Differentially Private ANOVA

Zachary Campbell, Adam Groce, Anna Ritz, and Andrew Bray

Reed College

Objectives

Our goal is to apply methods of differential privacy to a common query framework used on biological and social-science databases. We set out to acheive the following:

- To formalize ANOVA in a differentially-private manner
- Implement a differentially-private ANOVA algorithm
- To evaluate the performance of our algorithm with varying privacy parameters

Introduction

The objects of concern in this project are statistical databases. We will think of these specifically as databases populated with individuals'private, sensitive information (such as medical records, financial records, genome sequences, etc.). Differential privacy is a mathematical formulation within the field of database privacy that aims to allow queries on a statistical database, while offering strong privacy guarantees. That is, when one asks for quantitative information from a database (think mean, median, etc.), we want to gaurantee that this query will be executed in such a way that with high probability it does not leak any identifying information about individual records.

Motivation

It has been shown many times that simple anonymization of database records is not sufficient to protect the privacy of individuals who populate a databse. Perhaps most famously, in [1], researchers used anonymized and publicly available data released by Netflix, in addition to auxiliary databases, to identify specific users in the Netflix data. Similar attacks have shown effective on anonymized genetic data, leading to widespread data removal. The goal of differential privacy is to allow queries on these databases in such a way that privacy is gauranteed, no matter what auxiliary information an attacker might have. For another attack, see [2].

Differential Privacy

In the following, let x and y be databases in the universe $\mathcal{X}$. We define the notation $||x-y||_1$ to be a measure of how many records differ between databases x and y. We formally define differential privacy as follows: A randomized algorithm $\mathcal{M}$ with domain $\mathcal{X}$ is ϵ-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathcal{X}$ such that $||x-y||_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\epsilon)\Pr[\mathcal{M}(y) \in \mathcal{S}].$$

Idea

Data
Add noise
Privatized Data

Figure 1: In general, privacy is acheived by injecting random noise into the data. For more information see [3]

ANOVA

The setting is that we have a dataset that can be partitioned into k disjoint groups of individuals, based on some categorical predictor variable. The goal of an analysis of variance, or ANOVA, is to test if there is astatistical difference between the means of these groups. We now define the specifics of an ANOVA test. Let $\mathbf{D}$ be our database, and let $\{D_i \mid 1 \leq i \leq k\}$ be a partition of $\mathbf{D}$, where each D_i is what we think of as a "group." We denote the jth entry of group i by y_{ij}, the mean of group i by $\overline{y}_i$, and the mean of the entire database by $\overline{y}$. Also, let $x_i = |D_i|$. There are two important measures in the ANOVA framework that we will use:

$$\text{SSA} = \sum_{i=1}^{k} x_i(\overline{y}_i - \overline{y})^2,$$

$$\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{x_i}(y_{ij} - \overline{y}_i)^2.$$

Our algorithm

Our algorithm uses a technique called the Laplace mechanism, which is a powerful primitive in differential privacy. The idea is, given a query $f : \mathcal{D} \to \mathbb{R}^k$, we first analyze the sensitivity of f, denoted Δf. Intuitively, the sensitivity captures how much the output of our function can change for a change of input in a specified range. Using this sensitivity, we then draw noise from the Laplace distribution $\text{Lap}(\Delta f/\epsilon)$.

```
Algorithm 1 Differentially private ANOVA
  Input: Normalized database, $\epsilon$ value
  Output: $\epsilon$-differentially private $p$-value
  Compute $\text{SSA} + Y_1$ where $Y_1 \sim \text{Lap}(3/(\epsilon/2))$
  Compute $\text{SSE} + Y_2$ where $Y_2 \sim \text{Lap}(3/(\epsilon/2))$
  Compute MSE and MSA, and $F = \text{MSA}/\text{MSE}$
  Determine $p$-value based on this $F$-ratio
```

Figure 2: Our algorithm, which adds Laplacian noise based on the provable sensitivity of general sum-of-squares formulas, and returns a differentially-private p-value.

Results

Our results indicate that one can carry out an ANOVA test on reasonable database sizes with reasonable privacy guarantees.

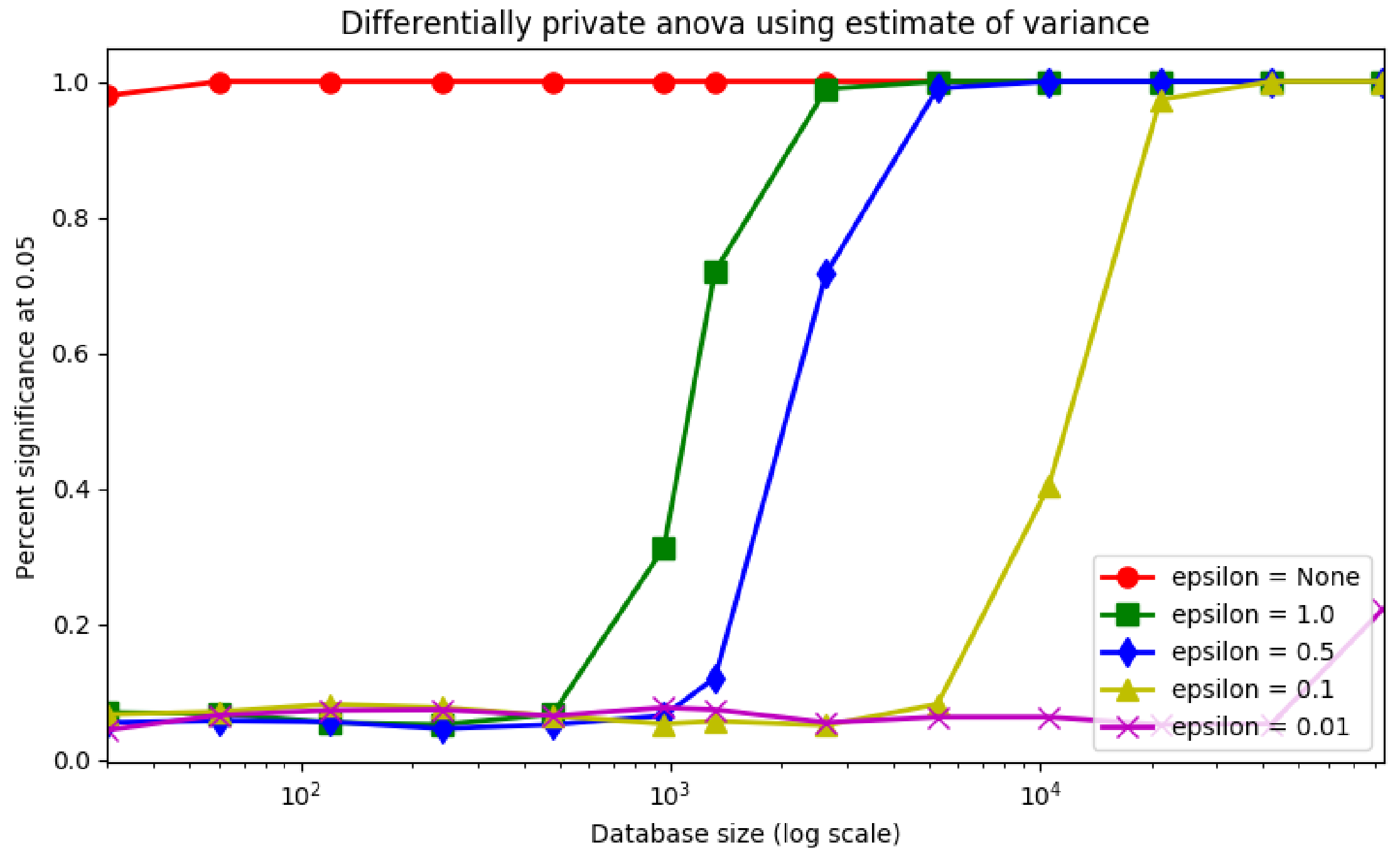

Figure 3: Differentially-private ANOVA with varying database sizes and epsilon values.

Conclusion

We show that for reasonable database sizes, one can execute differentially-private ANOVA tests with reasonable privacy gaurantees, without significant loss in accuracy. Of course, the model is not as accurate as when using non-private ANOVA, but that is the trade-off. As database size increases, the output of the differentially-private algorithm converges to the output of the non-private algorithm. Ultimately, we have shown that differentially-private ANOVA maintains strong utility.

Further Work

This project is only a small step into what is a very large area for exploration. There are certainly ways to get tighter bounds on the sensitivity of the sum-of-squares measurements, which would improve accuracy. We did make some attempt at implementing more complicated methods, but they require further work.

References

[1] Arvind Narayanan and Vitaly Shmatikov.
Robust de-anonymization of large sparse datasets.
In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[2] 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.

[3] Cynthia Dwork, Aaron Roth, et al.
The algorithmic foundations of differential privacy.
Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014.