Differentially Private Methods for Computing Confidence Intervals

Andrew Bray (Advisor)[†], Wenxin Du[†], Monica Moniot[‡], Canyon Foot[†], Adam Groce (Advisor)[‡]

[‡]Reed College Department of Computer Science, [†]Reed College Department of Mathematics

Differential Privacy

Imagine that a researcher publicly releases the maximum income of residents in a particular town. Anyone viewing the statistic could be confident that an acquaintance living in that town has an income no greater than the released value. In this way, even though the maximum is a summary statistic and we do not know who actually has that income, sensitive information has been revealed about everyone in the town. Differential Privacy provides a framework for such releases by giving a probabilistic guarantee that the inclusion of a particular person's data will not have a large effect on the statistic released. The strength of this guarantee is determined by the analyst's choice of the privacy parameter, . In practice, this typically means randomizing the output in some way, such as by adding (calibrated) noise to the true statistic. This process allows researchers to release informative statistics on sensitive databases without compromising the privacy of individuals.

Confidence Intervals

Confidence intervals are a popular technique for constructing a range in which some value of interest is likely to fall. When the each element of a sample $(\mathbf{X} = X_1, ..., X_n)$ is drawn independently from a normal distribution, confidence intervals can be constructed using only the sample mean $(\bar{X})$ and the sample standard deviation (s).

$$c(\mathbf{X}) = \left[\bar{X} \pm \frac{s}{\sqrt{n}} q_{n-1}\left(\frac{\alpha}{2}\right)\right].$$

When many intervals are constructed this way, $c(\mathbf{X})$ will contain the actual mean of the normal distribution, $(1-\alpha)\%$ of the time.

Our Objective

Produce differentially private confidence intervals for normal data that are small enough to be useful.

Making Queries Private

- A *query* is a function on the database such as mean or maximum.
- The Laplace mechanism takes the true value of a query Q and adds calibrated noise from a Laplace distribution:

$$\hat{Q}(D) = Q(D) + Lap(\frac{\Delta Q}{\epsilon})$$

where ΔQ is the sensitivity of the query.

- The exponential mechanism selects a value based on a utility function U that the analyst determines. The values are selected with probability proportional to:

$$\exp\left(\frac{\epsilon U(r, D)}{2\Delta U}\right)$$

Simulating Reference Distribution

The standard reference distribution quantiles are no longer appropriate since we are adding additional noise. Instead, we simulate the reference distributed to derive our critical values. With enough simulations, we can get as close as we want to the true reference distribution.

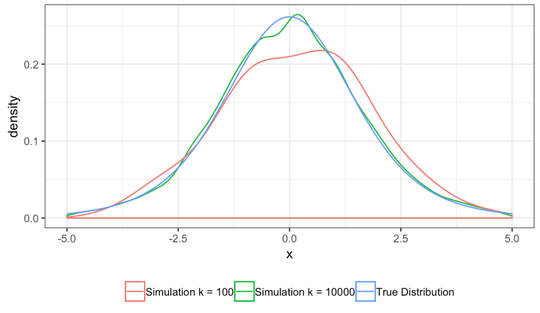

Figure: Example of simulated normal-Laplace distribution

'Noisy' Algorithms

To construct a differentially private confidence interval, we first construct private measure of the center and the spread. We found that the best measure of center was the sample mean, but that we could improve on the sample standard deviation. Instead, we calculate the mean absolute deviation:

$$\tilde{s} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|,$$

which has lower sensitivity.

Quantile Algorithms

We employed the exponential mechanism and constructed an algorithm that outputs differentially private estimate of sample quantiles. Because of the way the exponential mechanism selects the output, we are often able to get much more accurate estimates than we would with the Laplace mechanism.

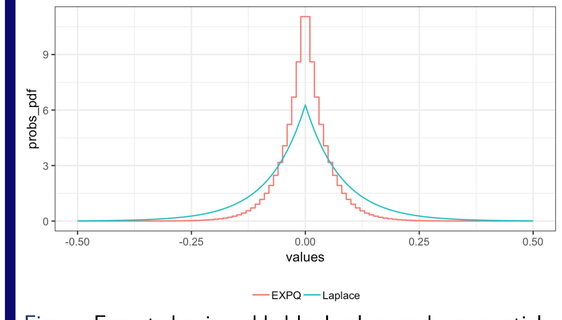

Figure: Expected noise added by Laplace and exponential mechanisms for estimating the mean

Our best quantile based algorithm estimates two qunatiles an equal distance away from the median. Their mean is the estimate of the center and the difference is used to estimate spread.

Results

All algorithms outperformed previously existed differentially private confidence interval algorithms for normal data. All algorithms produce private confidence intervals with correct coverage and reasonable width that allow the outputs to be practically useful. We compare to existing methods (see [1], [2], [3]) in the figure below:

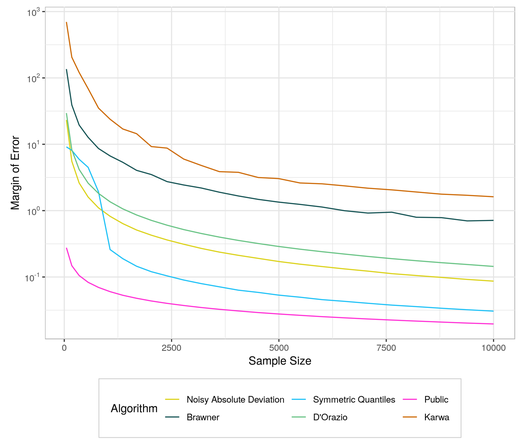

Figure: Graph of the performance of our methods against the public.

Acknowledgements

Funding was provided by the Reed College Science Research Fellowship for Faculty-Student Collaborative Research.

References

[1] T. Brawner and J. Honaker.
Bootstrap inference and differential privacy: Standard errors for free.
Unpublished Manuscript, 2018.

[2] V. D'Orazio, J. Honaker, and G. King.
Differential privacy for social science inference.
SSRN Electronic Journal, 01 2015.

[3] V. Karwa and S. P. Vadhan.
Finite sample differentially private confidence intervals.
CoRR, abs/1711.03908, 2017.