Can Data Obfuscation techniques be beneficial for preserving Data Utility unlike Differentially Private Algorithms?

Debolina Ghatak [a], Kouichi Sakurai[b], and Bimal K Roy[c]

[a]Indian Institute of Technology Jammu
[b]Kyushu University
[c]Indian Statistical Institute Kolkata

Abstract

The data we put into social media platforms everyday encounter a lot of privacy issues. One of the very famous and important tools to check if any available information from a database is secure or not is to check if the released information is differentially private or not. Note that differential privacy only ensures if a given algorithm is secure or not and the utility of data is completely lost except the algorithm in concern. For example, if one publishes the output of a differentially private k-means algorithm, the output only helps in understanding the clusters of the data but not any distributional layout of a particular attribute or relationships among variables. This calls for data obfuscation, where a data-set is perturbed in such a way that the resulting output can further be utilized for various statistical inferences. Here, we discuss a few famous techniques to obfuscate both discrete and numerical data and discuss their privacy concerns.

Introduction

The data-sets that are collected by banks, agencies, institutes, hospitals etc. carry various sensitive information which if released publicly can be used by some intruder to access some private information about his target individual and harm him in some manner. That is why it is very necessary to look after privacy protection of an individual before releasing any information about a data-set.

Sometimes one may have the misconception that hiding identifying variables like name, id number etc. can make the data private. But that is not always true. To explain, we give an example here. Consider a data-set of bank data including attributes like *NAME*, *GENDER*, *AGE*, *PINCODE*, *PROFESSION*, *MONTHLY SALARY*, *ACCOUNT BALANCE*. Now, here name is an identifying variable. So if we remove it before releasing the data-set in public, then one may think at first that an intruder who is trying to get the private information of his target individual may fail to do that as he cannot identify the indivual in the data-set. But, that is not always true. The intruder may have some prior information about his target individual; for example he may have some idea about the age, profession, gender and residential place of his target individual and finding the data-set he may see that all these information matches only one row. He then directly gets to know the salary and account balance of the individual. Thus, hiding identifying variables is not sufficient to ensure privacy of individuals. However, some privacy preserving measures are proposed and studied in literature. In the last decade the concept of differential privacy has earned a lot of importance in ensuring privacy protection for a database.

Differential Privacy

One of the very important tools to check if any available information from a database is secure or not is to check if the released information is *differentially private* or not.

Theoretically, an algorithm $\mathcal{K}$ with range set $\mathcal{S}$ is called ϵ-differentially private if for two databases D_1 and D_2 differing in only one row satisfy,

$$e^{-\epsilon} \leq \frac{P[\mathcal{K}(D_1) \in S]}{P[\mathcal{K}(D_2) \in S]} \leq e^{\epsilon}$$

for any subset S of $\mathcal{S}$. There are lots of mechanisms discovered to achieve differential privacy, some of which include the **Laplace mechanism**[1] where we add a Laplace noise independent of the data to the output to ensure differential privacy of the output function, or the **Geometric mechanism**[3] where similarly we add Geometric noise instead of Laplace to the output. Geometric noise is usually added when $\mathcal{S}$ is discrete.

Data Obfuscation and Some Results

Data Obfuscation refers to the type of data masking where some useful information about the complete data-set remains even after hiding the individual sensitive information. So, the main objectives of Data obfuscation are

(i) minimize risk of disclosure resulting from providing access to the data

(ii) maximize the analytic usefulness of the data.

Some famous obfuscation techniques for which both privacy and utility are discussed include the following:

- *Addition / Multiplication of noise* for numerical data.
- *PRAM* models for categorical data.
- Generating *synthetic data* or data simulation from estimated distribution.

Table 1: True and obfuscated values for 10 data points selected from the 445 data points

No.	TRUE	Uniform	Laplace
"1"	814	960.562	733.931
"2"	750	695.214	829.526
"3"	764	656.395	591.158
"4"	574	704.041	599.055
"5"	614	670.67	586.944
"6"	669	595.926	670.136
"7"	616	553.873	533.097
"8"	674	748.607	677.74
9"	714	595.295	658.648
"10"	740	883.885	764.591

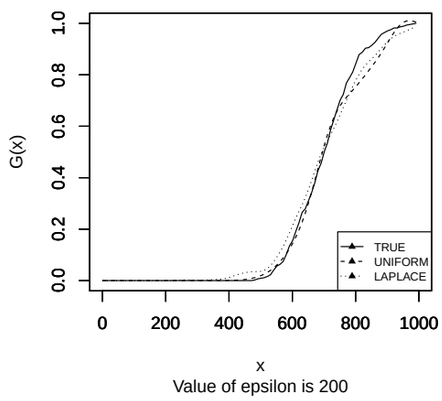

Figure 1: Showing true and estimated distribution curve

Addition or multiplication of noise to individual data values before releasing it is a famous technique to preserve privacy and along with that estimating density and distribution of original data from noisy data is well-studied in literature. Estimation of probabilities of falling into each category from post randomized data is also well-studied ad discussed by Nayak et al.[7]. Also synthetic data-sets, as described by Rubin(1993) [4] are data values simulated from the estimated copula of the original data-set and hence preserves the correlation structure among attributes which again implies preserving utility. Also the data values are simulated quantities and not true values which ensures privacy protection for data. Thus the obfuscation tools are not only useful in giving privacy guarantees but at the same time some utility of the data is preserved.

We demonstrate here the results of one of our analysis the details of which are given in Ghatak and Roy (2018)[2]. We collected a data-set of marks achieved by 445 students in the M.Stat 2nd yr program of Indian Statistical Institute Kolkata over 10yrs 2006-2015. Now since marks is a sensitive data, it cannot be released in its raw form in a data-set. So we add noise to the marks values before releasing the data in public such that even if an intruder gets successful in identifying the row of his target individual, the value of marks he will see wont be its true value but will be its obfuscated value.

Table 1 represents true and obfuscated values of 10 data points to show how the values are masked with Uniform and Laplace Errors. Then from the obfuscated values the true distribution can be estimated as shown in Figure 1. Thus one can see from the above results how one can ensure privacy and utility of a data-set at the same time using an obfuscation tool.

Security and Utility

The world of study of security in the last decade has focused very much on the idea of differential privacy along with another very important privacy measure named k-anonimity [5]. Most statistical functions including mean, median, k-means algorithm outputs, regression parameter estimation etc. has a differentially private versions of itself. Although these secured algorithms are very much talked about, they hardly take into consideration the aspect of utility of the data. Thus, the information released after applying necessary changes to make it differentially private mostly damages the utility of the data. Obfuscation tools on the other hand does not give a concrete measure for ensuring privacy which makes us often choose differential privacy over obfuscation. Wasserman and Zhou [6] discuss a few differentially private mechanisms that ensures some amount of utility of data. But these methods are applied for a limited type of data.

Conclusion and Future Prospect

Differential privacy and k-anonymity works effectively in preserving privacy but the algorithms often damage utility of data-sets. In future, we intend to find *obfuscation tools* that preserves the above privacy concerns along with the statistical utility of the data-set.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. page 265–284. In Proceedings of the Third Conference on Theory of Cryptography, 2006.

[2] D. Ghatak and B. Roy. Estimation of true quantiles from quantitative data obfuscated with additive noise. Journal of Official Statistics:In Press, 2018.

[3] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. page 1673–1693. In STOC, 2009.

[4] Donald B. Rubin. Discussion statistical disclosure limitation. pages 461–468. Journal of Official Statistics, 1993.

[5] L. Sweeney. k-anonymity: A model for protecting privacy. volume 10:5, pages 557—-570. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002.

[6] L. Wasserman and S. Zhou. A statistical framework for differential privacy. volume 105:489, pages 375–389. Journal of the American Statistical Association, 2010.

[7] T. K. Nayak C. Zhang and J. You. Measuring identification risk in microdata release and its control by post-randomization. Center for Disclosure Avoidance Research U.S. Census Bureau Washington DC 20233, 2016.

Can Data Obfuscation techniques be beneficial for preserving Data Utility unlike Differentially Private Algorithms?

Debolina Ghatak , Kouichi Sakurai , Bimal K Roy

Indian Institute of Technology Jammu , Kyushu University , Indian Statistical Institute Kolkata

Problem Description

Removal of identifying variables may not always ensure privacy protection in statistical data release.

Typical data-set View

I.D.	Gender	Age	Pin Code	Profession	M.Income (in 1000)	Account Balance
10101	"M"	43	700012	Worker	35	612342
10102	"M"	55	700043	Officer	90	5534567
10103	"F"	50	700003	Officer	70	3965478
10104	"F"	34	700082	Scholar	40	800432
10105	"F"	47	700055	Officer	120	1020045
10106	"M"	28	700100	Student	10	200654
10107	"F"	42	700049	Officer	60	1530128
10108	"F"	36	700082	Worker	25	983071
10109	"M"	34	700039	Worker	30	856313
10110	"F"	60	700053	Doctor	70	1709349
10111	"F"	29	700076	Scholar	25	481496
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Intruder Scheme

Prior information to intruder : My target individual stays near New Alipore, Kolkata, India which has pincode 700053 and is an old female doctor.
If he finds no other row in the data-set satisfying these conditions, he finds his target individual.

Privacy Protection

Famous techniques to guarantee privacy protection:

- *Differential Privacy:* For a data-base **D**, if an information **K(D)** is released, one way to check if it is secure is to check whether it can be shown to be ϵ-differentially private for some ϵ.
- **k**-*anonymity:* Looking at a released data-set **D**, observing any information associated with it, the intruder shall be confused among at least **k** individuals.

Utility

Although the above measures ensure privacy, often most of the utility of the data is lost.

Data Obfuscation

Data obfuscation refers to the type of data masking where some useful information about the complete data-set remains even after hiding the individual sensitive information.

Objectives:

- Minimize risk of disclosure resulting from providing access to the data.
- Maximize the analytic usefulness of the data.

Obfuscation applied on real-data (Additive Noise Model)

Table 1: True and masked values for 10 data points selected from the 445 data points of marks of students of an institute

No.	TRUE	Uniform	Laplace
"1"	814	960.562	733.931
"2"	750	695.214	829.526
"3"	764	656.395	591.158
"4"	574	704.041	599.055
"5"	614	670.67	586.944
"6"	669	595.926	670.136
"7"	616	553.873	533.097
"8"	674	748.607	677.74
9"	714	595.295	658.648
"10"	740	883.885	764.591

Note that: Noise was generated independent of data and added to data values.

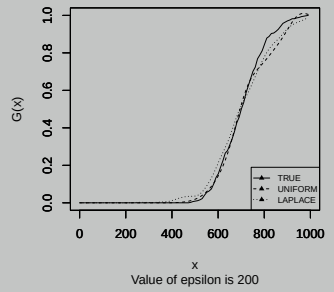

Figure 1: Showing true and estimated distribution curve of data

Discussion

In future, we aim to find methods of obfuscation that satisfy the privacy concerns of differential privacy, k-anonimity but presrves utility of data as well.