psilence: an R library for differential privacy

James Honaker*, Thomas Brawner*, Christian Covington*, Ira Globus-Harris

CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1565387. *Supported by SaTC.

Objectives

We have built a modular R library of routines for releasing differentially private statistical summaries and models. This can be used independently by researchers, or integrated into our larger deployed system for analysis of sensitive data. Differential privacy offers an attractive approach to enabling data sharing among social science researchers. Our desiderata are:

- **Accessibility by non-experts:** researchers in the social sciences should be able to use the library and larger system to share and explore data with no involvement from experts in data privacy, computer science, or statistics.
- **Generality:** the system should be applicable and effective on a wide variety of heterogeneous datasets hosted in a repository such as the Harvard Dataverse.
- **Workflow-compatibility:** the system should fit naturally in the workflow of its users (e.g. researchers in the social sciences), and be positioned to offer clear benefits (e.g. more access to sensitive data or less risk of an embarrassing privacy violation) rather than being an impediment.

Motivation

Researchers in all experimental and empirical fields are increasingly expected to widely share the data behind their published research, to enable other researchers to verify, replicate, and extend their work. Indeed, data-sharing is now often mandated by funding agencies. However, many of the datasets in the social and health sciences contain sensitive personal information about human subjects.

1. Numerous data sets, such as surveys, that have been "deidentified" via traditional means are increasingly being deposited in publicly accessible data repositories at risk of reidentification.
2. Numerous other data sets are not made available at all, or only with highly restrictive provisions.

Statistics

The initial set of dp-algorithms were chosen to give immediate utility for social science research:

- Univariate descriptive statistics, such as means, quantiles, histograms, and approximate cdfs.
- Basic statistical estimators and procedures, such as matching algorithms and difference-of-means tests for causal inference, and low-dimensional linear, logit, probit and poisson regression.
- Transformations for creating new features (variables) out of combinations of others.
- Simple handling of missing data
- Composition theorems for formally tracking the total privacy loss from a workload of statistics.

Library Architecture

There is a three tier hierarchy to the class structure that allows for easy reuse and verification:

1. **Statistics** define the quantities to release
   - Formal sensitivity of the release
   - Underlying true function
   - How to post-process
2. **Mechanisms** produce differentially private estimates of a statistic
   - Only location to touch the data
   - Evaluate all output
   - Generalize across statistics
3. **Randomization** functions provide random numbers for mechanisms
   - Do not rely on additional libraries beyond core R distribution
   - Use cryptographically secure randomness
   - Handle finite precision floating point algebra

Deployment Architecture

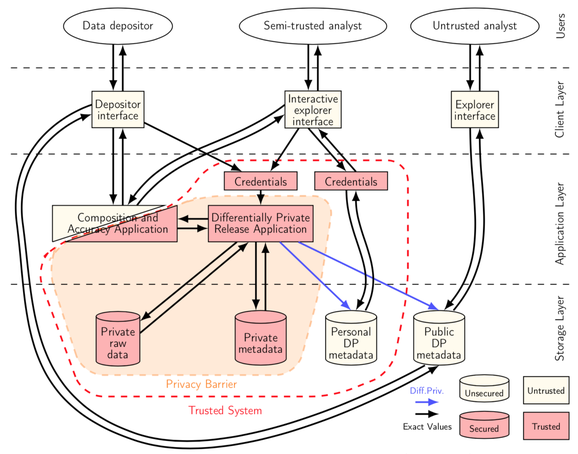
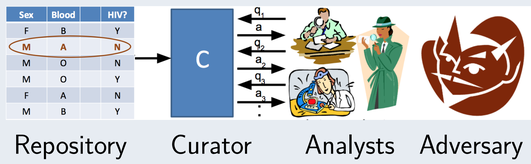


Figure 2: Architecture Diagram for Library (Center) Integrated into Deployed PSI System Including Interfaces

Differential Privacy and Queries

Differential Privacy guarantees that if a dataset changes by one observation, the distribution of answers released to queries would be indistinguishable.



Thus any one individual's information is masked and safe from reidentification by analysts or adversaries.

Differential Privacy

Differential Privacy, deriving from roots in cryptography, is a formal, mathematical conception of privacy preservation. It guarantees that any released statistical result does not reveal information about any one single individual (Dwork, McSherry, Nissim, Smith 2006). These algorithms inject a precisely calculated quantity of noise to any statistical query to mask the possible contribution of any one individual to the result. It is provable that no possible combination of queries or model results can tease out the information of any individual.

Differential privacy ensures that using an individual's data will not reveal essentially any personally identifiable information, or even whether the individual's information was used at all.

Budgeting Tool

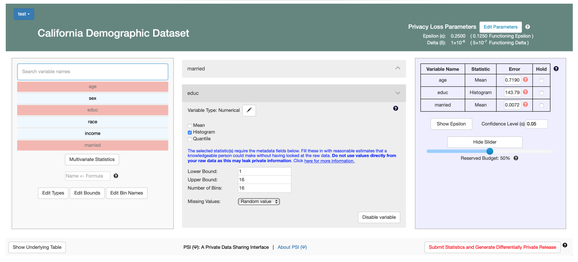


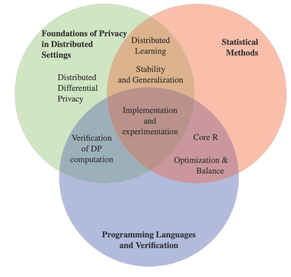
Figure 1: GUI Interface for Library Budgeting DP Releases

We have developed a privacy budgeting tool GUI interface to facilitate library usage that intuitively exposes the privacy-accuracy tradeoff to the user.

To ensure that we get the most utility out of the global privacy budget, we use the "approximate optimal composition theorem" which in fact was developed for the purpose of our privacy budget tool.

Acknowledgments

This work is part of the "TWC: Large: Collaborative: Computing Over Distributed Sensitive Data" project at Harvard, supported by NSF grant CNS-1565387.



Contact Information

- Code: `https://github.com/privacytoolsproject/PSI-Library`
- Prototype: `http://psiprivacy.org/about`
- Group: `http://privacytools.seas.harvard.edu`
- Email: `privacytools-info@seas.harvard.edu`
- Direct Email: `james@hona.kr`
- Web: `http://privacytools.seas.harvard.edu`
- Email: privacytools-info@seas.harvard.edu