Integral Privacy Compliant Statistics Computation

Navoda Senavirathne[1,2(✉)] and Vicenç Torra[1,2]

[1] School of Informatics, University of Skövde, Skövde, Sweden
navoda.senavirathne@his.se, vtorra@ieee.org
[2] Hamilton Institute, Maynooth University, Maynooth, Ireland

Abstract. Data analysis is expected to provide accurate descriptions of the data. However, this is in opposition to privacy requirements when working with sensitive data. In this case, there is a need to ensure that no disclosure of sensitive information takes place by releasing the data analysis results. Therefore, privacy-preserving data analysis has become significant. Enforcing strict privacy guarantees can significantly distort data or the results of the data analysis, thus limiting their analytical utility (i.e., differential privacy). In an attempt to address this issue, in this paper we discuss how "integral privacy"; a re-sampling based privacy model; can be used to compute descriptive statistics of a given dataset with high utility. In integral privacy, privacy is achieved through the notion of stability, which leads to release of the least susceptible data analysis result towards the changes in the input dataset. Here, stability is explained by the relative frequency of different generators (re-samples of data) that lead to the same data analysis results. In this work, we compare the results of integrally private statistics with respect to different theoretical data distributions and real world data with differing parameters. Moreover, the results are compared with statistics obtained through differential privacy. Finally, through empirical analysis, it is shown that the integral privacy based approach has high utility and robustness compared to differential privacy. Due to the computational complexity of the method we propose that integral privacy to be more suitable towards small datasets where differential privacy performs poorly. However, adopting an efficient re-sampling mechanism can further improve the computational efficiency in terms of integral privacy.

Keywords: Privacy-preserving statistics · Privacy-preseving data analysis · Descriptive statistics

1 Introduction

Privacy preserving data analysis has become a strong requirement with the use of sensitive data in data analysis. The privacy requirement remains such that no analysis done on sensitive data should lead to any disclosure of sensitive information. Several definitions of what privacy means have been introduced

C. Pérez-Solà et al. (Eds.): DPM 2019/CBT 2019, LNCS 11737, pp. 22–38, 2019.
https://doi.org/10.1007/978-3-030-31500-9_2

in the literature. They are computational definitions that permit us to build algorithms to provide solutions satisfying these privacy guarantees. Examples of such definitions include k-anonymity and differential privacy.

In [1], the concept of *integral privacy* (IP) was introduced with respect to machine and statistical learning models, which focuses on how the models are affected as the underlying data changes. In a real world scenario, the collected data we use for analysis may update over time. And this brings up the requirement to regenerate the data analysis results. An adversary who has access to previous and new (regenerated) data analysis results should not be able to infer any sensitive information despite of having access to auxiliary information. The privacy model suggests achieving privacy through releasing stable/robust results that are less likely to change due to small perturbation done to training data. The *stability* of the results are defined in terms of how many different combinations of data (generators) can be used to construct the same result.

In this paper, we study how to apply IP in order to compute descriptive statistics. In particular, we will consider mean, median, IQR, standard deviation, variance, count, sum, min and max. We have proposed a method based on data discretization and re-sampling to compute integrally private statistics. Also, we compare the differentially private statistics with the ones obtained with our approach for their robustness (variability) and accuracy.

The structure of the paper is as follows. In Sect. 2 we review the related work followed by Sect. 3 which explains the preliminary concepts. Section 4 describes the methodology. Evaluation and results are presented in Sect. 5. Section 6 contains the discussion and the paper finishes with a section on conclusions and lines for future work.

2 Related Work

Over the years, many different privacy models have been introduced to attain privacy preserving data analysis. Among them differential privacy [2] stands out due to its mathematical rigour. Differential privacy is considered in the context of statistics, mainly with respect to statistical database systems [3]. In an interactive setting, the data curators want to ensure answering queries submitted by the users does not lead to any form of disclosure. Dwork et al. discuss differentially private statistical estimators and how they can be applied to obtain privacy preserving statistics [4]. Also, in another work Dwork et al. explore the relationship between robust statistics and differential privacy [5]. Even though differential privacy provides a very strong, theoretically sound privacy guarantee, there are some practical limitations [6]. Intuitively, differential privacy states that any possible result of an analysis should be almost equally likely regardless of the presence or absence of specific data records. This goal is achieved through controlled random noise addition. This diminishes the utility of the final outputs greatly. Also, differential privacy is being criticized for its complexity in implementing differentially private mechanisms, the difficulty of adopting such mechanisms into other algorithms, deciding on privacy parameter ϵ, difficulty in

estimating the sensitivity of an arbitrary function etc. Therefore, a solution is required that is compliant with the "indistinguishability" principle while capable of providing results with high utility. With that goal in mind in this work, we implement integral privacy [1] in the context of statistics in order to compute descriptive statistics. In some previous works, it is shown that the concept of IP can also be applied in the context of machine learning model section where stable models can be selected to achieve privacy [7,8].

3 Preliminaries

Differential privacy (DP) is the most commonly used privacy model in statistical and machine learning domains. The privacy guarantee of DP is such that the existence of any individual record can not be determined by examining the results of a function that was executed on two neighbouring datasets, which are differing from each other based on a single record. In other words, the result of a function does not change too much as a response to an addition or deletion of one record. This is achieved by introducing some uncertainty to the final result. Formally DP is defined as below.

Definition 1. *A randomized algorithm A is said to be ϵ-differentially private, if for all neighbouring data sets X and X', and for all events $E \subseteq Range(A)$,*

$$Pr[A(X) \in E] \leq e^{\epsilon} Pr[A(X') \in E]$$

Laplacian noise addition is one of the most commonly used mechanisms to implement DP in the case of numerical data. The noise is calibrated based on the "sensitivity" or the maximum variation a function can take [2].

Definition 2. *Let A be a real valued function; then, the global sensitivity of A is defined by*

$$\Delta A = \max_{d(X,X')=1} ||A(X) - A(X')||_1.$$

At the end, the noisy result is computed as $A(X) + Lap(\frac{\Delta A}{\epsilon})$ for $\epsilon > 0$. Here, ϵ is the privacy parameter.

The concept of integral privacy (IP) was first introduced in [1] with respect to machine and statistical learning models, that focuses on how the privacy of the models are affected by the changes done to the underlying dataset. It states that by observing the regenerated results (due to the dataset modification), an adversary with some auxiliary information can infer the modifications done to the input data (data addition and deletion) as it is being reflected by the final results. In [7], an adversarial model is explained with respect to machine learning model selection known as "model comparison attacks", that can be avoided by adhering to IP conditions. The idea is that, when the adversary has information on the previous ML model and the new ML model (trained after the changes applied to the input data) along with full/partial access to

the training data used to generate the previous ML model; they can be used together in order to determine which input data have specifically resulted in the given ML model, or to derive an idea on how input data have been changed. Therefore, generating robust/stable results that are less likely to be affected by the input data modification is significant for privacy. The goal of IP is to protect from intruders learning about the database and about the set of modifications applied. DP achieves the above mentioned privacy requirement through random noise addition whereas, IP achieves it by releasing the least susceptible result for input modification.

IP is based on the concept of "generators of an output". Let P be the population (or an estimation of this population) in a given domain $\mathcal{D}$. Let A be an algorithm or a function that given a data set $S \subseteq P$ computes an output $A(S)$ that belongs to another domain $\mathcal{G}$. Then for any $G \in \mathcal{G}$ and some previous knowledge S on the generators, the set of possible generators of G is the set defined by $Gen(G, S) = \{S' | S \subseteq S' \subseteq P, A(S') = G\}$.

The following definition formalizes integral privacy. It is to protect inferences by an intruder who (i) has some partial knowledge S on the original database and on S' the database obtained after modification, (ii) has knowledge on the algorithm/function A applied to both databases, and (iii) on the output of this algorithm when applied to the original database (say, G) and the one obtained when applied to the modified database (say G').

Definition 3. *Let $G, G' \in \mathcal{G}$, let A be the algorithm to compute the function, let $S, S' \subseteq P$ be some background knowledge on the data sets used to compute G and G', and let*

$$\mathbb{M} = \cup_{g \in Gen(G,S), g' \in Gen(G',S')} \{g' \ominus g\}.$$

Then integral privacy is satisfied when the set $\mathbb{M}$ is large *and*

$$\cap_{m \in \mathbb{M}} m = \emptyset.$$

The null intersection is to avoid that all generators share record/s. This would imply that there is a minimum set of modifications that can be inferred from G and G'.

4 Methodology

Inferential analysis of aggregated statistics can be used to obtain a variety of sensitive information about the underlying dataset. Compared to DP, IP looks at privacy preservation from a slightly different angle. As explained in the previous section, the main goal here is to select a statistical or a machine learning model that can be represented by multiple generators. In other words, these are different combinations of input data samples with no shared records among them. In this case, it is infeasible to determine exactly what input data has resulted the specific output even though the adversary has access to crucial auxiliary information.

The implementation of IP achieves this through re-sampling and discretization of outputs. When deriving the answer for a given statistical query (e.g., mean), the proposed integral privacy based method selects the most recurrent result which can be generated by unique input data samples with no intersection among them.

In order to implement the above, it is required to construct the distribution of the outputs of a given function $A()$ considering all possible combinations of the input dataset. As generating all possible combinations are computationally expensive, a re-sampling based approximation method is used to build a sampling distribution of the outputs. A t number of re-samples (S_i) are drawn from the original dataset P and then a specific function $A()$ is computed for each of the re-sample as $m_i = A(S_i)$. In the end, the distribution of function outputs (m_i) is built based on the relative frequency of occurrence of each output. Here, t is a user defined parameter.

A user defined parameter k, is used to define the level of recurrence (frequency). In this context, k works as a frequency threshold. All the responses with the frequency of occurrence greater than k are selected as a candidate response. Then the responses (m_i) with no intersection among its generators are filtered out, and the one with the highest frequency of occurrence or the least error can be selected as the final answer. Parameter k can take any value ≥ 2.

However, it becomes challenging when IP needs to be applied on statistical databases, due to the fact that the range of a function A could be such that $A(S_i) \in \mathbb{R}$. This does not guarantee recurrence in output values as most of the outputs can be unique. Our solution to this problem is applying rounding based data discretization on input data as well as to the final result before determining the relative frequencies. By using data discretization, a continuous data set can be mapped to a finite, discrete set. We discuss our solutions for input and output discretization below.

1. Input discretization - We apply microaggregation (MA) a masking technique where the input data are divided into micro-clusters, and then they are replaced by the cluster representatives. Parameter y defines the number of minimum data points required to form a micro-cluster. As the cluster centroid is used to replace the original values that fall into the particular cluster, the uniqueness of data records is concealed, thus preserving the privacy of the released data. The basic idea is to generate homogeneous clusters over the original data in a way the distance between clusters are maximized. As the value of y increases, more distortion is applied to data and vice versa. Application of microaggregation on a numerical dataset transforms the data into a discrete space.
2. Output discretization - The output values of a given function are rounded off in order to limit the number of unique responses. This improves the frequency of occurrence of a given response value with respect to different data re-samples. In this case, the final answer is rounded-off to a decimal number with fewer digits, r (E.g., 2 decimal points).

As explained above, in order to implement IP, it is required to obtain re-samples of data from the original dataset. In this work bootstrapping is used as

Data: P: Data set;
n: number of re-samples;
A: function;
k: minimum frequency threshold for generators;
Result: Integrally private function results

P^D := MA(P,y) ▷ Input discretization using microaggregation (MA)
for $i = 1$ **to** n **do**
  $S_i := bootstrapSample(P^D)$
  $m_i := A(S_i)$ ▷ Compute function A() for given S_i
  $md_i := round(m_i, r)$ ▷ Output discretization with r number of decimals
  E_M := add (md_i, S_i) ▷ Add the results to the re-sample function space
end
for *each unique* $md_i \in E_M$ **do**
  $Frequency_i := frequency(md_i)$ ▷ Derive the distribution of function results from E_M
  $DistributionOfResults :=$ add $(E_M, Frequency_i, generatorList = append(concat(S_i)))$ ▷ generatorList is upadted by concatenating all items in re-samples S_i
end
for *each* $md_i \in DistributionOfResults$ **do**
  if $frequency(md_i) \geq k \wedge intersection(generatorList_i) == \emptyset$ **then**
    $CandidateResultList := add\ (md_i)$
  end
end
return $CandidateResultList$;

Algorithm 1. Integrally private statistics computation.

the re-sampling technique [9]. It consists of drawing s observations with replacement from the original data. Here s denotes the size of the original data. Sampling with replacement causes the replication of some observations while the exclusion of the others. On average, in a bootstrap sample, there are $0.632 * s$ unique observations. In this case, bootstrapping is selected as it draws samples that match the size of the original dataset (due to sampling with replacement). In this way, when IP is used to compute counting queries, it leads to the correct answer. Initially, we also experimented with sub-sampling technique that generates samples without replacement. Results observed in both cases are very similar (bootstrap results are marginally better than sub-sampling), except for the counting queries. Hence, we opted bootstrapping as the re-sampling technique.

To build each sample $S_i{}^*$, s instances are selected with replacement from the original data set (P). The samples are same in size as $s = |P|$. This process is repeated n times to generate the bootstrap distribution.

Algorithm 1 summarizes our method. The algorithm returns an empty list when there are no integrally private results for the function A, given the dataset and other user defined parameters. In that case, the discretization parameters (y in microaggregation, r rounding), number of re-samples (n) or the frequency

threshold (k) can be adjusted to generate IP results. However, this can result in high computational cost (when increasing n) or high distortion of the results (when increasing y, r).

The above mentioned method is applied to compute IP solutions for some descriptive statistics. We focus our work in the following ones; mean, median, IQR, standard deviation, variance, count, sum, min and max. The experiments obtained using Algorithm 1 are described in the next section.

In IP, privacy is defined by the notion of stability, which leads to release of the least susceptible data analysis results towards the changes in the input data. Here, stability is explained by the relative frequency of different generators (re-samples of data) that lead to the same data analysis results. Highly stable results are recurring with respect to different data re-samples obtained from the original dataset. Integrally private output $f(x)$ can be considered as stable on the dataset x, if the same result appears more than a given frequency threshold k with respect to unique data re-samples drawn from x which does not share any common data instances among them.

5 Results and Evaluation

This section is focused on evaluating the effectiveness of our approach when computing a set of descriptive statistics. Here, we describe the experimental setting (data in Sect. 5.1, evaluation in Sect. 5.3), analysis of the results and comparison with differential privacy (Sect. 5.4) respectively.

5.1 Data

Six synthetic datasets (1-dimensional) and two real world datasets are used to evaluate the results. The parameters used for creating the synthetic data distributions are described along with the dataset dimension in Table 1. Abalone and breast cancer datasets are downloaded from the UCI data repository.

Table 1. Dataset descriptions.

Dataset	Instances × Columns	Description
Norm I	1000×1	Normally distributed with $\mu = 1, \sigma = 1$
Norm II	1000×1	Normally distributed with $\mu = 1, \sigma = 5$
Exp I	1000×1	Exponentially distributed with $\lambda = 1$
Exp II	1000×1	Exponentially distributed with $\lambda = 0.2$
Unif I	1000×1	Uniformly distributed in range ($\min = 0, \max = 100$)
Unif II	1000×1	Uniformly distributed in range ($\min = 0, \max = 1000$)
Abalone Dataset	4177×8	UCI data repository
Breast Cancer	683×7	UCI data repository

5.2 Experimental Setup

Algorithm 1 is implemented for calculating the descriptive statistics compliant with IP. Nine basic descriptive statistics have been considered. They are mean, median, standard deviation, min, max, interquartile range (IQR), sum and variance. Algorithm 1 is used to calculate the descriptive statistics compliant with IP. In this case, the number of re-samples (n) extracted from the original dataset is set to 1000 for synthetic datasets, 5000 for the Abalone dataset and 700 for the breast cancer dataset which is roughly closer to the number of instances (i). For both IP and DP, before reporting the final statistics, 10 iterations are carried out per each statistic and then the mean values are reported with their standard deviation and mean absolute relative error (ARE) for evaluation purpose.

For calculating DP statistics Laplacian mechanism is used. As explained in the preliminaries section, to calibrate the noise for DP, we need to derive the sensitivity of the functions. To compute the maximum variation a function can take (global sensitivity), it is essential to know the lower and upper bounds for the domain of a given dataset.

The global sensitivity of a function can be very large, causing high distortion to the computed results. Because of that local sensitivity derived from the dataset is used for some functions (i.e., median, max, min, IQR). For normally and exponentially distributed data the minimum and maximum bounds for the datasets lies between $(-20, 20)$ whereas, for the unif I the values range from $(0, 100)$ and for Unif II from $(0, 1000)$. Given that the Abalone and Breast cancer datasets are biological datasets, no strict domain bounds are introduced to pre-process the data as it presents very limited chances of being boundless. Therefore, when computing function's sensitivity min and max values of the respective datasets are used.

When computing differential privacy statistics mechanisms introduced in the literature are used to estimate the global/local sensitivity of the statistical functions as mentioned below. For median, max and min functions techniques introduced in [10] are used with local sensitivity. For mean calculation *noisy average clamping down* algorithm introduced in [11] is used whereas, for sum queries maximum value in the domain (i.e., in this case the max value of the specific dataset) is used. For IQR calculation *Scale* algorithm proposed in [12] is used. For variance and standard deviation the min and max values computed using the above techniques are used to estimate the function's sensitivity. For counting queries the global sensitivity is set to 1.

For IP, 18 different dataset instances are evaluated based on the data distribution type and discretization parameters. Two discretization phases are used as *output discretization (Out Dis:)*, and *both input and output discretization (in/out Dis:)* combined with input discretization levels, low (L) and high (H). In low discretization level parameter y for microaggregation is set to 2 whereas for high discretization parameter y is set to 20. In all the cases, rounding parameter is set to 2 at the output discretization phase. For DP different data distributions are used with differing ϵ values which indicates the amount of privacy.

5.3 Evaluation Criteria

For the evaluation purposes three measures are used as mentioned below. Here, $A()$ indicates the statistic value to be computed (e.g., mean(), median()), P indicates the original dataset, S_i indicates the re-samples, IP{} indicates integrally private value selection, true value indicates the real statistic value computed on the original dataset and private value indicate the mean IP or DP compliant statistic value. When computing absolute relative error (ARE) the distance between the true value and the private value is divided by maximum among 1 or true value to avoid division by zero. A lower ARE indicate less distorted IP/DP results.

$$IP\ Mean\ Statistic\ Value\ With\ SD = \frac{\sum_{j=1}^{10} IP\{A(S_1)\dots A(S_i)\}}{10} \pm SD \quad (1)$$

$$DP\ MeanStatistic\ Value\ With\ SD = \frac{\sum_{j=1}^{10} \{A(P) + Lap(\frac{\Delta A}{\epsilon})\}}{10} \pm SD \quad (2)$$

$$Absolute\ Relative\ Error(ARE) = \frac{|True\ Value - Private\ Value|}{max\{1, True\ Value\}} \quad (3)$$

5.4 Results and Discussion

Variability/Robustness of the Results. Tables 2 and 3 respectively show IP statistics and DP statistics computed on the synthetic dataset. In this case, we wanted to check the variation of the final results among different iterations. The same is illustrated by Fig. 1. By observing the results, few interesting facts can be noted. Relative to DP in IP the variability of the results is low in many instances. This is indicated by the $\pm SD$ values. Further, this indicates that as opposed to adding Laplacian noise to achieve DP, re-sampling based IP provides more stable/robust answers with less variability despite different iterations. However, DP performs better than IP when calculating sum() and mean(). This behaviour is expected as re-sampling does not provide a correct approximation of the total values. Also, in the case of mean() computation DP results have low variability compared to IP.

In the case of uniformly distributed data IP reports many "NA" values. This indicates that at least there has been one iteration where an IP result was not available with respect to the provided set of parameters. Uniformly distributed data contains integers as opposed to the other two distributions. This might require us to increase the discretization level or the number of re-samples and recheck for IP results. For example, in output discretization we have limited the number of decimal points to 2. In the case of integer data, rounding to the nearest integer or a multiple of some value can be used to avoid "no response (NA)" issue. Moreover, it is noted that robustness of the answers and the discretization level in IP or ϵ in DP has no prominent relationship.

Accuracy of the Results. Variability of the results does not indicate the quality of the computed statistics alone. To measure how accurate the results are absolute relative error (ARE) can be used. Tables 4 and 5 contain a detailed picture of the ARE rate for different data instances. Generally speaking, IP reports a lower error rate compared to DP. However, as discussed earlier with respect to uniformly distributed data sum() and variance() functions fails to find IP compliant solutions within the defined set of parameters. Table 6 shows the summation of ARE rate after excluding the sum() and the variance(). As it can be seen clearly, DP reports a very high ARE rate compared to IP in all the cases. Thus, IP can be seen as the preferable solution in both robustness and accuracy wise.

Different Discretization Methods for IP. When computing IP compliant statistics, three discretization methods are used as, (a) output discretization, (b) output discretization with minimum input discretization and (c) output discretization with high input discretization. Based on the results from Table 4, it can be seen that the output discretization is enough to produce IP results with minimum ARE rate. Sum of ARE is reported as 16.47, 22.28 and 38.74 under the discretization scenario (a), (b) and (c) respectively. However, by using both input and output discretization the frequency of occurrence (k) in a given result can be increased. In other words, this provides a high degree of privacy as having a higher number of generators increase the uncertainty of exactly figuring out the set of generators of a given result. When IP is used with integer data, output discretization required to be more carefully selected to avoid "no response (NA)" scenarios. Usually, increasing the rounding base or the number of re-samples can be seen as an answer to this.

Comparison of IP with DP on Real World Datasets. Here, we carried out the same experiment on the Abalone dataset where IP and DP are used to compute the descriptive statistics. As depicted by Fig. 2, IP reports a much less ARE rate compared to DP. Further, for statistics like count, mean, median, SD, and IQR, the ARE rates are negligible. The highest amount of the errors in IP are reported by the variable "V8" (number of rings) which is an integer attribute. As mentioned earlier, by adjusting the rounding parameters or the number of re-samples the error rate can be further reduced. DP statistics are calculated with $\epsilon = 4$ which should provide a very high data utility. However, compared to the IP solution, in the DP case the error is much higher for all the statistics except the count and the sum values.

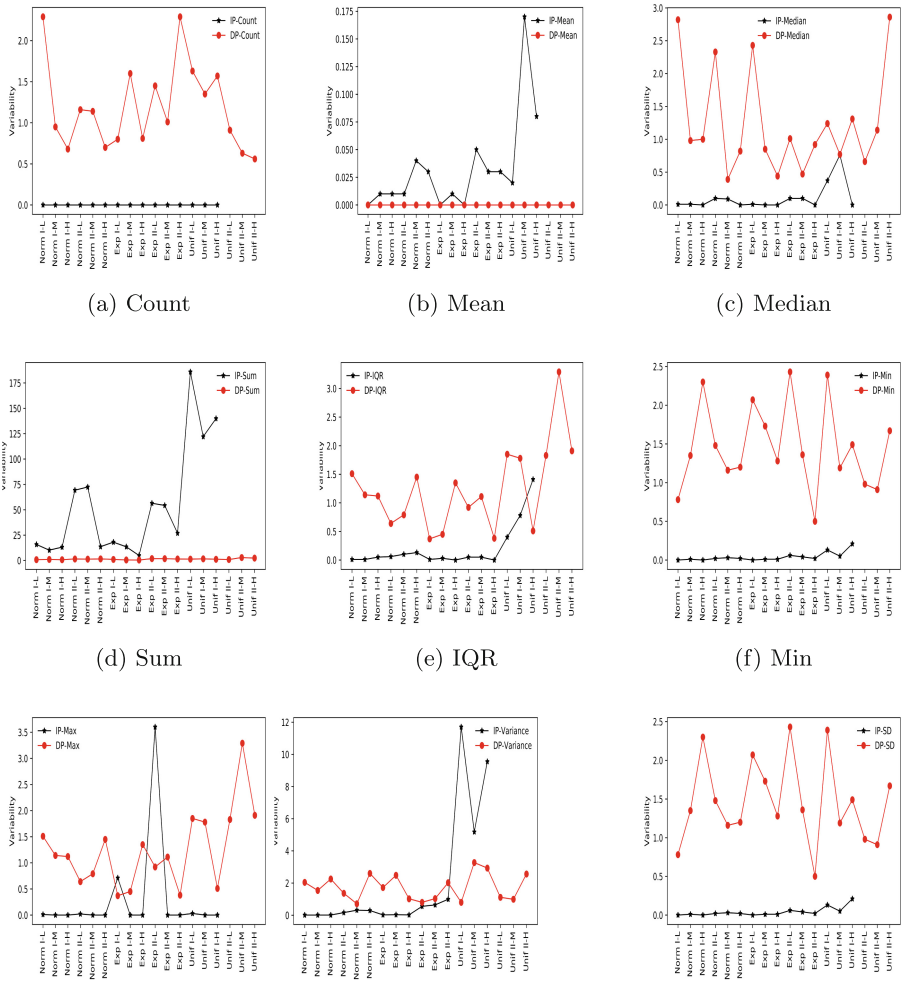
Moreover, with respect to IP we collected the frequency of occurrence (k) of the selected IP results for descriptive statistics computed over the 8 variables of the Abalone dataset. In other words, out of 5,000 re-samples what was the average rate of occurrence (ARO) of the selected IP statistic over the 8 variables (number of generators). It is respectively, 5000 for count, 3994 for mean, 3793 for median, 4251 for SD, 4652 for min, 4446 for max, 3891 for IQR, 4245 for variance and 9 for sum. For a total of 5,000 re-samples (approximately the size of the

Table 2. IP statistics and their standard deviation computed for synthetic datasets with different discretization methods. 1,000 re-samples are used for the computation.

Dataset	Count	Mean	Median	SD	Min	Max	IQR	Sum	Variance
Norm I Out Dis:	1000 ± 0	1.04 ± 0	1.1 ± 0.01	0.99 ± 0	-1.77 ± 0.05	3.97 ± 0.01	1.33 ± 0.01	668.9 ± 15.77	0.98 ± 0.01
Norm I in/out Dis:(L)	1000 ± 0	1.04 ± 0.01	1.1 ± 0.01	0.98 ± 0.01	-1.77 ± 0	3.96 ± 0	1.32 ± 0.01	678.03 ± 10.17	0.98 ± 0.01
Norm I in/out Dis:(H)	1000 ± 0	1.04 ± 0.01	1.09 ± 0	0.99 ± 0	-1.33 ± 0	3.4 ± 0	1.31 ± 0.05	670.87 ± 13.12	0.98 ± 0.01
Norm II Out Dis:	1000 ± 0	0.99 ± 0.01	1.05 ± 0.1	4.87 ± 0.02	-16.54 ± 2.78	16.52 ± 0.02	6.45 ± 0.06	680.42 ± 69.27	23.72 ± 0.16
Norm II in/out Dis:(L)	1000 ± 0	1 ± 0.04	0.98 ± 0.09	4.87 ± 0.03	-15.39 ± 1.42	16.54 ± 0	6.48 ± 0.1	661.09 ± 72.42	23.86 ± 0.31
Norm II in/out Dis:(H)	1000 ± 0	1 ± 0.03	0.92 ± 0	4.85 ± 0.02	-11.47 ± 0	12.69 ± 0	6.4 ± 0.13	718.88 ± 13.59	23.56 ± 0.29
Exp I Out Dis:	1000 ± 0	1.06 ± 0	0.73 ± 0.01	1.1 ± 0	0 ± 0	8.67 ± 0.71	1.18 ± 0.01	693.78 ± 18.03	1.21 ± 0.02
Exp I in/out Dis:(L)	1000 ± 0	1.06 ± 0.01	0.73 ± 0	1.1 ± 0.01	0 ± 0	8.2 ± 0	1.18 ± 0.03	675.47 ± 13.49	1.17 ± 0.03
Exp I in/out Dis:(H)	1000 ± 0	1.06 ± 0	0.72 ± 0	1.08 ± 0.01	0.01 ± 0	5.33 ± 0	1.15 ± 0	680.61 ± 4.89	1.15 ± 0.02
Exp II Out Dis:	1000 ± 0	5.34 ± 0.05	3.59 ± 0.1	5.39 ± 0.06	0 ± 0	32.89 ± 3.6	5.98 ± 0.05	3470.72 ± 56.4	29.06 ± 0.55
Exp II in/out Dis:(L)	1000 ± 0	5.32 ± 0.03	3.62 ± 0.1	5.41 ± 0.04	0 ± 0	35.3 ± 0	5.98 ± 0.05	3479.54 ± 54.26	29.51 ± 0.64
Exp II in/out Dis:(H)	1000 ± 0	5.35 ± 0.03	3.76 ± 0	5.4 ± 0.02	0.06 ± 0	25.45 ± 0	5.84 ± 0	3446.93 ± 26.87	29.2 ± 0.99
Unif I Out Dis:	1000 ± 0	50.69 ± 0.02	50.62 ± 0.37	28.95 ± 0.13	0.05 ± 0	99.93 ± 0.03	51.73 ± 0.4	NA $\pm$ 185.88	834.24 ± 11.7
Unif I in/out Dis:(L)	1000 ± 0	50.57 ± 0.17	50.75 ± 0.76	28.89 ± 0.05	0.06 ± 0	99.93 ± 0	52.05 ± 0.78	NA $\pm$ 122.08	844.04 ± 5.18
Unif I in/out Dis:(H)	1000 ± 0	50.75 ± 0.08	51.02 ± 0	28.91 ± 0.21	1.59 ± 0	99.1 ± 0	51.35 ± 1.41	NA $\pm$ 139.89	841.39 ± 9.55
Unif II Out Dis:	$1000 \pm$ NA	$486.36 \pm$ NA	$468.79 \pm$ NA	$303.57 \pm$ NA	$0.62 \pm$ NA	$997.81 \pm$ NA	$546.37 \pm$ NA	NA $\pm$ NA	NA $\pm$ NA
Unif II in/out Dis:(L)	$1000 \pm$ NA	$484.6 \pm$ NA	$475.7 \pm$ NA	$303.36 \pm$ NA	$0.6 \pm$ NA	$997.44 \pm$ NA	$540.01 \pm$ NA	NA $\pm$ NA	NA $\pm$ NA
Unif II in/out Dis:(H)	$1000 \pm$ NA	$488.25 \pm$ NA	$483.47 \pm$ NA	$304.25 \pm$ NA	$5.37 \pm$ NA	$988.37 \pm$ NA	$552.93 \pm$ NA	NA $\pm$ NA	NA $\pm$ NA

Table 3. DP statistics and their standard deviation computed for synthetic datasets with different ϵ values.

Dataset	Count	Mean	Median	SD	Min	Max	IQR	Sum	Variance
Norm I ($\epsilon = 0.01$)	1100.7 ± 2.29	0.59 ± 0	0.61 ± 2.82	19.78 ± 0.78	8.47 ± 1.01	7.86 ± 1.51	7.86 ± 1.51	1435.68 ± 0.79	5.62 ± 2.04
Norm I ($\epsilon = 2$)	1001.11 ± 0.95	0 ± 0	1.51 ± 0.98	1.19 ± 1.35	-2.46 ± 2.57	3.92 ± 1.14	3.92 ± 1.14	1039.42 ± 1.05	1.95 ± 1.54
Norm I ($\epsilon = 4$)	1000.38 ± 0.68	0 ± 0	1.78 ± 1	1.88 ± 2.3	-1.88 ± 0.4	4.91 ± 1.12	4.91 ± 1.12	1038.11 ± 0.75	-0.47 ± 2.25
Norm II ($\epsilon = 0.01$)	1100.02 ± 1.16	3.51 ± 0	1.65 ± 2.33	115.8 ± 1.48	486.82 ± 2.9	208.31 ± 0.64	208.31 ± 0.64	2651.68 ± 1.46	146.96 ± 1.36
Norm II ($\epsilon = 2$)	1000.87 ± 1.14	0.02 ± 0	0.44 ± 0.39	5.41 ± 1.16	-17.04 ± 2.53	17.47 ± 0.79	17.47 ± 0.79	1004.16 ± 1.29	24.73 ± 0.71
Norm II ($\epsilon = 4$)	1000.88 ± 0.7	0.01 ± 0	1.38 ± 0.82	5.46 ± 1.2	-16.99 ± 0.54	16.97 ± 1.45	16.97 ± 1.45	998.8 ± 1.51	24.63 ± 2.6
Exp I ($\epsilon = 0.01$)	1099.28 ± 0.8	0.9 ± 0	0.83 ± 2.43	30.45 ± 2.07	0.08 ± 0.76	167.36 ± 0.37	167.36 ± 0.37	1962.91 ± 1.03	8.29 ± 1.72
Exp I ($\epsilon = 2$)	1001.56 ± 1.6	0.01 ± 0	1.35 ± 0.85	0.85 ± 1.73	0.59 ± 0.95	9.34 ± 0.45	9.34 ± 0.45	1066.92 ± 0.53	-0.24 ± 2.48
Exp I ($\epsilon = 4$)	1000.16 ± 0.81	0 ± 0	1.44 ± 0.44	1.23 ± 1.28	-0.18 ± 1.38	8.9 ± 1.35	8.9 ± 1.35	1065.72 ± 0.44	0.53 ± 1.02
Exp II ($\epsilon = 0.01$)	1100.38 ± 1.45	3.94 ± 0	3.47 ± 1.01	127.57 ± 2.43	0.98 ± 0.76	842.91 ± 0.92	842.91 ± 0.92	9264.07 ± 1.8	184.24 ± 0.8
Exp II ($\epsilon = 2$)	998.98 ± 1.01	0.03 ± 0	3.68 ± 0.47	5.22 ± 1.36	-0.97 ± 0.75	43.32 ± 1.11	43.32 ± 1.11	5352.31 ± 1.79	29.73 ± 1.03
Exp II ($\epsilon = 4$)	1000.85 ± 2.29	0.02 ± 0	4.52 ± 0.92	6.38 ± 0.5	0.66 ± 1.09	42.02 ± 0.38	42.02 ± 0.38	5342.18 ± 1.43	30.81 ± 2.02
Unif I ($\epsilon = 0.01$)	1099.71 ± 1.63	10.05 ± 0	57.54 ± 1.24	345.43 ± 2.39	7.71 ± 0.42	107.37 ± 1.85	107.37 ± 1.85	60624.79 ± 1.33	1838.98 ± 0.8
Unif I ($\epsilon = 2$)	999.31 ± 1.35	0.1 ± 0	49.95 ± 0.77	31.06 ± 1.19	0.02 ± 0.8	100.09 ± 1.78	100.09 ± 1.78	50678.05 ± 1.62	843.86 ± 3.27
Unif I ($\epsilon = 4$)	1000.32 ± 1.57	0.08 ± 0	50.53 ± 1.31	30.16 ± 1.49	0.39 ± 1.18	99.55 ± 0.51	99.55 ± 0.51	50652.64 ± 1.2	842.97 ± 2.93
Unif II ($\epsilon = 0.01$)	1099.71 ± 0.91	100.22 ± 0	593.97 ± 0.66	3457.43 ± 0.98	12.87 ± 1.13	1106.95 ± 1.83	1106.95 ± 1.83	587122.17 ± 0.92	191734.16 ± 1.11
Unif II ($\epsilon = 2$)	1001.13 ± 0.63	0.99 ± 0	472.36 ± 1.14	319.76 ± 0.91	1.15 ± 0.86	999.76 ± 3.29	999.76 ± 3.29	487829.21 ± 2.88	92763.76 ± 0.99
Unif II ($\epsilon = 4$)	999.79 ± 0.56	$0.74 \pm \pm 0$	475.33 ± 2.86	312.99 ± 1.67	0.77 ± 1.48	997.72 ± 1.91	997.72 ± 1.91	487580.65 ± 2.32	92515.39 ± 2.56



(a) Count (b) Mean (c) Median

(d) Sum (e) IQR (f) Min

(g) Max (h) Variance (i) SD

Fig. 1. Standard deviation of IP and DP statistics over multiple iterations. Each synthetic data distribution is tagged as L, M and H which indicate Low, Medium and High privacy levels. For IP, L indicates output discretization, M indicates low input discretization combined with output discretization and H indicates high input discretization with output discretization. For DP L, M and H respectively indicate ϵ values 0.01,2 and 4.

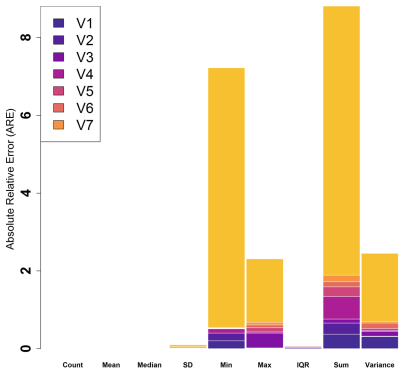
Table 4. Absolute Relative Error (ARE) of IP statistics computed on synthetic datasets with different discretization methods.

Dataset	Count	Mean	Median	SD	Min	Max	IQR	Sum	Variance
Norm I Out Dis:	0.00	0.00	0.01	0.00	0.05	0.00	0.00	0.35	0.01
Norm I in/out Dis:(L)	0.00	0.00	0.01	0.01	0.05	0.01	0.01	0.35	0.01
Norm I in/out Dis:(H)	0.00	0.00	0.00	0.00	0.28	0.15	0.02	0.35	0.01
Norm II Out Dis:	0.00	0.01	0.03	0.01	1.09	0.01	0.00	0.30	0.04
Norm II in/out Dis:(L)	0.00	0.00	0.04	0.01	1.71	0.01	0.02	0.32	0.10
Norm II in/out Dis:(H)	0.00	0.00	0.09	0.03	3.82	0.97	0.04	0.27	0.21
Exp I Out Dis:	0.00	0.00	0.01	0.01	0.00	0.08	0.00	0.36	0.02
Exp I in/out Dis:(L)	0.00	0.00	0.01	0.01	0.00	0.20	0.00	0.37	0.02
Exp I in/out Dis:(H)	0.00	0.00	0.00	0.01	0.01	0.92	0.02	0.37	0.04
Exp II Out Dis:	0.00	0.01	0.07	0.01	0.00	1.61	0.01	1.79	0.14
Exp II in/out Dis:(L)	0.00	0.01	0.04	0.01	0.00	1.01	0.01	1.79	0.32
Exp II in/out Dis:(H)	0.00	0.02	0.09	0.00	0.03	3.48	0.12	1.82	0.00
Unif I Out Dis:	0.00	0.06	0.05	0.03	0.03	0.01	0.09	NA	5.54
Unif I in/out Dis:(L)	0.00	0.06	0.17	0.09	0.03	0.01	0.15	NA	4.50
Unif I in/out Dis:(H)	0.00	0.12	0.41	0.07	0.86	0.22	0.38	NA	1.79
Unif II Out Dis:	0.00	0.94	3.22	0.18	0.03	0.03	0.23	NA	NA
Unif II in/out Dis:(L)	0.00	2.63	3.11	0.40	0.01	0.12	4.54	NA	NA
Unif II in/out Dis:(H)	0.00	0.89	10.22	0.50	2.58	2.39	5.14	NA	NA

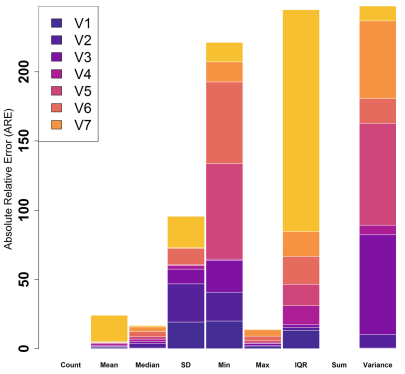
Table 5. Absolute Relative Error (ARE) of DP statistics computed on synthetic datasets with different ϵ values.

Dataset	Count	Mean	Median	SD	Min	Max	IQR	Sum	Variance
Norm I ($\epsilon = 0.01$)	0.10	0.43	0.44	19.03	5.56	0.97	4.89	0.39	4.76
Norm I ($\epsilon = 2$)	0.00	1.00	0.38	0.20	0.33	0.02	1.94	0.00	1.00
Norm I ($\epsilon = 4$)	0.00	1.00	0.63	0.90	0.01	0.23	2.68	0.00	1.48
Norm II ($\epsilon = 0.01$)	0.10	2.42	0.58	112.32	272.23	48.09	151.27	1.60	126.32
Norm II ($\epsilon = 2$)	0.00	0.94	0.53	0.54	0.82	0.23	8.26	0.01	0.99
Norm II ($\epsilon = 4$)	0.00	0.95	0.33	0.59	0.85	0.10	7.88	0.00	0.89
Exp I ($\epsilon = 0.01$)	0.10	0.16	0.10	29.73	0.04	39.72	124.53	0.87	7.28
Exp I ($\epsilon = 2$)	0.00	1.02	0.57	0.24	0.32	0.09	6.12	0.00	1.47
Exp I ($\epsilon = 4$)	0.00	1.03	0.66	0.14	0.10	0.02	5.79	0.00	0.68
Exp II ($\epsilon = 0.01$)	0.10	1.34	0.18	123.70	0.53	201.52	627.16	3.79	158.97
Exp II ($\epsilon = 2$)	0.00	5.11	0.01	0.19	0.52	1.00	27.97	0.02	0.54
Exp II ($\epsilon = 4$)	0.00	5.12	0.78	0.99	0.35	0.68	27.00	0.01	1.65
Unif I ($\epsilon = 0.01$)	0.10	39.13	6.38	320.44	4.15	1.85	41.60	9.64	1024.67
Unif I ($\epsilon = 2$)	0.00	48.73	0.57	2.11	0.01	0.03	36.15	0.05	4.32
Unif I ($\epsilon = 4$)	0.00	48.75	0.04	1.20	0.21	0.11	35.74	0.02	3.41
Unif II ($\epsilon = 0.01$)	0.10	373.32	111.39	3193.41	6.62	27.34	420.31	96.24	101990.65
Unif II ($\epsilon = 2$)	0.00	469.02	0.05	16.21	0.31	0.46	339.99	0.48	510.80
Unif II ($\epsilon = 4$)	0.00	469.26	2.77	9.35	0.11	0.05	338.46	0.24	256.13

dataset) following input and output discretization number of generators seems to be very high showing that the chances to distinguish the exact data records used to compute a given statistic are minimal. However, it is being repeatedly shown that the for summation queries IP might not be an ideal solution.

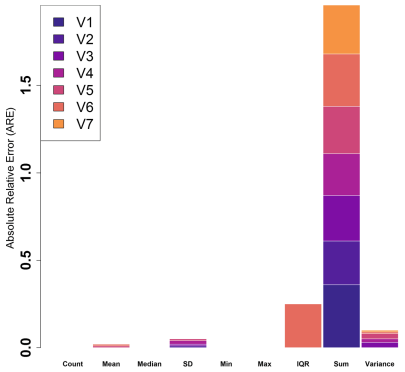


(a) IP with output discretization and minimum input discretization (MA y=2), 4177 re-samples are used to compute the IP statistics.

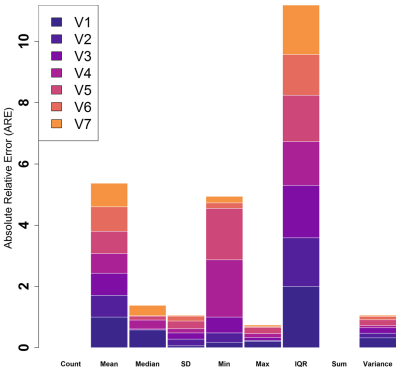


(b) DP with $\epsilon = 4$

Fig. 2. Absolute relative error (ARE) for descriptive statistics computed over the numerical variables of the Abalone dataset.



(a) IP with output discretization and minimum input discretization (MA y=2), 700 re-samples are used to compute the IP statistics.



(b) DP with $\epsilon = 4$

Fig. 3. Absolute relative error (ARE) for descriptive statistics computed over the numerical variables of the Breast Cancer dataset.

Figure 3 depicts the computation of IP and DP statistics on UCI breast cancer dataset. To comparatively evaluate the results ARE rates are computed per variable. The results show the same pattern as the Abalone dataset. Compared to DP the ARE rates are low for IP except for calculating the sum. The poor performance of the IP with respect to summation can be attributed to use of re-sampling.

These results shows us that IP results have a high utility value compared to DP in most of the cases. Therefore, this method can be used for releasing aggregated statistics without compromising the privacy of the sensitive data. However, in order to use this in terms of large scale databases, it is required to improve the computational efficiency of the process further.

Table 6. Summation of absolute relative error for different statistics computed using IP and DP

Privacy model	Count	Mean	Median	SD	Min	Max	IQR
IP	0	4.75	17.58	1.38	10.58	11.23	10.78
DP	0.6	1468.73	126.39	3831.29	293.07	322.51	2207.74

6 Conclusion

In this paper, we have discussed how to provide integral privacy for descriptive statistics computation while maintaining the robustness and the utility of the final results. We have proposed a re-sampling and discretization based approach to achieve integral privacy and empirically shown that integral privacy based solution works better than the differential privacy based solution in most of the cases. Especially, it is noted that the proposed solution can easily be used with small datasets where differential privacy usually fails in terms of utility. However, further work is required to minimize the computational cost and to introduce a formal method to derive the minimum number of re-samples required to achieve integral privacy for a given dataset. And also, we hope to develop an inference attack to assess the effectiveness of integral privacy in our future work.

References

1. Torra, V., Navarro-Arribas, G.: Integral privacy. In: Foresti, S., Persiano, G. (eds.) CANS 2016. LNCS, vol. 10052, pp. 661–669. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48965-0_44
2. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
3. Blum, A., Dwork, C., McSherry, F., Nissim, K.: Practical privacy: the SuLQ framework. In: Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 128–138. ACM (2005)

4. Dwork, C., Smith, A.: Differential privacy for statistics: what we know and what we want to learn. J. Priv. Confid. **1**(2) (2010)
5. Dwork, C., Lei, J.: Differential privacy and robust statistics. In: STOC, vol. 9, pp. 371–380 (2009)
6. Clifton, C., Tassa, T.: On syntactic anonymity and differential privacy. In: 2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW), pp. 88–93. IEEE (2013)
7. Senavirathne, N., Torra, V.: Integrally private model selection for decision trees. Comput. Secur. **83**, 167–181 (2019)
8. Senavirathne, N., Torra, V.: Approximating robust linear regression with an integral privacy guarantee. In: 2018 16th Annual Conference on Privacy, Security and Trust (PST), pp. 1–10. IEEE (2018)
9. Efron, B.: Bootstrap methods: another look at the Jackknife. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics), pp. 569–593. Springer, New York (1992). https://doi.org/10.1007/978-1-4612-4380-9_41
10. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Megías, D.: Individual differential privacy: a utility-preserving formulation of differential privacy guarantees. IEEE Trans. Inf. Forensics Secur. **12**(6), 1418–1429 (2017)
11. Li, N., Lyu, M., Dong, S., Yang, W.: Differential privacy: from theory to practice. Synth. Lect. Inf. Secur. Priv. Trust. **8**(4), 1–138 (2016)
12. Dwork, C., Lei, J.: Differential privacy and robust statistics. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, 31 May–June 2 2009, pp 371–380 (2009)