted protocol. While the templates and structure provide support for collaboration, details for protecting the data itself (via encryption or some privacy mechanism) is not discussed.

III. Methodology

Our DP application outputs a differentially private contingency table, which is a histogram of counts for all possible attributes settings. We use either a simple *cell-based* or *k-d tree partitioning* method. The cell-based method adds Laplacian noise to each histogram cell (or bin) independently using the perturbation (privacy) ϵ parameter. The k-d tree method adds noise in two stages. First, it applies cell-based noise using an ϵ parameter. Next, it partitions the dataset based on both an entropy threshold (ET) and information gain (IG) parameter, and applies Laplacian noise to each partition independently.

The parameter ϵ determines the level of privacy, with a lower value providing more privacy. For the k-d tree algorithm, the ET parameter determines the entropy, or uniformity, within each partition. The lower the value of ET, the less uniform partitions will be. In our experiments, we vary ϵ over the range $[0.1, 2.0]$ and ET over the range $[0.4, 1.0]$. These ranges represent very low and very high privacy, and cover the true entropies of the datasets. For all our experiments, we fix the IG parameter at 0.0001.

For comparison with the cell-based and k-d tree algorithms, we also generated DP histograms from our datasets using wavelet transforms [19]. We generated *basic* and *adhoc* DP histograms with the same ϵ parameter range. The *adhoc* histograms were created with the $Privelet^*$ algorithm from [19], and all variables were considered as having a small domain. The *basic* histograms were created with the same type of algorithm as our cell-based histograms (i.e., adding Laplacian noise directly to the contingency table). Because the wavelet code and cell-based algorithm compute the λ parameter for Laplacian noise differently, however, we have included both cell-based and *basic* histograms for comparison.

We generate 25 DP histograms for each set of parameter settings, and compare the mean results from these 25 sample histograms to the original. We make two assumptions about the use of DP in this context. The first is that the differentially private histogram is generated once for all variables in a given set and released. That is, we assume non-interactivity (i.e. all queries are know apriori). The second is that once a party has the differentially private histogram, they are free to do with

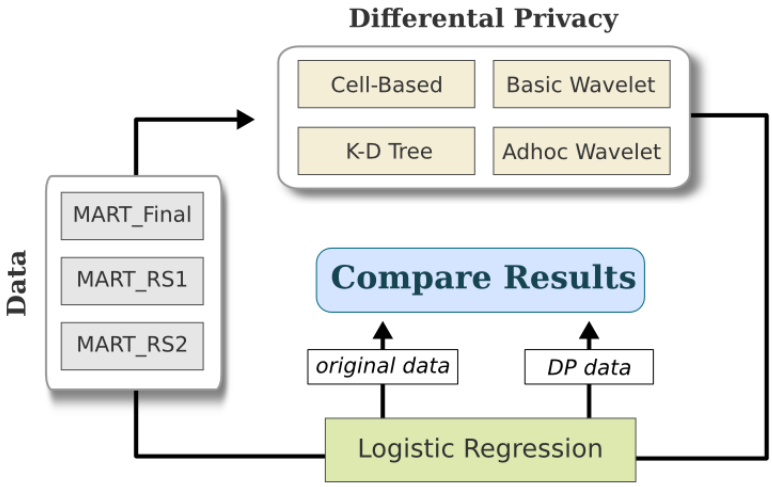


Fig. 1: Experiment Flow Chart

it what they please. This includes reconstructing data records from the histogram bins. Figure 1 illustrates a flow chart of the experiment setup.

A. The Data

Our datasets contains 5,887 participants who have answered a subset of 332 questions from a sexual health survey. The participants are a convenience sample of individuals within the USA. The survey contains multiple modules and most modules are standardized sexual health *scales* – a questionnaire with which you hope to measure a psychological construct using multiple questions (i.e., items or indicators) [26]. The steps for developing a scale include the development/writing of questions/items that are relevant to the construct. After a large number of subjects answer all the questions, factor analysis is used to select items that are relevant to the construct and eliminate ones that are not. One of the steps in factor analysis is to remove items that are very skewed (e.g., if 90% of participants answer that they 'can be shy at times').

B. Use Cases

The use cases we examine are derived from The Kinsey Institute's work in [8], which looks at predictors of unprotected sex and unplanned pregnancy. The specific use cases use logistic regression to evaluate the likelihood of a participant reporting an unplanned pregnancy and the likelihood of a participant reporting having had unsafe sex in the last 12 months (both binary outcomes). When generating our differentially private histograms, we use the following predictor variables:

- **age** - Age of the participant (31 levels). Range is from 18-80 years old.
- **employ** - Employment status (4 levels): full-time, part-time, unemployed, temp/seasonal.
- **gender** - Gender (2 levels): male, female
- **income** - Income level (4 levels): poor, lower, middle, upper
- **relation** - Relationship status (3 levels): none, exclusive, non-exclusive
- **sis5** - Use of safe sex products causes arousal loss (4 levels): strongly agree, agree, disagree, strongly disagree
- **sis8** - Risk of pregnancy inhibits arousal (4 levels): strongly agree, agree, disagree, strongly disagree

Our *response*, or predicted variables are:

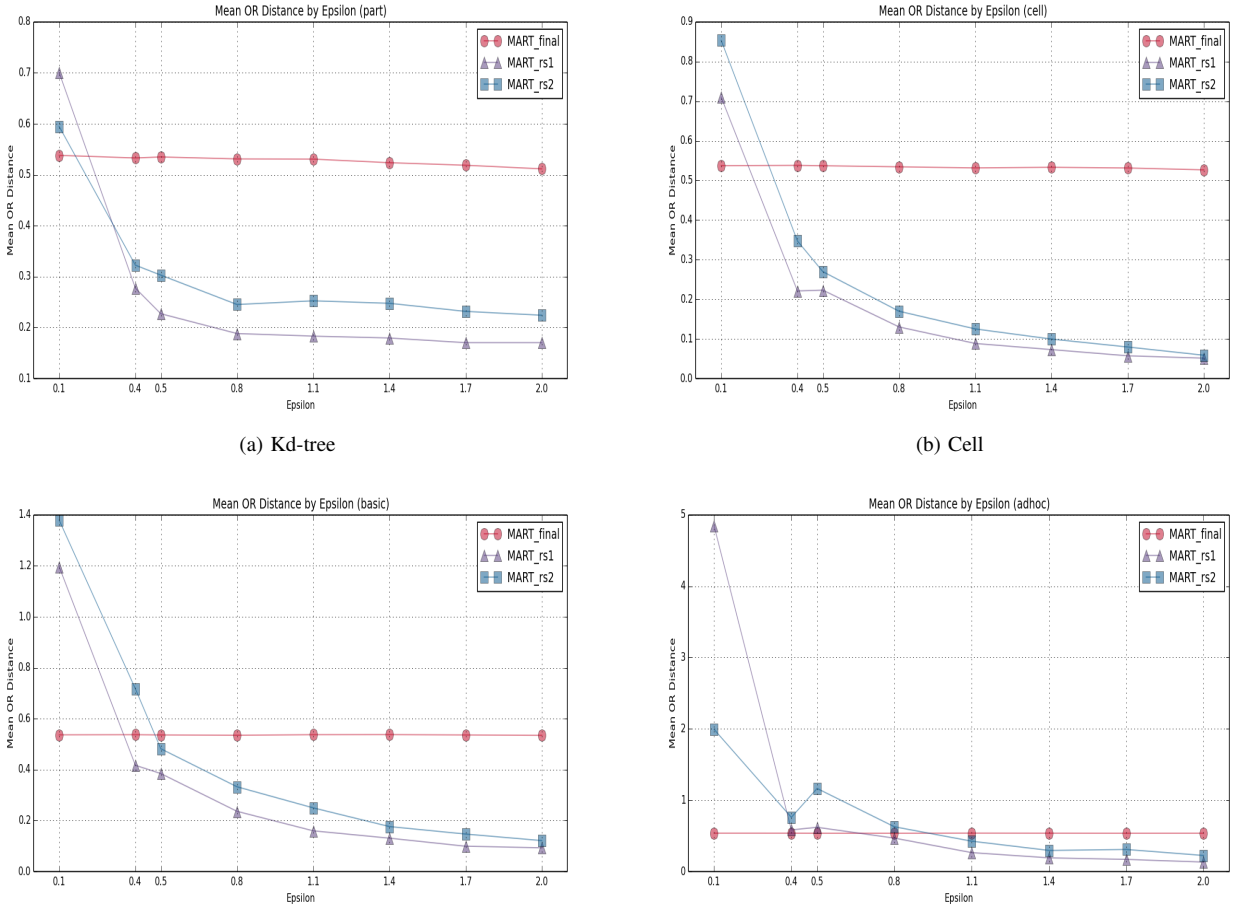
- **kisbq18** - Had unprotected vaginal sex in the last 12 months (binary)
- **kisbq20** - Ever had an unplanned pregnancy (binary)

In addition to the full set of 9 variables (7 predictors, 2 response), we generate DP histograms and measure utility with two reduced datasets (derived from the full dataset). These reduced datasets each include two predictors and one response variable, which allows us to evaluate the DP algorithms on lower-dimensional data. The variables included in each dataset are:

- **Full Set**: 7 predictor variables (age, gender, relation, employ, income, sis5, sis8) and 2 response variables (kisbq18, and kisbq20).
- **Reduced Set 1 (RS1)**: 2 predictor variables (income, sis5) and 1 response variable (kisbq18).
- **Reduced Set 2 (RS2)**: 2 predictor variables (relation, sis8) and 1 response variable (kisbq20).

C. Utility Measures

The size of the database directly impacts the accuracy of a DP query. When $n \ll y$, where n is the number of records within the database and y is the number of cells covered by the query, then the query results will be inaccurate [4], [12]. For the Full Set, RS1 and RS2, there are 190,464, 32, and 24 cells



(a) Kd-tree

(b) Cell

(c) Basic

(d) Adhoc

Fig. 2: Logistic Results: Odds Ratio Distance

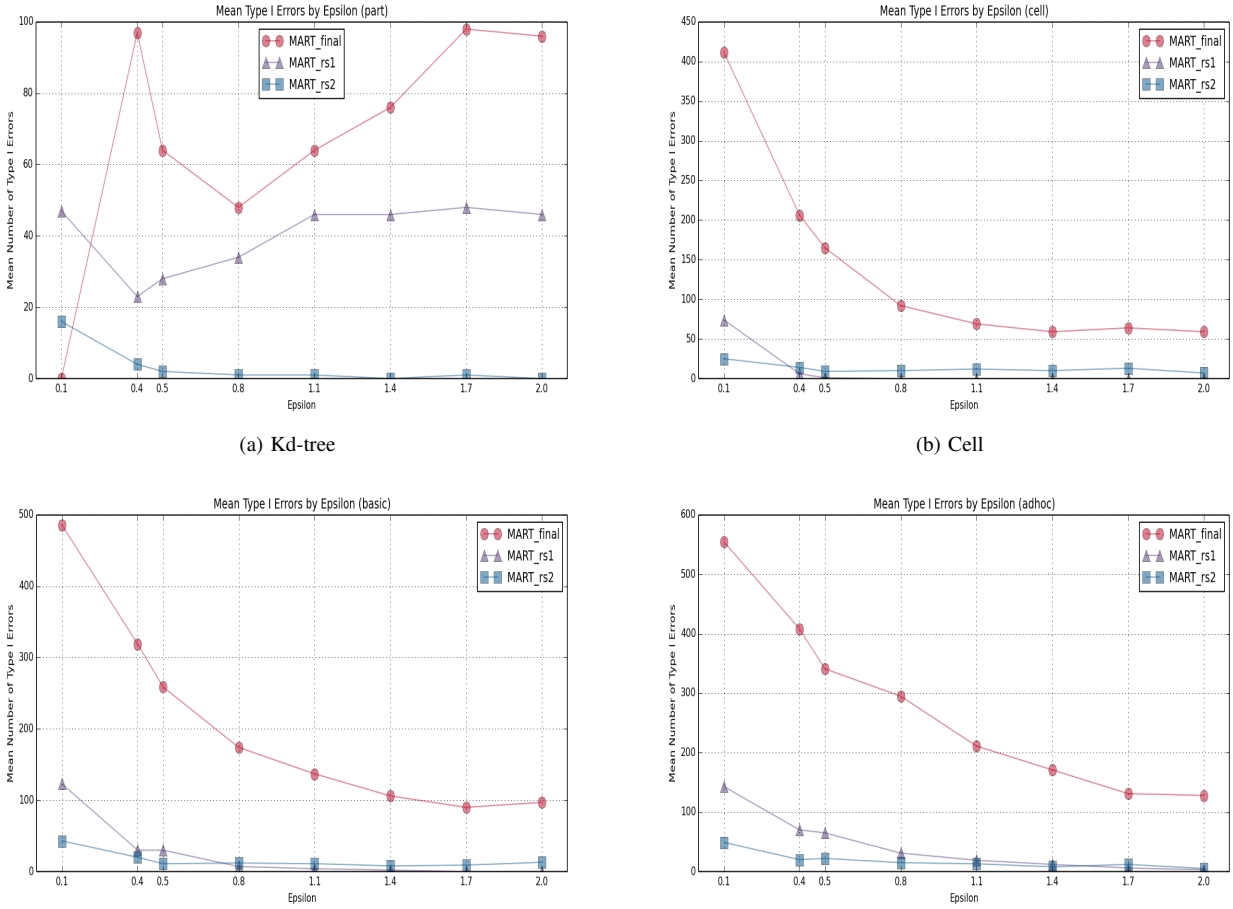
covered by query respectively [1]. To evaluate the effect that the size of the database has on the accuracy of our results, we compare the results across the three datasets.

To evaluate whether or not the DP algorithms preserve utility, we measure the mean absolute deviation between results obtained from the original and differentially private data. These analyses fall into two different categories: (1) multivariate logistic regression, (2) feature importance. We generated 25 differentially private histograms for every dataset (full, RS1, RS2), DP algorithm (cell-based, k-d tree), ϵ value $[0.1, 0.4, 0.5, 0.8, 1.1, 1.4, 1.7, 2.0]$, and ET value $[0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, E_i]$ (where E_i is the entropy of the original histogram).

Logistic Regression For every DP histogram, we ran several different multivariate logistic regressions corresponding to the use cases presented in [8]. Data was separated by gender and by response variables. For the full dataset, there were 2 response variable (kisbq18 and kisbq20), making a total of 4 regressions. There are 12 predictor levels for each regression, which gives 48 runs per parameter setting(s) (ϵ, ET). The reduced datasets each contained a single response variable, so only 2 regressions (male and female) were run per reduced set. There are 6 predictor levels for each regression, which gives a total of 24 runs. For every set of parameter values and regression, we measured the mean absolute deviation between the 25 DP odds ratios of each predictor level (e.g., full-time for `employ`) to the original regression results. We use Bonferroni correction to address the issue of multiple testing within our experiments.

Feature Importance. We compare the relative importances of features (i.e., dimensions, variables) in the DP histograms to those in the original data. Using the extremely randomized trees (ERT) classifier from the sklearn library [27], we compute a rank order of each feature in the dataset from most to least important for prediction. Similar to the logistic and classifier utility methods, the data are first split by gender and use case (unsafe sex, unplanned pregnancy). Next, an ERT classifier generates and trains randomized decision trees on the data. Features are ranked by their relative predictive power within the forest of decision trees, and this ranking is compared to the original results. If the original and DP rankings match exactly, we consider utility to have been preserved in the DP histogram

[1]The number of cells that are covered by a query was calculated by taking the product of the number of different responses for each variable. See Section III-B for a description of the variables.



(a) Kd-tree

(b) Cell

(c) Basic

(d) Adhoc

Fig. 3: Logistic Results: Type I Errors

(i.e., a good run).

IV. Results

Before evaluating the results, we first compare the size of our database n to the number of cells that are covered for each use case. Recall that $n = 5887$ in the original dataset, while the cell coverage for the Full Set use case is 190,464 cells. Therefore, as we decrease the amount of noise that is added, we do not expect the odds ratio distances for the logistic regressions to change significantly. We expect similar results for Feature Importance for the Full Set use case, irrespective of the algorithm. Recall that larger values for ϵ result in less noise.

When considering the reduced set use cases, RS1 and RS2, the number of records n is significantly larger than the cell coverage. Therefore, as we alter the ϵ parameter to decrease noise, we expect a decrease in the distance between the original and DP odds ratios. In addition we expect a higher percentage of Feature Importance experiments to match the results from the original dataset.

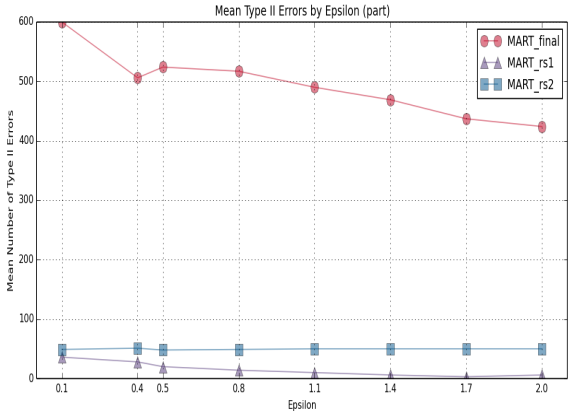
A. *Logistic*

Figures 2-4 illustrate the logistic results for the Kd-tree partitioning, cell-based, basic wavelet and adhoc wavelet algorithms respectively. These results include average distance measures and Type I and Type II errors.[2] For the Full, RS1 and RS2 datasets, there are 1300, 350 and 300 experimental runs respectively for each parameter settings. These numbers denote the maximum number of possible errors for per parameter setting(s) for each dataset.

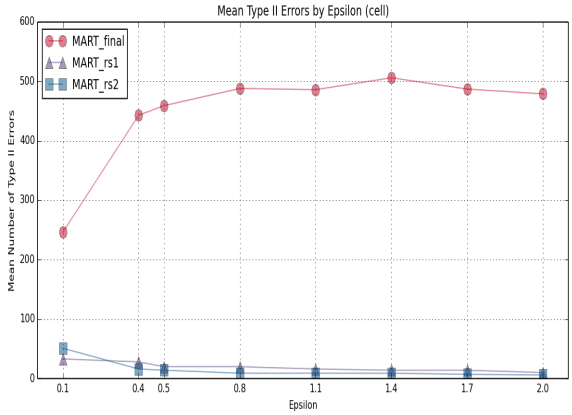
As shown in Figure 2, as expected for all algorithms, the odds ratio distance for the Full dataset has little or no change as the noise level decreases. For RS1 and RS2, the average distance decreases and levels off at $\epsilon >= .8$, with the Cell-based algorithm providing the best results with average distances less than .1.

Figure 3 shows that the Type I error rate gradually reduces for all datasets for the cell, basic and adhoc algorithms, with the error rate approaching zero for the reduced sets when $\epsilon >= .4$, and approximately 4% for the Full dataset for

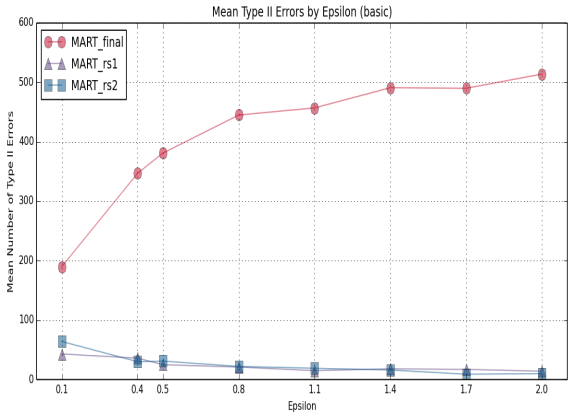
[2]Type I error is an incorrect rejection of the null hypothesis (i.e. false positive). Type II error is a failure to reject an untrue null hypothesis (i.e. false negative).



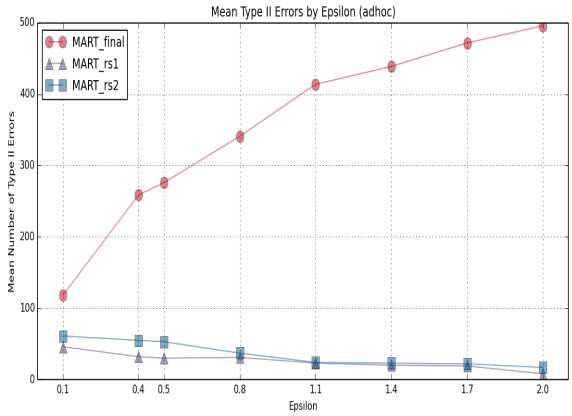
(a) Kd-tree



(b) Cell

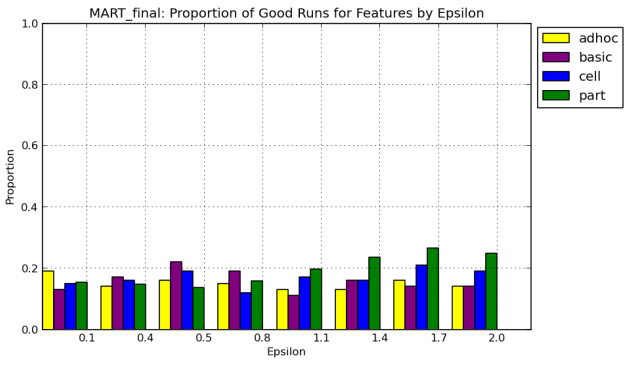


(c) Basic

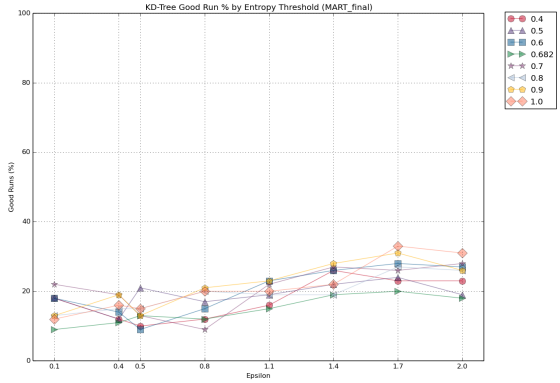


(d) Adhoc

Fig. 4: Logistic Results: Type II Errors



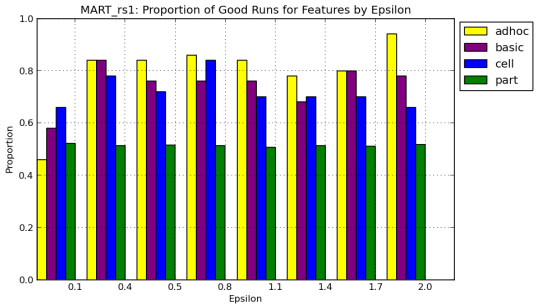
(a) All Algorithms



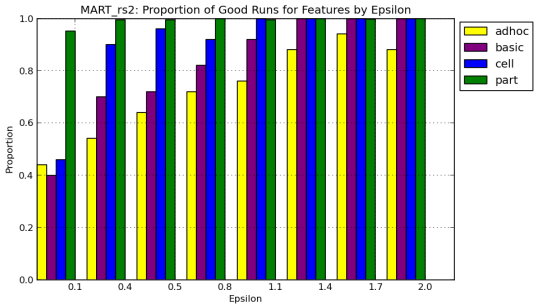
(b) Kd-tree

Fig. 5: Feature Importance results for MART_final for all algorithms.

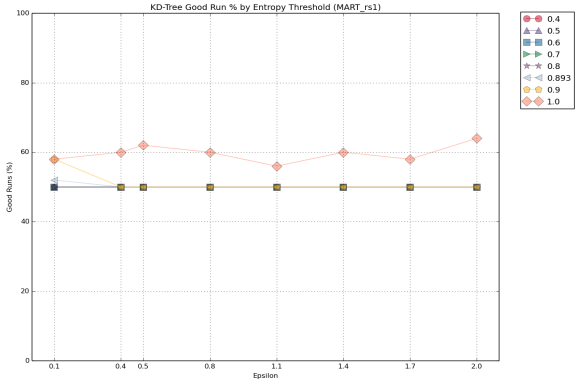
$\epsilon > 1.1$. The kd-tree algorithm generates DP data that gives mixed results. For example the rate of Type I errors decreases for only one dataset. These mixed results may be attributed to its noisy partitioning feature. Recall that kd-tree has a two-phase process for generating DP histograms. The first phase generates a synthetic dataset, and indices are taken from the noisy dataset and used to partition the data in the second phase. While using the synthetic dataset helps to conserve the privacy budget, it increases the inaccuracy of the partitioning and acts as another source of noise.



(a) Mart_rs1



(b) Mart_rs2



(c) Kd-tree MART_rs1

Fig. 6: Feature Importance results for MART_rs1 and MART_rs2 for all algorithms.

Figure 4 show similar Type II error rates for algorithms. For example, for the Full dataset, the error rate is never less than 30%, and for the non-kd-tree algorithms, the rate increases as the noise decreases. Recall that since the number of records in the original database is far less than the number of cells, we don't expect good performance from either algorithm. Also note that at lower values of ϵ more records are added to the noisy dataset. These synthethic records generate more false positives as shown in Figure 3 and fewer false negatives.

B. Feature Importance

We compare the relative importances of features (i.e., dimensions, variables) in the DP histograms to those in the original data. Using the extremely randomized trees (ERT) classifier from the sklearn library [27], we compute a rank order of each feature in the dataset from most to least important for prediction. Similar to the logistic and classifier utility methods, the data are first split by gender and use case (unsafe sex, unplanned pregnancy). Next, an ERT classifier generates and trains 250 randomized decision trees on the data. Features are ranked by their relative predictive power within the forest of decision trees, and this ranking is compared to the original results. If the original and DP rankings match exactly, we consider utility to have been preserved in the DP histogram (i.e., a good run).

For the full dataset, MART_final, we observed similar performance for all algorithms with k-d tree performing slightly better for values of $epsilon \geq 1.1$ (see Figure 5). Note that the results for the k-d tree partitioning algorithm are aggregated across all values for the entropy threshold parameter, and that the proportion of good runs reaches 33% for some $epsilon$ and entropy threshold combinations (see Figure 5).

As shown in Figure 6, the performance for the reduced sets is greatly improved with as much as 94% of the feature orders being preserved for MART_rs1 and 100% by all algorithms for MART_rs2. The performance of the k-d tree algorithm follows a similar trend for MART_rs1 in that the best performance occurs when the entropy threshold is 1 (see Figure 6). The k-d tree algorithm outperforms the other algorithms for MART_rs2 dataset for high noise values, $epsilon \leq .8$ (see Figure 6).

V. Conclusion and Future Work

This work presented an empirical evaluation of a differentially private behavioral science dataset. The goal of the analysis was to better understand whether DP data could be shared and within what context. Therefore, we sought to identify the data characteristics and the analytical results that are preserved even when DP noise is applied to a dataset. In addition, we wanted to identify any limitations of the evaluated algorithms and possible improvements. One challenge to assessing the utility of DP data is specifying what constitutes equivalent results. While our results show that in some cases, distance measures and error rates approach zero, this does not sufficiently articulate equivalence. When discussing these results with our team of behavioral scientists, we find that this measure of utility may not translate. Thus, more work is

needed to bridge the gap between a quantitative measure of accuracy and the way in which the behavioral scientist uses and evaluates the quality of their results.

Our results confirm that dimensionality is a major challenge for DP algorithms, especially when the number of records in the database is sufficiently less than the number of cells covered by the query. The k-d tree partitioning algorithm attempts to preserve such properties by using an entropy threshold to control the partitioning algorithm. One feature of k-d tree algorithm is that partitioning keys are derived from a perturbed dataset. While this feature allows us to conserve our privacy budget, it may be an additional source of noise and variance for the cells within a partition. This may explain why the DP histograms that were produced by the k-d tree algorithm for our reduced data sets had larger distance values and error rates than the other algorithms.

One very interesting finding is that all algorithms produced noisy histograms that had strong results for Feature Importance. While most of our social science collaborators tend to favor regression-based analysis, these results indicate that there is a privacy-preserving incentive for using other data mining and machine learning techniques for data analysis, especially when coupled with DP data. Such techniques may be used by reseachers during the process of collecting preliminary data and trying to identify features for their model.

VI. Acknowledgements

This work is funded by NSF grants CNS-1012081 and CNS-1117763.

References

[1] D. O. H. Code of Federal Regulations and H. SERVICES, "Title 45 public welfare, part 46, protection of human subjects," [45CFR46.101(b)(2)]
Revised January 15, 2009
Effective July 14, 2009.

[2] L. M. Beskow, L. Dame, and E. J. Costello, "Certificates of confidentiality and the compelled disclosure of research data," *Science (New York, NY)*, vol. 322, no. 5904, p. 1054, 2008.

[3] A. Solomon, R. Hill, E. Janssen, S. A. Sanders, and J. R. Heiman, "Uniqueness and how it impacts privacy in health-related social science datasets," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ACM, 2012, pp. 523–532.

[4] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*. Springer, 2008, pp. 1–19.

[5] ——, "Differential privacy," in *Automata, languages and programming*. Springer, 2006, pp. 1–12.

[6] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan, "On the complexity of differentially private data release: efficient algorithms and hardness results," in *Proceedings of the 41st annual ACM symposium on Theory of computing*. ACM, 2009, pp. 381–390.

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

[8] J. A. Higgins, A. E. Tanner, and E. Janssen, "Arousal loss related to safer sex and risk of pregnancy: Implications for women's and men's sexual health," *Perspectives on sexual and reproductive health*, vol. 41, no. 3, pp. 150–157, 2009.

[9] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, p. 14, 2010.

[10] C. C. I. Aggarwal and P. S. Yu, "A survey of randomization methods for privacy-preserving data mining," in *Privacy-Preserving Data Mining*, ser. Advances in Database Systems. Springer US, 2008, vol. 34, pp. 137–156.

[11] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '09. New York, NY, USA: ACM, 2009, pp. 19–30. [Online]. Available: http://doi.acm.org/10.1145/1559845.1559850

[12] F. K. Dankar and K. El Emam, "The application of differential privacy to health data," in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. ACM, 2012, pp. 158–166.

[13] Y. Xiao, L. Xiong, and C. Yuan, "Differentially private data release through multidimensional partitioning," in *Secure Data Management*. Springer, 2010, pp. 150–168.

[14] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim, "Private coresets," in *Proceedings of the 41st annual ACM symposium on Theory of computing*. ACM, 2009, pp. 361–370.

[15] M. Götz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke, "Privacy in search logs," *arXiv preprint arXiv:0904.0682*, 2009.

[16] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private record matching using differential privacy," in *Proceedings of the 13th International Conference on Extending Database Technology*. ACM, 2010, pp. 123–134.

[17] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas, "Releasing search queries and clicks privately," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 171–180.

[18] F. McSherry and I. Mironov, "Differentially private recommender systems: building privacy into the net," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 627–636.

[19] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 8, pp. 1200–1214, 2011.

[20] Y. Xiao, L. Xiong, L. Fan, S. Goryczka, and H. Li, "Dpcube: Diferentially private histogram release through multidimensional partitioning," in *Transactions on Data Privacy*, to appear.

[21] K. El Emam and F. Dankar, "Protecting privacy using k-anonimity," *Journal of American Medical Informatics Association*, vol. 15, no. 5, pp. 627–637, 2008.

[22] J. Brown, J. Holmes, K. Shah, K. Hall, L. R., and R. Platt, "Distributed health data networks," *Med Care*, vol. 48, no. 6 Suppl, pp. S45–S51, 2010.

[23] S. N. Murphy, V. Gainer, M. Mendis, S. Churchill, and K. I., "Strategies for maintaining patient privacy in i2b2," *Journal of American Medical Informatics Association*, vol. 13, no. Suppl, pp. 103–108, 2011.

[24] C. A. Kushida, R. Nichols, D.A. amd Jadrnicke, R. Miller, W. J. K., and K. Griffin, "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies," *Med Care*, vol. 13, no. Suppl, pp. S82–S101, 2012.

[25] C. E. Bredfeldt, A. L. Butani, R. Pardee, P. Hitz, S. Padmanabhan, and G. Saylor, "Managing personal health information in distributed research environments," *BMC Medical Informatics and Decision Making*, vol. 13, no. 116, 2013.

[26] L. Clark and D. Watson, "Constructing validity: Basic issues in objective scale development," *Psychological assessment*, vol. 7, no. 3, pp. 309–319, 1995.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.