An Empirical Analysis of the Utility of a Differentially Private Social Science Dataset

Raquel Hill, Michael Hansen ({ralhill,mihansen}@indiana.edu)
School of Informatics and Computing
Indiana University, Bloomington, IN 47405

Erick Janssen, Stephanie A. Senders, Julia R. Heiman
({ejanssen,sanders,jheiman}@indiana.edu)
Kinsey Institute for Research in Sex, Gender and Reproduction
Indiana University, Bloomington, IN 47405

Li Xiong ({lxiong@mathcs.emory.edu)
Department of Mathematics and Computer Science
Emory University, Atlanta, GA 30322 USA

June 14, 2013

Abstract

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. In addition, we assess the effectiveness of this mechanism to produce a privacy preserving dataset that conserves the use of the data for making statistical inferences.

1 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 outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, insurability, or reputation." [2] Disclosure of research data may be compelled by subpoena or by institutional interpretations of "open-records laws [1]. 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 evaluate contains medical, sexual, demographic, and psychological data. Participants can be vulnerable to privacy violations if, for example, an adversary already has some information about the person he or she is targeting [8].

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 technique 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. Such protection is worth exploring for the protection against subpoena and other exposure that is outside of the control of the data owner. While there are many theoretical results ('in vitro') for DP and utility outside of actual data cases [3, 4, 6, 5], 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.

2 Methodology

We take a use case based approach to evaluate the utility of our differntially private data. The use cases that we examine are derived from The Kinsey Institute's work in [7], which examines predictors of unprotected vaginal sex and unplanned pregnancy.

The specific use cases employ multi-variate logistic regression to evaluate the likelihood of a participant reporting an unplanned pregnancy (a binary outcome) and the likelihood of a participant reporting having had unsafe vaginal sex in the last 12 months. In addition to the full set of 9 variables (7 predictors, 2 response), we use two reduced predictive sets. These reduced sets each include two predictors and one response variable, which enables us to evaluate the DP algorithms on lower dimensional data.

To evaluate whether the DP mechanism preserves data utility, we performed performed multivariate logistic regression on both the original and differentially private data. The resulting odds ratios (OR) and statistical significances for each level of each predictor were compared against regressions with the differentially private datasets. For all sets of parameters, we judged regression results for each predictor level (e.g., an exclusive relationship status) as significantly different from the original in two ways: a one sample t-test for the OR values, and if the level changed from statistically significant to insignificant (or vice versa).

Our DP application is implemented in Python, and outputs a differentially private histogram using either a simple *cell-based* or *KD-tree partitioning* method [9]. The cell-based method adds Laplacian noise to each histogram bin independently using the perturbation (privacy) ϵ parameter. The KD-tree method partitions the dataset based on an entropy threshold (ET) and information gain (IG) parameter, and then applies Laplacian noise to the partitions.

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. The second is that once a party has the differentially private histogram, they are free to do with it what they please. This includes reconstructing data records from the histogram bins.

3 Results

Across all datasets and parameter settings, only 4.7% of the variable levels were similar to the originals. Most of the utility-preserving levels were from the reduced datasets (about 88%). There were twice as many cell-based levels as k-d tree overall, though the full dataset had more utility-preserving k-d tree than cell-based levels.

For the reduced sets, average effect size decreased as ϵ increased (cell-based), and as ϵ and entropy threshold increased (k-d tree). Effect size for the full dataset, however, was not strongly correlated with ϵ. Having an entropy threshold at or above the original histogram entropy was strong predictor of utility preservation across all datasets.

4 Acknowledgement

This work is funded by NSF grant CNS-101 2081.

References

[1] Laura M. Beskow, Lauren Dame, and E. Jane Costello. Certificates of confidentiality and the compelled disclosure of research data. *Science (New York, NY)*, 322(5904):1054, 2008.

[2] DEPARTMENT OF HEALTH Code of Federal Regulations and HUMAN SERVICES. Title 45 public welfare, part 46, protection of human subjects. [45CFR46.101(b)(2)]
Revised January 15, 2009
Effective July 14, 2009.

[3] Cynthia Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[4] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

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

[6] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil 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*, pages 381–390. ACM, 2009.

[7] Jenny A. Higgins, Amanda E. Tanner, and Erick 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*, 41(3):150–157, 2009.

[8] Ayla Solomon, Raquel Hill, Erick Janssen, Stephanie A. Sanders, and Julia 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*, pages 523–532. ACM, 2012.

[9] Yonghui Xiao, Li Xiong, and Chun Yuan. Differentially private data release through multidimensional partitioning. In *Secure Data Management*, pages 150–168. Springer, 2010.