PSIlence

The R library for the Privacy Tools Project

1 Introduction to PSIlence

PSIlence is a package for the R statistical language, that implements differentially private algorithms for releasing statistics while safeguarding the privacy of individuals in the raw data. Differential privacy provides a framework to reason about and bound the potential leakage of information about an individual in a data set. For example, a differentially private mean age of a group of people will not change significantly if you replace one person in the original group with another, thereby preventing inference about any individual in the dataset.

PSIlence provides the underlying algorithms for PSI(Ψ): a Private data Sharing Interface, a system created to enable researchers in the social sciences and other fields to share privacy-sensitive datasets they may not want to release to the public, and to explore datasets they may not otherwise have access to, all with the strong privacy protections of differential privacy (Project, Paper). It can also be used by itself in any application that requires differentially private mechanisms.

PSI and the PSIlence package were developed by the Privacy Tools Project, an interdisciplinary research group at Harvard, MIT and Boston University.

2 Brief Overview of Differential Privacy

Differential privacy is a rigorous mathematical framework for making statistical information about private datasets available. It allows a researcher to calculate statistics on a population, while bounding the probability of exposing any data about the individuals in the population, thus preserving their privacy. Differentially private estimates of various statistics are available in this package. For example, the `mean.release()` function releases a mean of a variable, while adding a precise amout of noise to guarantee ϵ differential privacy [1].

"ϵ differential privacy" depends on a privacy loss parameter ϵ, which is chosen by the user and which represents the degree of privacy preservation guaranteed to each observation (i.e. individual) in the data when releasing information. The value of ϵ is a description of how private the statistic release is, and is typically valued between 0 and 1 — as the value gets smaller, the level of privacy protection grows. However, greater privacy protection means more noise must be added to the true statistic to achieve the desired amount of privacy. Therefore, as ϵ grows smaller, the privacy protection becomes greater, but the accuracy of the statistical release becomes weaker. This is the key tradeoff in differential privacy: privacy protection vs. accuracy. The researcher may change the value of the privacy loss parameter ϵ as they wish, but they need to be aware of this trade-off. See section 6 for examples of how a user might vary the value of ϵ [2].

There is a second definition of differential privacy known as "$\epsilon - \delta$ differential privacy," which includes an additional privacy loss parameter δ, as well as ϵ. The big difference between ϵ and δ is that, mathematically, ϵ is a multiplicative bound on information leakage, whereas δ is an additive bound. When δ (the additive bound) is large enough, it allows for complete publication of a row of the database, while a large value of ϵ (the multiplicative bound) does not allow for complete publication of data. In other words, δ is a bound on the probability of a catastrophic privacy failure. The value of δ should be very small, on the order of 2^{-30}, and generally not greater than $1/n^2$ (where n is the size of the database). Similar to ϵ, as the value of δ grows smaller, the level of privacy increases, but the accuracy decreases. PSIlence uses a default delta value of 2^{-30}, and it is suggested that users do not set their own value of δ unless they are comfortable with the concepts of differential privacy. When the ϵ definition of differential privacy is used, δ is set to 0. Note that this means ϵ differential privacy offers a stronger privacy guarantee, because it indicates that the probability of a catastrophic privacy failure is 0 [2].

2.1 Motivating Examples for Using Differential Privacy

Example #1: Re-identification by linkage

In the late 1990s, the Massachusetts-based Group Insurance Company was responsible for purchasing health insurance for state employees and collected patient data of 135,000 state employees and their families. This data included the attributes in the left-most circle in Figure 1. The voter registration file for any city in the United State is public and can be procured at any time from the local city hall, and contains attributes in the right-most circle in Figure 1. The attributes of these two datasets overlap in zipcode, birthdate, and sex.

Researcher Latanya Sweeney got these two datasets, combined them on the overlapping attributes, and found that the medical records of the Governor of Massachusetts at the time, William Weld, were in the dataset. Knowing that Weld was from Cambridge, Massachusetts, Sweeney was able to determine that he was the only male with his birthdate from Cambridge in the dataset, and was able to determine that he was admitted to a hospital for cancer treatment [3].

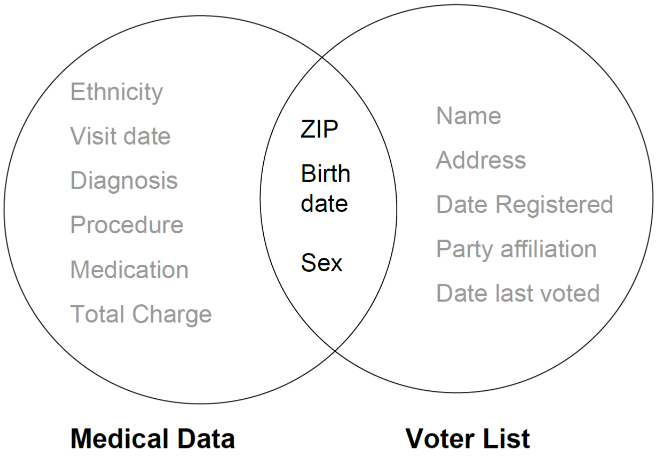

Figure 1: The overlapping attributes allowing for re-identification by matching

This example demonstrates re-identification by matching records on shared attributes. Sweeney's work shows that using differential privacy to alter the released information to map to many possible people, thereby making matches between records ambiguous, can prevent this kind of attack.

Example #2: Similarity Attack

At the 2017 Bar-Ilan Winter School on Applied Cryptography and Cybersecurity, Vitaly Shmatikov proposed an example of a similarity attack on data that had supposedly been "anonymized" [4].

Say we have a patient dataset where the sensitive attributes of zipcode and age have been anonymized by removing that last few digits, a common anonymization technique:

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Table 1: "Anonymized" patient data

In Table 1, no individual column is particularly sensitive or discloses a large amount of information, leading someone to belief the data is private enough. However, say we knew something about a particular individual in the dataset, Bob:

Bob	
Zip	Age
47678	27

Table 2: An example individual

With the information in Table 2, we can look at the patient dataset and see that there are 3 individuals living in a zipcode starting with 476** that are in their 20s, any of whom could be Bob. Again, this seems private.

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer

Table 3: Potential candidates that can be identified as Bob, even with anonymization

But by narrowing down the possibilities of which patient could be Bob to 3 possibilities, we still learn something about Bob:

1. Bob's salary is between $20K and $40K
2. Bob has a stomach-related disease

The "anonymization" in this example does not consider the semantics of sensitive values. If differential privacy were used here, such privacy violations would be prevented because the data would not be made public and open to attacks.

3 Getting Started

To get familiar with the statistics that are available with PSIlence and learn how to use them, you can explore the vignettes (R programming guides) in the `vignettes` folder of the library.

To start getting differentially private estimates of statistics on your data, first choose a statistic you would like to calculate. **The currently available statistics are: mean, variance, covariance, and histogram.** Once you choose a statistic to calculate, create an instance of that statistic with the `new` method, run that statistic's `release` method on your data, and print the `result`. Let's see an example:

Say we want to get a differentially private mean of the age of the people in the dataset `myData`:

```
exampleMean <- dpMean$new(var.type='numeric', variable='age', n=10000,
                          epsilon=0.1, rng=c(0, 100))
exampleMean$release(myData)
print(exampleMean$result)
```

See section 4 for more information on the parameters to specify in the `new` method.

4 Parameters to Specify for Differentially Private Estimation

When using PSIlence, the user must manually specify:

1. The variable type (of the values on which the statistic is going to be calculated)
2. The variable of interest in the dataframe for the statistic
3. The number of observations (n)
4. The desired ϵ value (see section 4 for how to select a value for ϵ)
5. The range of the data of interest (as a list of 2 numbers, the minimum value of the data and the maximum value), if the data is numeric

Optional parameter specification includes:

6. The desired δ value (default is 2^{-30})

δ is optional because it is only used when the stability mechanism is used (see section 7), and has a default value that is generally suitable in most cases. Users should not define their own value of δ unless they are comfortable with concepts of differential privacy.

The parameters above must be specified when the differentially private statistic (ex. `dpMean` or `dpHistogram`) is instantiated (with the `new` method), so the library can calculate the appropriate sensitivity and accuracy values to guarantee the desired level of privacy, and this must be done before the library accesses the data.

5 How to Interpret Results in PSIlence

Once you run the `release` method on the desired statistic, the differentially private estimate and information about the statistic can be found in `$result`. The result contains the following fields:

- The `$release` is the differentially private estimate of the statistic.
- The `$variable` is the variable on which the statistic was calculated.

- The `$accuracy` is the a bound on the difference between the differentially private estimate and the true value. If *alpha* is input as a parameter into the statistic, we say "with probability $1 - alpha$, the differentially private value of the statisitic is within *accuracy* of the true value."

- `$epsilon` is the differential privacy parameter of the statistic.

- The `$interval` is the $(1 - alpha)\%$ confidence interval on the true value of the statistic. If *alpha* is input as a parameter into the statistic, we say "with $(1 - alpha)\%$ confidence, the true value of the statistic in within this interval."

6 Use Cases for Differential Privacy

Example #1:
Say a researcher has done a study on the prevalence of HIV in a specific region of the United States. The researcher is worried about the privacy of the individuals in the population and does not want their data to be leaked. The researcher can use the PSIlence library to release the statistics of their study to guarantee the privacy of their study participants' data. The PSIlence library would allow them to release a mean age, for example, without risking revealing the age of any individual participant. Or the library would allow them to create a histogram of the number of people with and without HIV without revealing the status of any specific individual.

Example #2:
Say a researcher is studying a certain company, and they want to publish a histogram of the different types of employees, but they do not want to reveal the identity of employees with uncommon job titles. The researcher can use PSIlence to create a differentially private histogram of job titles, which will have noise added to each histogram bin to hide to true number of employees in each category, protecting individuals' privacy.

How to set ϵ:
Say a researcher conducts a study on an at-risk population and collects data. They want to publish the findings of their study, but they do not want to publish the raw statistics about the population (such as the mean age or a histogram of their nationalities), because it would reveal too much about the vulnerable population. The researcher would choose a small value for epsilon (on the order of 0.1 or smaller), to ensure a high degree of privacy for the calculations they release. This small value of epsilon and high degree of privacy means that the values of the statistics released will not be completely accurate for the population, so the researcher does not risk revealing too much about their sensitive population.

The following table can help you choose an ϵ value that corresponds with your required level of privacy:

Privacy Level	ϵ
Revealing the data would not cause harm, but you choose to keep it confidential	1
Revealing the data would cause harm to individuals	0.25
Revealing the data would likely cause serious harm to individuals	0.05

It is recommended that the value of ϵ never exceed 1 [5].

7 What is a mechanism?

Calculations are made "differentially private" by *mechanisms*. There exist many options for mechanisms, and the mechanism chosen for the calculation determines the way the calculation is made private.

The most commonly used mechanism is the **Laplace mechanism**, which adds noise to a statistic release by sampling from the Laplace distribution. For example, say you want to calculate a differentially private mean, using the `dpMean` statistic. After specifying the parameters from section 4, you would enter the data vector, and the `dpMean` statistic would first calculate the standard mean (by summing the values and dividing by the number of values). Then the `dpMean` statistic would use the Laplace mechanism to select a random amount of noise from the Laplace distribution to add to the mean value, to make the estimate of the mean differentially private and ready for release.

Another mechanism used in the PSIlence library is the **stability mechanism**. In the library, the stability mechanism is implemented to be used specifically for the histogram statistic, and should not be used for any other function except by users who are confident in their understanding of the mechanism. However, in general, the stability mechanism is one which takes advantage of "stable" functions, i.e. ones where the function output is constant in some neighborhood around the input database. When calculating a histogram, for example, the `dpHistogram` statistic will find the raw counts for each bin, and then the stability mechanism will remove any empty bins, add Laplace noise to the remaining bins (similar to the Laplace mechanism), and then will calculate an accuracy threshold and will remove any bins from the histogram with a count below that threshold. Removing empty bins and bins with a low count is what adds stability to the differentially private histogram, because keeping a bin with a small count may give away information about the presence of a bin in one database that may not exist in a neighboring database.

8 Resources for Differential Privacy

- The 2017 Bar-Ilan University Winter School on Cryptography
  - The 7th BIU Winter School focused on Differential Privacy and posted all of their lectures.
  - For a good overview of the basics of Differential Privacy, see Katrina Ligett's Introduction to Differential Privacy lecture and Basic Tools lecture.
- The Algorithmic Foundations of Differential Privacy by Cynthia Dwork and Aaron Roth
- The Complexity of Differential Privacy by Salil Vadhan

9 Resources for Developers

- Datacamp Introduction to R tutorial
- Advanced R by Hadley Wickham
- R Packages by Hadley Wickham
- PSIlence on GitHub

References

[1] Wood, Alexandra, et al. "Differential Privacy: A Primer for a Non-Technical Audience." *Vanderbilt Journal of Entertainment amp; Technology Law*, vol. 21, no. 1, ser. 209, 2018. 209, doi:10.2139/ssrn.3338027.

[2] Dwork, Cynthia, and Aaron Roth. *The Algorithmic Foundations of Differential Privacy.* Now Publishers Inc, 2014.

[3] Sweeney, Latanya. "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, May 2002.

[4] The BIU Research Center on Applied Cryptography and Cyber Security. "The 7th BIU Winter School: The Anonymization/ De-Identification Paradigm - Vitaly Shmatikov." YouTube, 8 Mar. 2017, www.youtube.com/watch?v=im1Rpvb0e8c.

[5] "Budgeting Tool." Privacy Tools Project, psiprivacy.org/.