Adaptive Differentially Private Histogram of Low-Dimensional Data

Chengfang Fang Ee-Chien Chang*

School of Computing
National University of Singapore
{c.fang, changec}@comp.nus.edu.sg

Abstract. We want to publish low-dimensional points, for example 2D spatial points, in a differentially private manner. Most existing mechanisms publish noisy frequency counts of points in a fixed predefined partition. Arguably, histograms with adaptive partition, for example V-optimal and equi-depth histograms, which have smaller bin-widths in denser regions, would provide more statistical information. However, as the adaptive partitions leak significant information about the dataset, it is not clear how differentially private partitions can be published accurately. In this paper, we propose a simple method based on the observation that the sensitivity of publishing the *sorted* sequence of a dataset is independent of the size of dataset. Together with isotonic regression, the dataset can be reconstructed with high accuracy. One advantage of the proposed method is its simplicity, in the sense that there are only a few parameters to be determined. Furthermore, the parameters can be estimated solely from the privacy requirement ϵ and the total number of points, and hence do not leak information about the data. Although the parameters are chosen to minimize the earth mover's distance between the published data and original data, empirical studies show that the proposed method also achieves high accuracy w.r.t. to some other measurements, for example range query and order statistics.

1 Introduction

The popularity of personal devices equipped with location sensors leads to a large amount of location data being gathered. Such data contain rich information and would be valuable if they can be shared and published. As the data may reveal location of an identified individual, it is important to anonymize the data before publishing. The recently developed notion of differential privacy [5] provides a strong form of privacy assurance regardless of the background information held by the adversaries. Such assurance is important, as many case studies and past events have shown that a seemingly anonymized dataset together with additional knowledge held by the adversary could reveal information on individuals.

Most studies on differential privacy focus on publishing statistical values, for instance, k-means [3], private coreset [7], and median of the database [20].

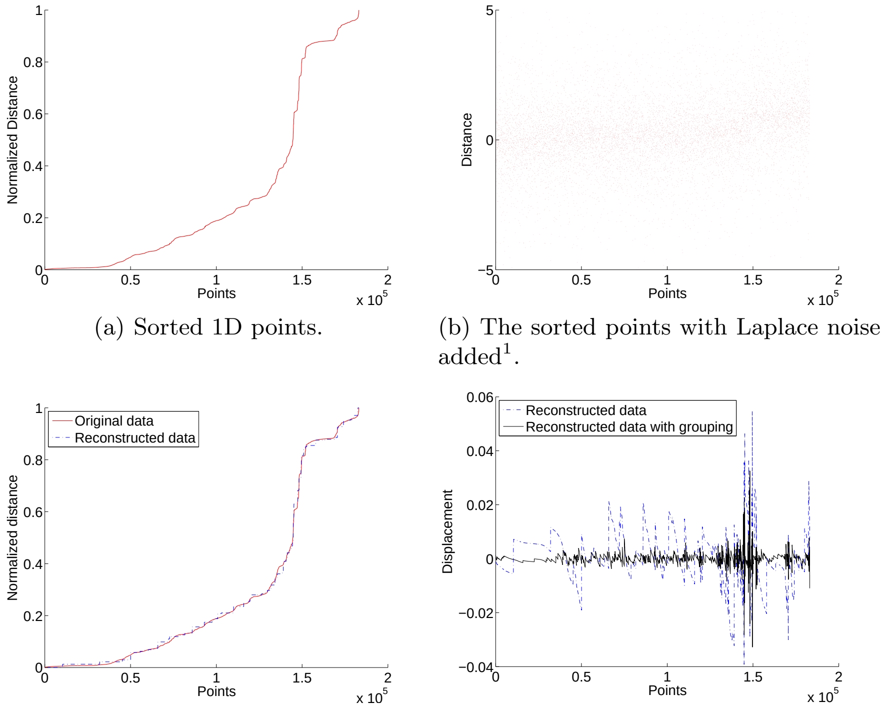
* Chang is partially supported by Grant TDSI/09-003/1A.

Publishing specific statistics or data-mining results is meaningful if the publisher knows what the public specifically wants. However, there are situations where the publishers want to give the public greater flexibility in analyzing and exploring the data, for example, using different visualization techniques. In such scenarios, it is desired to *"publish data, not the data mining result"* [8].

The histogram of a dataset contains rich information that can be harvested by subsequent analysis. In the context of different privacy, *parallel composition* can be exploited to treat non-overlapping bins independently and thus achieving high accuracy. There are a number of research efforts [14, 2] investigating the dependencies of frequencies counts of fixed overlapping bins, where parallel composition can not be directly applied. Such overlapping bins are interesting as different domain partition could lead to different accuracy and utility. For instance, Xiao et al. [28] proposes publishing wavelet coefficients of an equi-width histogram, which can be viewed as publishing a series of equi-width histograms with different bin-widths, and is able to provide higher accuracy in answering range queries compare to a single equi-width histogram.

It is generally well accepted that equi-depth histogram and V-optimal histogram provide more useful statistical information compare to equi-width histogram [21, 22], especially for multidimensional data. These histograms are adaptive in the sense that the domain partitions are derived from the data such that denser regions will have smaller bin-widths and the sparser regions will have larger bin-widths, as illustrated in Fig. 7(b). Since the bin-widths are derived from the dataset, they leak information about the original dataset. There are relatively few works that consider adaptive histogram in the context of differential privacy. One exception is the work by Xiao et al. [29]. Their method consists of two steps where firstly synthetic data are generated from the differentially private equi-width histogram. After that, a k-d tree (which can be viewed as an adaptive histogram) is generated from the synthetic data, and the noisy counts are then released with the partition. Machanavajjhala et al. [16] proposed a mechanism that publishes 2D histograms with varying bin-widths, where the bin-widths are determined from a previously released similar data. The histograms generated are not adaptive in the sense that the partitions do not depend on the data to be published.

In this paper, instead of publishing the noisy frequency counts in equi-width bins, we propose a method that directly publishes the noisy data, which in turn leads to an adaptive histogram. To illustrate, let us first consider a dataset consisting of a set of real numbers from the unit interval, for example, the normalized distance of Twitter users' locations [1] to New York City (Fig. 1(a)). We observe that sorting, as a function that takes in a set of real numbers from the unit interval and outputs the sorted sequence, interestingly has sensitivity one (Theorem 1). Hence, the mechanism that first sorts, and then adds independent Laplace noise of $\mathrm{LAP}(1/\epsilon)$ to each element achieves ϵ-differential privacy. Fig. 1(b) shows the noisy output data after the Laplace noise has been added to the sorted sequence. Although seemingly noisy, there are dependencies to be exploited because the original sequence is sorted. By using isotonic regression,



(a) Sorted 1D points.

(b) The sorted points with Laplace noise added[1].

(c) Reconstructed with isotonic regression.

(d) The differences of the reconstructed points from the original.

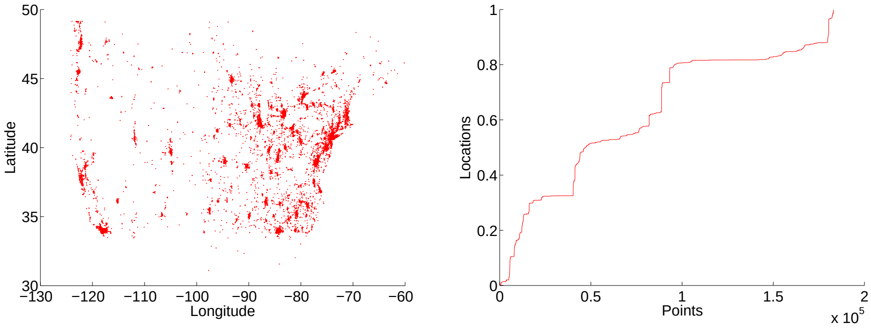
Fig. 1. Overview of proposed method.

the noise can be significantly reduced (Fig. 1(c)). To further reduce noise, before adding the Laplace noise, consecutive elements in the sorted data can be grouped and each point is replaced by the average of its group. Fig. 1(d) shows the difference of the original and the reconstructed points with and without grouping.

To extend the proposed method to higher dimension data, for example, location data of 183,072 Twitter users in North America as shown in Fig. 2(a), we employ *locality-preserving mapping* to map the multidimensional data to one-dimension (Fig. 2(b)), such that any two close points in the one-dimension domain are mapped from two close multidimensional points. After that, the publisher can apply the proposed method on the 1D points, and publish the reverse mapped multidimensional points.

One desired feature of our scheme is its simplicity: there is only one parameter, the group size, to be determined. The group size affects the accuracy in three ways: (1) its effect on the *generalization error*, which is introduced due to

[1] To avoid clogging, only 10% of the points (randomly chosen) are plotted for Fig. 1(b) and Fig. 2(a).



(a) Locations of Twitter users.

(b) The sorted 1D images of the location data.

Fig. 2. Twitter location data and their 1D images of a locality-preserving mapping.

averaging; (2) its effect on the level of Laplace noise to be added by the differentially private mechanism; and (3) its effect on the number of constraints in the isotonic regression. Based on our error model, the optimal parameter can be estimated without knowledge of the dataset distribution. In contrast, many existing methods have many parameters whose optimal values are difficult to be determined differentially privately. For instance, although the equi-width histogram has only one parameter, i.e. the bin-width, its value significantly affects the accuracy, and it is not clear how to differentially privately obtain a good choice of the bin width.

Our error model utilizes the earth mover's distance (EMD) to measure the accuracy of the published data. Some existing works measure the accuracy of a histogram by its distance, such as L_2 distance or KL divergence, to a reference equi-width histogram. One limitation of this measurement is that the reference histogram can be arbitrary and thus arguably ill-defined. If the reference bin-width is too small, each bin will contain either one or no point, which leads to significantly large distance from a seemingly accurate histogram. On the other hand, if its bin-width is too large, the reference histogram would be over generalized. In contrast, EMD measures the distance of the published data and original points, where the "reference" is the original points and thus well-defined. We conduct empirical studies to compare against a few related known methods: equi-width histogram, wavelet-based method [28] and smooth sensitivity based median-finding [20]. Although our method is designed to minimize the EMD, it also attains high accuracy w.r.t. other measurements. Empirical studies shows that our method outperforms the wavelet-based method w.r.t. accuracy of range-query, even for ranges with large sizes. It is also comparable to the smooth sensitivity based method in publishing median.

Organization: We first describe some background materials in the next section. In Section 3 we present our main ideas and mechanism, and show that the proposed mechanism achieves differential privacy in Section 4. Next, in Section

5, we formulate and analyze how the group size affects the accuracy and derive a strategy to choose the group size based on this model. In Section 6, we compare our mechanism with three known mechanisms: (1) equi-width histogram, (2) wavelet-based method, and (3) smooth sensitivity based median-finding. In Section 7, we discuss the extensions and limitations of our method. Lastly, we describe related works in Section 8 and conclude in Section 9.

2 Background

2.1 Differential Privacy and Laplace Noise

We treat a database as a multi-set (i.e. a set with possibly repeating elements), and consider two datasets D_1 and D_2 of size n to be neighbors when D_2 can be obtained from D_1 by replacing one element, i.e. $D_1 = \{x\} \cup D_2 \setminus \{y\}$ for some x and y. Differential privacy with this definition of neighborhood is known as the *bounded differential privacy* [6, 13].

A randomized algorithm (also known as *mechanism*) $\mathcal{A}$ achieves ϵ differential privacy if,

$$Pr[\mathcal{A}(D_1) \in S] \le exp(\epsilon) \times Pr[\mathcal{A}(D_2) \in S]$$

for all $S \subseteq Range(\mathcal{A})$, where $Range(\mathcal{A})$ denotes the output range of the algorithm $\mathcal{A}$, and for any pair of neighboring datasets D_1 and D_2.

For a function $f : D \to \mathbb{R}^k$, the *sensitivity* [5] of f is defined as

$$\Delta(f) := \max \|f(D_1) - f(D_2)\|_1,$$

where the maximum is taken over all pairs of neighboring D_1 and D_2. It is shown [6] that the mechanism $\mathcal{A}$

$$\mathcal{A}(D) = f(D) + (Lap(\Delta(f)/\epsilon))^k$$

achieves ϵ-differential privacy, where $(Lap(\Delta(f)/\epsilon))^k$ is a vector of k independently and randomly chosen values from the Laplace distribution with standard deviation $2\Delta(f)/\epsilon$.

2.2 Isotonic Regression

Given a sequence of n real numbers $a_1, \ldots, a_n$, the problem of finding the least-square fit $x_1, \ldots, x_n$ subjected to the constraints $x_i \le x_j$ for all $i < j \le n$ is known as the isotonic regression. Formally, we want to find the $x_1, \ldots, x_n$ that minimizes

$$\sum_{i=1}^{n} (x_i - a_i)^2, \qquad \text{subjected to } x_i \le x_j \text{ for all } 1 \le i < j \le n.$$

The unique solution can be efficiently found using pool-adjacent-violators algorithms in $O(n)$ time [10]. When minimizing w.r.t. ℓ-1 norm, there is also an efficient $O(n \log n)$ algorithm [25]. There are many variants of isotonic regression, for example, variants with a smoothness component in the objective function [27, 17].

2.3 Locality-Preserving Mapping

A locality-preserving mapping $T : [0,1]^d \rightarrow [0,1]$ maps d-dimensional points to the unit interval, while preserving locality. In this paper, we seek a mapping that, if the mapped points $T(x)$, $T(y)$ are "close", then x and y are "close" in the d-dimensional space. More specifically, there is some constant c s.t. for any x, y in the domain of the mapping T,

$$\|x - y\|_2 \leq c \cdot (\|T(x) - T(y)\|)^{1/d}. \tag{1}$$

The well-known Hilbert curve [9] is a locality-preserving mapping. It is shown that for any 2D points x, y in the domain of T, $\|x - y\|_2 \leq 3\sqrt{|T(x) - T(y)|}$. Niedermeier et al. [19] showed that with careful construction, the bound can be improved to $2\sqrt{|T(x) - T(y)|}$ for 2D points and $3.25\sqrt[3]{\|T(x) - T(y)\|}$ for 3D points. In our construction, for simplicity, we use Hilbert curve in our experiments.

Note that it is challenging in preserving locality "in the other direction", that is, any two "close" points in the d-dimensional domain are mapped to "close" points in the one-dimensional range [18]. Fortunately, in our problem, such property is not required.

2.4 Datasets

We conduct experiments on two datasets: locations of Twitter users [1] (herein called the Twitter location dataset) and the dataset collected by Kaluža et al. [12] (herein called Kaluža's dataset). The Twitter location dataset contains over 1 million Twitter users' data from the period of March 2006 to March 2010, among which around 200,000 tuples are labeled with location (represented in latitude and longitude) and most of the tuples are in the North American continent, concentrating in regions around the state of New York and California. Fig. 2(a) shows the cropped region covering most of the North American continent. The cropped region contains 183,072 tuples. The Kaluža's dataset contains 164,860 tuples collected from tags that continuously record the location information of 5 individuals. While some of the tuples consist of many attributes, in our experiments, only the 2D location data are being used.

3 Proposed Approach

Before receiving the data, the publisher has to make a few design choices. The publisher need to decide on a locality-preserving mapping T, and the strategy (which is represented as a lookup table) of determining the group size from the privacy requirement ϵ and the size of dataset n. Now, given the dataset D of size n, and the privacy requirement ϵ, the publisher carries out the following:

A1. The publisher maps each point in D to a real number in the unit interval $[0,1]$ using T, and lookups the group size based on n and ϵ. Let $T(D)$ be the set of transformed points. For clarity in exposition, let us assume that k divides n.

A2. The publisher sorts the mapped points, divides the sorted sequence into groups of k consecutive elements, and then for each group, determines its average over the k elements. Let the averages be $S = \langle s_1, \dots, s_{n/k} \rangle$.

A3. The publisher releases $\widetilde{S} = S + (\text{Lap}(\epsilon^{-1})/k)^{(n/k)}$ and the group size k.

A public user may extract information from the published data as follow:

B1. The user performs isotonic regression on $\widetilde{S}$ and obtains $\text{IR}(\widetilde{S})$, and then replaces each element $\widetilde{s}_i$ in $\text{IR}(\widetilde{S})$ with k points of value $\widetilde{s}_i$. Let P be the set of resulting points.

B2. The user maps the data point back to the original domain, that is, computes $\widetilde{D} = T^{-1}(P)$. Let us call $\widetilde{D}$ the reconstructed data.

Note that the public user is not confined to performing step B1 and B2. The user may, for example, incorporates some background knowledge to enhance accuracy. To relieve the public from computing step B1 and B2, the regression and the inverse mapping can be carried out by the publisher on behalf of the users. Nevertheless, the raw data $\widetilde{S}$ should be (although it is not necessary) published alongside the reconstructed data for further statistical analysis.

4 Security Analysis

In this section, we show that the proposed mechanism (Step A1 to A3) achieves differential privacy. The following theorem shows that sorting, as a function, interestingly has sensitivity 1. Note that a straightforward analysis that treats each element independently could lead to a bound of n, which is too large to be useful.

Theorem 1. *Let $S_n(D)$ be a function that on input D, which is a multi-set containing n real numbers from the unit interval $[0, 1]$, outputs the sorted sequence of elements in D. The sensitivity of S_n w.r.t. the bounded differential privacy is 1.*

Proof. Let D_1 and D_2 be any two neighboring datasets. Let $\langle x_1, x_2 \dots x_i \dots x_n \rangle$ be $S_n(D_1)$, i.e. the sorted sequence of D_1. WLOG, let us assume that an element x_i is replaced by a larger value A to give D_2, for some $1 \leq i \leq n-1$ and $x_i < A$. Let j to be largest index s.t. $x_j < A \leq 1$. Hence, the sorted sequence of D_2 is:

$$x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_j, A, x_{j+1}, \dots, x_n.$$

The L_1 difference due to the replacement is,

$$\begin{aligned}
&\|S_n(D_1) - S_n(D_2)\|_1 \\
&= |x_{i+1} - x_i| + |x_{i+2} - x_{i+1}| + |x_j - x_{j-1}| + |A - x_j| \\
&= (x_{i+1} - x_i) + (x_{i+2} - x_{i+1}) + (x_j - x_{j-1}) + (A - x_j) \\
&= A - x_i \leq 1.
\end{aligned}$$

We can easily find an instance of D_1 and D_2 where the difference $A - x_i = 1$. Hence, the sensitivity is 1. □

Since the sensitivity is 1, the mechanism $S_n(D)+Lap(1/\epsilon)^n$ enjoys ϵ-differential privacy. Also note that the value of n is fixed. Hence, the size of D is not a secret and is made known to the public.

The following corollary shows (proof omitted) that grouping (in Step A2) has no effect on the sensitivity.

Corollary 1. *Consider a partition* $H = \{h_1, h_2 \dots h_m\}$ *of the indices* $\{1, 2, \dots, n\}$. *Let* $S_H(D)$ *be the function that, on input* D*, which is a multi-set containing* n *real numbers from the unit interval* $[0, 1]$*, outputs a sequence of* m *numbers:*

$$y_i = \sum_{j \in h_i} x_j,$$

for $1 \leq i \leq m$ *where* $\langle x_1, x_2, \dots, x_n \rangle$ *is the sorted sequence of* D*. The sensitivity of* S_H *is 1.*

Note that the grouping in step A2 is a special partition with equal-sized h_i's, whereas Corollary 1 gives a more general result where H can be any partition. From Corollary 1, the proposed mechanism achieves ϵ-differential privacy.

5 Analysis and Parameter Determination

The main goal of this section is to analyze the effect of the privacy requirement ϵ, dataset size n and the group size k on the error in the reconstructed data, which in turn provides a strategy in choosing the parameter k given n and ϵ.

Intuitively, when n and ϵ are fixed, the choice of parameter k affects the accuracy in following three ways: (1) a larger k decreases the number of constraints in isotonic regression, which leads to lower noise reduction; (2) a larger k reduces the effect of the Laplace noise; and (3) a larger k introduces higher generalization error due to averaging.

Our analysis consists of the following parts. We first describe our utility function in Section 5.1. In Section 5.2, we consider the case where $k = 1$ and empirically show that the expected error of a typical dataset can be well approximated by the expected error on a synthetic equally-spaced dataset. Let us call this error $Err_{n,\epsilon}$. Next in Section 5.3, we investigate and estimate the generalization error due to the averaging and show that with a reasonable assumption on the dataset distribution, the expected error can be approximated by $\frac{k}{4n}$. Let us call this error $Gen_{n,k}$. Finally, in Section 5.4, we consider the general case of $k \geq 1$ and give an approximation of the expected error in terms of $Err_{n,\epsilon}$ and $Gen_{n,k}$.

5.1 Error function

We use an error function based on the earth mover's distance(EMD) [24] to quantify the utility of the published data. The EMD between two pointsets of equal size is defined to be the minimum cost of bipartite matching between the two sets, where the cost of an edge linking two points is the cost of moving one point to the other. Hence, EMD can be viewed as the minimum cost of transforming one pointset to the other. Different variants of EMD differ on how the cost is defined. In this paper, we adopt the typical definition that defines the cost as the Euclidean distant between the two points.

In one-dimensional space, the EMD between two sets D and $\widetilde{D}$ is simply the L_1 norm of the differences between the two respective sorted sequences, i.e. $\|S_n(D)-S_n(\widetilde{D})\|_1$, which can be efficiently computed. Recall that $S_n(D)$ outputs the sorted sequence of elements in D. In other words,

$$\mathrm{EMD}(D,\widetilde{D}) = \sum_{i=1}^{n} |p_i - \widetilde{p}_i|, \tag{2}$$

where p_i's and $\widetilde{p}_i$'s are the sorted sequence of D and $\widetilde{D}$ respectively. Note that this definition assumes D and $\widetilde{D}$ have the same number of points, which is ensured by step B1 of our scheme.

Given a dataset D and the published dataset $\widetilde{D}$ of a mechanism $\mathcal{M}$ where $|D| = |\widetilde{D}| = n$, let us define the *normalized error* as $\frac{1}{n}\mathrm{EMD}(D,\widetilde{D})$ and denote $Err_{\mathcal{M},\mathcal{D}}$ the expected normalized error,

$$Err_{\mathcal{M},D} = Exp\left[\ \frac{1}{n}\ \mathrm{EMD}(D,\widetilde{D})\ \right], \tag{3}$$

where the expectation is taken over the randomness in the mechanism.

Our mechanism publishes $\widetilde{D}$ based on two parameters: the privacy requirement ϵ and the group size k. Therefore, let us write $Err_{\epsilon,k,D}$ for the expected normalized error of the dataset published in Step B2.

5.2 Effects on Isotonic Regression

Let us consider the expected normalized error when $k = 1$, in other words, we first consider the mechanism without grouping. In such case, the reconstructed dataset is $\mathrm{IR}(S_n(D) + \mathrm{Lap}(\epsilon^{-1})^n)$. Thus, the expected normalized error is

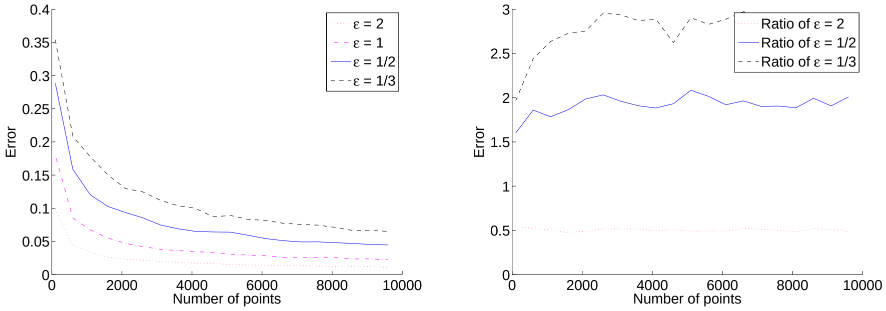
$$Err_{\epsilon,1,D} = Exp\left[\frac{1}{n}\ \mathrm{EMD}(D, \mathrm{IR}(S_n(D) + \mathrm{Lap}(\epsilon^{-1}))^n)\ \right].$$

To estimate the above expected error, we compute the expected normalized error on a few datasets of varying size n: (1) Multi-sets containing elements with the same value 0.5 (herein called repeating single-value dataset), (2) sets containing equally-spaced numbers $(i/(n-1))$ for $i = 0, \ldots, n-1$ (herein call equally-spaced dataset), (3) sets containing n randomly chosen elements from the Twitter location data [1], and (4) sets containing n randomly chosen elements from the Kaluža's data [12].

Fig. 4(a) shows the expected error $Err_{1,1,D}$ for the four datasets with different n. Each sample in the graph is the average over 500 runs. Observe that the error on equally-spaced data well approximates the errors on the two real-life dataset (Twitter location dataset and Kaluža's dataset). Hence, we take the error on the equally-spaced dataset as an approximation of the errors on other datasets. For abbreviation, let $Err_{\epsilon,n}$ denote the expected error $Err_{\epsilon,1,D}$ where D is the equally-spaced dataset with n points. Based on experiences on other datasets, we suspect that the expected error depends on the difference of the minimum and the maximum element in D, and the repeating single-value dataset is the extreme case whose error could be served as a lower bound as shown in Fig. 4(a).

Fig. 3(a) shows the expected error $Err_{\epsilon,1,D}$ for dataset on equally-spaced points for different ϵ and n, and Fig. 3(b) shows the ratios of error for different ϵ to $Err_{1,n}$. The results agree with the intuition that when ϵ is increased by a factor of c, the error would approximately decrease by factor of c, that is,

$$Err_{\epsilon,1,D} \approx \frac{1}{c} Err_{c\epsilon,1,D}. \tag{4}$$



(a) The normalized error $Err_{\epsilon,n}$.

(b) The ratio of $Err_{\epsilon,n}$ to $Err_{1,n}$.

Fig. 3. The normalized error for different security parameter ϵ on equally-spaced dataset, each sample is the average of 500 runs.

5.3 Effect on Generalization Noise

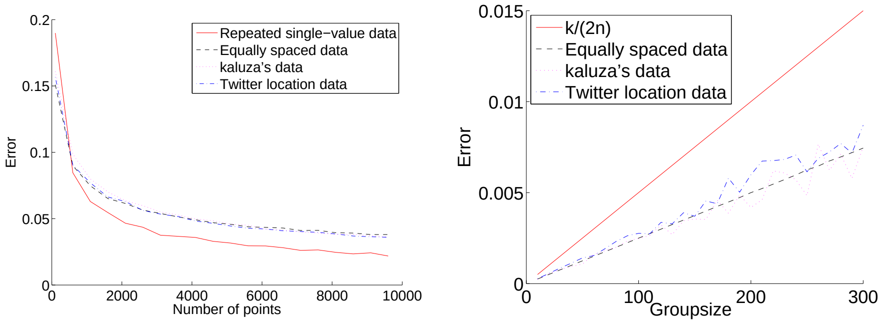
When $k > 1$, the grouping introduces a *generalization error*, which is incurred when all elements in a group are represented by their mean. Before giving formal description of generalization error, let us introduce some notations.

Given a sequence $D = \langle x_1, \ldots, x_n \rangle$ of n numbers, and a parameter k, where k divides n, let us call the following function *downsampling*:

$$\downarrow_k (D) = \langle s_1, \ldots, s_{(n/k)} \rangle,$$

where each s_i is the average of $x_{k(i-1)+1}, \ldots, x_{ik}$. Given a sequence $D' = \langle s'_1, \ldots, s'_m \rangle$ and k, let us call the following function *upsampling*,

$$\uparrow_k (D') = \langle x'_1, \ldots, x'_{mk} \rangle,$$



(a) Expected normalized error $Err_{1,1,D}$. (b) Normalized generalization error $Gen_{D,k}$.

Fig. 4. The expected normalized error and normalized generalization error with $\epsilon = 1$ on different dataset D.

where $x'_i = s'_{\lfloor (i-1)/k \rfloor + 1}$ for each i.

The *normalized generalization error* is defined as,

$$Gen_{D,k} = \frac{1}{n} \| D - \uparrow_k (\downarrow_k (D)) \|_1 .$$

It is easy to see that, for any k and D of size n, the normalized generalization error is at most $k/(2n)$. However, this bound is often an overestimate. Fig. 4(b) shows the generalization error of different group size a dataset containing $10,000$ equally-spaced values, a dataset containing $10,000$ numbers randomly drawn from the transformed Kaluža's dataset, and a dataset of $10,000$ numbers randomly drawn from the transformed Twitter location data.

Observe that, empirically, the generalization error can be well approximated by $\frac{k}{4n}$. To see that such approximation holds for a typical dataset, consider the following partition of the unit interval: $0 = p_0 < p_1 < p_2, \ldots, p_{(n/k)-1} < p_{n/k} = 1$. Let us consider a sorted sequence S of elements in dataset D, where the $jk+1, jk+2, \ldots (j+1)k$-th elements in S are uniformly independent and identically distributed over $[p_j, p_{j+1})$ for $j = 0, 1, \ldots, (n/k) - 1$. We can verify that the expected generalization error $Gen_{D,k} \approx \frac{k}{4n}$ with simulations. Hence, we approximate the generalization error by $\frac{k}{4n}$ and denote it as $Gen_{n,k}$.

5.4 Determining the group size k

Now, let us combine the components and build an error model of how k affects the accuracy. First, grouping reduces the number of constraints by a factor of k. As suggested by Fig. 4(a), when the number of constraints decreases, the error reduction from isotonic regression decreases. On the other hand, recall that the regression is performed on the published values divided by k (see the role of k in Step A3). This essentially reduces the level of Laplace noise by a factor of k. Hence, the accuracy attained by grouping k elements is "equivalent" to the accuracy attained without grouping but with the privacy parameter ϵ increased by a factor of k. These two components can be estimated in terms of $Err_{\epsilon,n}$ as follow:

$$Err_{\epsilon,k,D} \approx \frac{1}{k} Err_{\epsilon,n/k}.$$

For general k, the reconstructed dataset is

$$\widetilde{D} = \uparrow_k (\mathrm{IR}(\widetilde{S})),$$

where $\widetilde{S}$ is an instance of $\downarrow_k (S_n(D)) + \mathrm{Lap}(1)^{n/k}$. Now, we have,

$$\begin{aligned}
&\mathrm{EMD}\ (D, \widetilde{D}) = \|S_n(D) - \uparrow_k (\mathrm{IR}(\widetilde{S}))\|_1 \\
&= \|S_n(D) - \uparrow_k (\downarrow_k (S_n(D)) + \uparrow_k (\downarrow (S_n(D))) - \uparrow_k (\mathrm{IR}(\widetilde{S}))\|_1 \\
&\le n \cdot Gen_{D,k} + \| \uparrow_k (\downarrow_k (S_n(D))) - \uparrow_k (\mathrm{IR}(\widetilde{S}))\|_1 \\
&= n \cdot Gen_{D,k} + k \cdot \| \downarrow_k (S_n(D)) - \mathrm{IR}(\widetilde{S})\|_1 \\
&= n \cdot Gen_{D,k} + k \cdot \mathrm{EMD}(\downarrow_k (S_n(D)), \mathrm{IR}(\widetilde{S})).
\end{aligned} \tag{5}$$

Note that the first term $n \cdot Gen_{D,k}$ is a constant independent of the random choices made by the mechanism. Also note that the second term is the EMD between the down-sampled dataset and its reconstructed copy obtained using group size 1. Thus, by taking expectation over randomness of the mechanism, we have

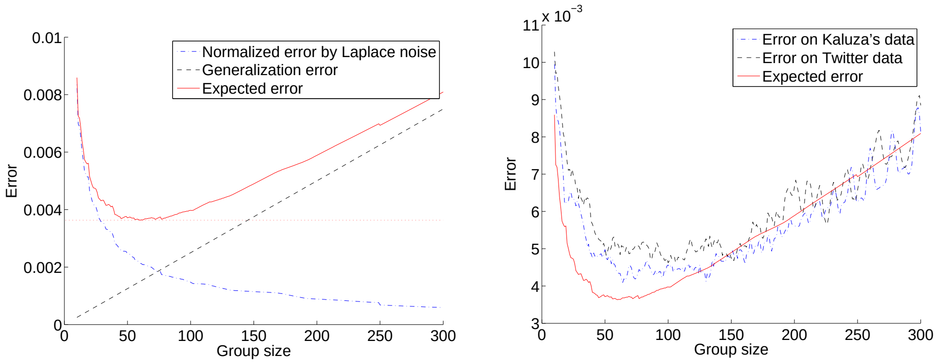
$$Err_{\epsilon,k,D} \le Gen_{D,k} + \frac{1}{k} Err_{\epsilon,1,\downarrow_k(D)}. \tag{6}$$

In other words, the expected normalized error is bounded by the sum of normalized generalization error, and the normalized error incurred by the Laplace noise. Fig. 5(a) shows the three values versus different group size k for equally-spaced data of size 10,000. The minimum of the expected normalized error suggests the optimal group size k.

Fig. 5(b) illustrates the expected errors for different k on the Twitter location data with 10,000 points. The red dotted line is $Err_{\epsilon,k,D}$ whereas the blue solid line is the sum in the right-hand-side of the inequality (6). Note that the differences between the two graphs are small. We have conducted experiments on other datasets and observed similar small differences. Hence, we take the sum as an approximation to the expected normalized error,

$$Err_{\epsilon,k,D} \approx Gen_{n,k} + \frac{1}{k} Err_{\epsilon,n/k}. \tag{7}$$

Now, we are ready to find the optimal k given ϵ and n. From Fig. 4(a) and Fig. 4(b) and the approximation given in equation (7), we can determine the best group size k when given the size of the database n and the security requirement ϵ. From the parameter ϵ, we can obtain the value $\frac{1}{k} Err_{n/k,e}$ for different k. From the database's size n, we can determine $Gen_{n,k}$ which is $\frac{k}{4n}$. Thus, we can approximate the normalized error $Err_{k,D}$ with equation (7) as illustrated in Fig. 5(a). Using the same approach, the best group size given different n and ϵ can be calculated and is presented in table 1.



(a) The expected error.

(b) Compare with the actual error.

Fig. 5. The expected error derived based on the equally-spaced dataset and the comparison with actual error on the Kaluža's dataset with $\epsilon = 1$.

Table 1. The best group size k given n and ϵ

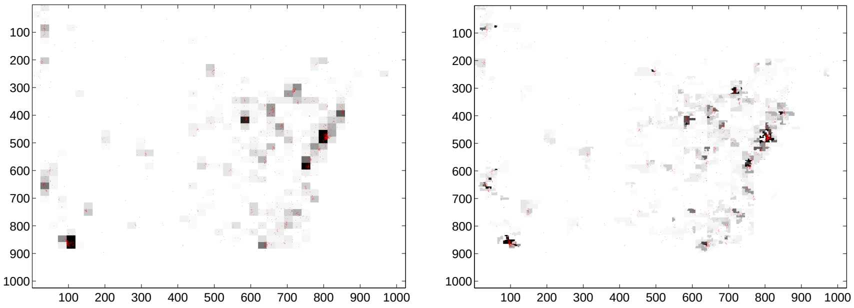
	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 3$
n= 2,000	44	29	20	12
n= 5,000	59	37	27	18
n= 10,000	79	51	36	27
n= 20,000	121	83	61	41
n= 100,000	234	150	98	73
n= 180,000	300	177	110	94

6 Comparisons

In this section, we compare the performance of the proposed mechanism with three known mechanisms w.r.t. different utility functions. We first compare the mechanism that outputs equi-width histograms. Next, we investigate the wavelet-based mechanism proposed by Xiao et al. [28] and measure accuracy of range queries. Lastly, we consider the problem of estimating median, and compare with a mechanism based on smooth sensitivity proposed by Nissim et al. [20]. We do not conduct experiments to compare with the k-d tree method [29] because it is designed for high dimensional data and it is not clear how to apply it to low dimension effectively. For comparison purposes, we empirically choose the best parameters for the known mechanisms, although this apriori information is not available to the publisher. We remark that the parameter k of our proposed mechanism is chosen from Table 1.

6.1 Equi-width Histogram

We want to compare the performance of our method with the equi-width histogram method. Fig. 6(a) shows a differentially private equi-width histogram. To



(a) Equi-width method.

(b) Proposed method.

Fig. 6. Visualization of the density functions, where the darker region corresponds to higher value. The superposing red dots are randomly selected from original data points for comparison purposes.

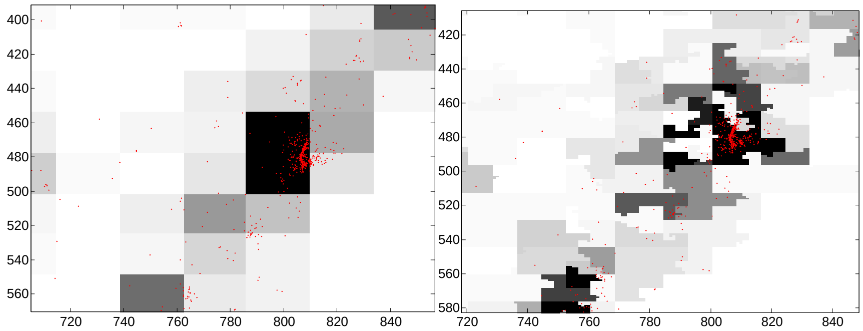
visualize the reconstructed points of our method as a histogram, we construct the bins in the following way: let B be the set of *distinct*-points in D, and we construct the Voronoi diagram of B. The cells in the Voronoi diagram are taken to be the bins of a histogram as depicted in Fig. 6(b).

To facilitate comparison, we treat the histograms as estimations of the underlying probability density function f, and use the statistical distance between density functions as a measure of utility. The value of $f(x)$ can be estimated by the ratio of the number of samples, over the width of the bin where x belongs to, with some normalizing constant factor.

In this section, we qualify the mechanism's utility by the distance between the two density functions: one that is derived from the original dataset, and the other that is derived from the mechanism's output.

Fig. 6(a) and 6(b) show the estimated density function from the Twitter's location dataset, by equi-width histogram mechanism and by our mechanism. For comparisons, 1% of the original points are plotted on top of the two reconstructed density functions. Fig. 7(a) and 7(b) show the zoom-in view of the dense region around New York City. Observe that the density function produced by our mechanism has "variable-sized" cells and thus is able to adaptively capture the fine details.

The statistical difference, measured with ℓ_1-norm and ℓ_2-norm, between the two estimated density functions derived from the original and the mechanism's output are shown in Table 2. We remark that it is not easy to determine the optimal bin-width for the equi-width histogram prior to publishing. Fig. 8 shows that the optimal bin-width differs significantly for three different datasets. For comparison purposes, we empirically choose the best parameters to the advantage of the compared algorithms, although such parameters could be dependent on the dataset.



(a) Zoom in view of Fig. 6(a).

(b) Zoom in view of Fig. 6(b).

Fig. 7. A more detailed view of the density functions.

6.2 Range Query

We consider the scenario where a dataset is to be published, and subsequently used to answer a series of range queries, where each range query asks for the total number of points within a query range. Publishing an equi-width histogram would not attain high accuracy if the size of the query ranges varies drastically. Intuitively, wavelet-based techniques [28] are natural solutions to address such multi-scales queries. However, there are many parameters, including the bin-widths at various scales and the amounts of privacy budget they consume, to be determined prior to publishing.

To apply the proposed method in this scenario, given a query, we obtain the number of points within the range from the estimated density function (as described in Section 6.1) by accumulating the probability over the query region and then multiplying by the total number of points.

We compare the range query results of the wavelet-based mechanism, the equi-width histogram mechanism and our mechanism on the 1D Twitter data, and on the 2D Twitter location dataset. To incorporate the knowledge of the database's size n, the total number of points is adjusted to n for the histogram mechanism and the DC component of the wavelet transform is set to be exactly n for the wavelet mechanism. For each range query, the absolute difference between the the true answer and the answer derived from the mechanism's output is taken as the error. We compare the results over different query range sizes and for each query range. For each range size s, 1,000 randomly chosen queries of size s are asked, and the corresponding errors are recorded. More precisely, the center of a 1D query range of size s is chosen uniformly at random in the continuous interval $[\frac{s}{2}, 1-\frac{s}{2}]$, whereas the center of a 2D query range of size s is chosen uniformly at random in the region $[\frac{s}{2}, 1-\frac{s}{2}] \times [\frac{s}{2}, 1-\frac{s}{2}]$.

To determine the parameters for the two compared mechanisms, we conduct experiments on a few selected values and choose the values to the advantage of the compared mechanisms. For the equi-width histogram, the only parameter is the number of bins (n_1). For the wavelet-based mechanism, the parameter we

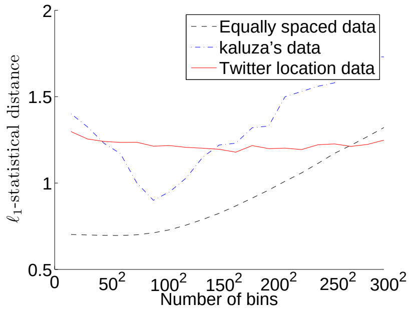
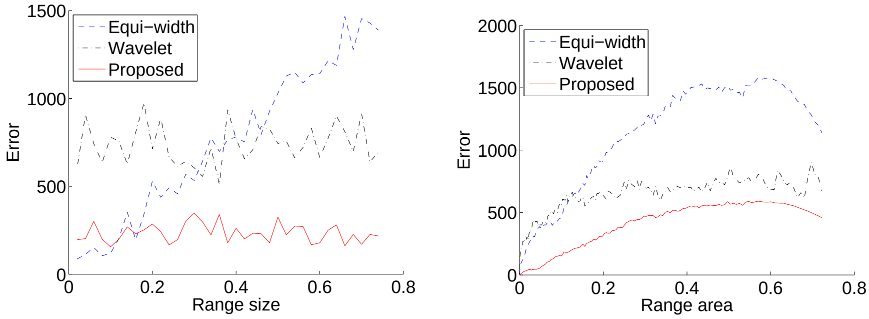


Fig. 8. The statistical differences versus bin-widths for different dataset D with $\epsilon = 1$ and $n = 60,000$.

	equi-width	proposed method
ℓ_1-norm	1.23	1.13
ℓ_2-norm	0.25	0.20

Table 2. The statistical differences of the two methods.



(a) 1D range query.

(b) 2D range query.

Fig. 9. Comparison of range query performance.

considered is the number of bins (n_2) of the histogram whereby wavelet transformation is performed on, that is, the number of bins in the "finest" histogram. From our experiments, we choose $n_1 = 1000$ and $n_2 = 1024$ for the 1D data, and $n_1 = 40 \times 40$ and $n_2 = 512 \times 512$ for the 2D data. The parameter k for our mechanism is looked up from Table 1. The choice of group size k according to Table 1 is 177 ($n = 180,000, \epsilon = 1$). The average errors of the range query is shown in Fig. 9(a) and 9(b).

Observe that our proposed method is less sensitive to the query range in the 1D case as expected because the accuracy of our range query results depend only on the boundary points, as opposed to the equi-width histogram method where errors are induced by each bins within the range. The wavelet-based mechanism outperforms the equi-width histogram mechanism in larger size range queries, but performs badly for small range due to the accumulation of noise.

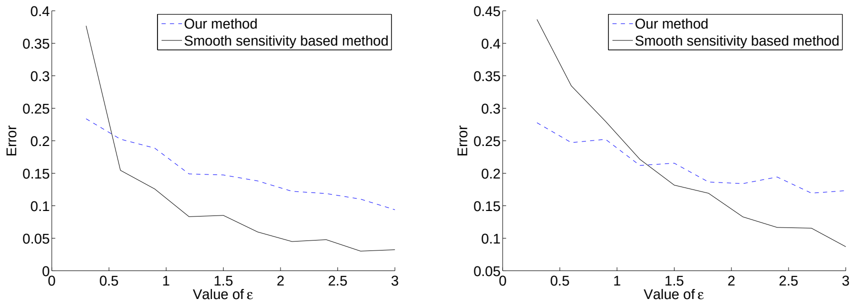
6.3 Median

Finding the median accurately in a differentially private manner is challenging due to the high "global sensitivity": there are two datasets that differ by one element but having a completely different median. Nevertheless, for many instances, their "local sensitivity" are small. Nissim et al. [20] showed that in general, by adding noise proportional to the "smooth sensitivity" of the database instance, instead of the global sensitivity, can also ensure differential privacy. They also gave an $\Theta(n^2)$ algorithm that find the smooth sensitivity w.r.t. median.

Our mechanism outputs the sorted sequence differentially privately, and thus naturally gives the median. Compare to the smooth sensitivity-based mechanism, our mechanism provides more information in the sense that it outputs the whole sorted sequence. Furthermore, our mechanism can be efficiently carried out in $O(n \log n)$ time.

We conduct experiments on synthetic datasets of size 129 to compare the accuracy of both mechanisms. The experiments are conducted for different local sensitivity and different ϵ values. To construct a dataset with a particular local sensitivity, 66 random numbers are generated with the exponential distribution and then scaled to the unit interval. The dataset contains the 66 random numbers and 63 ones. Fig. 10(a) and 10(b) shows the noise level with different ϵ on datasets that has a local sensitivity of 0.1 and 0.3.

When the local sensitivity of the median is high, our mechanism tends to provide a better result. In addition, our mechanism performs well under higher requirement of security: when the ϵ is smaller, the accuracy of our mechanism decreases slower than the smooth sensitivity-based method.



(a) Local sensitivity of 0.1.

(b) Local sensitivity of 0.3.

Fig. 10. The error of median versus different ϵ from two datasets.

7 Discussion and Future Work

7.1 Hybrid Method

The proposed mechanism can be viewed as the publishing of a "equi-depth" histogram, where the "depth" is the group size. Potentially, our proposed method and equi-width histogram could complement each other, by alternatively publishing one after another. For example, we can first apply an equi-width histogram to get a coarse distribution of the data, followed by our method for a "zoom-in" view. Alternatively, we can apply equi-width histogram after our method to "break" the stepping effect of isotonic regression.

7.2 Effect of Dimension

We rely on a locality-preserving mapping T to extend the mechanism to higher dimension. Although it is shown that the distant between two d dimensional points x, y is preserved and bounded by $c(\|T(x) - T(y)\|)^{1/d}$ (see Section 2.3), the curse-of-dimensionality is still in play in higher dimension. Firstly, to our best knowledge, there is no known efficient constructions for dimensions higher than 3. Secondly, the exponential factor $1/d$ amplifies the error: for example, Fig. 5 shows that our scheme can reduce the error of $\|T(x)-T(y)\|$ to 0.005, where y is the reconstructed point for x. When d is 2, we can have $\|x - y\|_2$ bounded by $0.07c$; when d is 3, the bound is increased to $0.17c$. We are unable to verify the performance in higher dimension due to the lack of efficient construction, and leave the accurate extension to higher dimensional data as future work.

8 Related Work

There are extensive works on privacy-preserving data publishing. The recent survey by Fung et al. [8] gives a comprehensive overview on various notions, for example, k-anonymity [26], ℓ-diversity [15], and the recently proposed concept of differential privacy [5].

Xiao et al. [28] proposed a mechanism of adding Laplace noise to the coefficients of a wavelet transformation of an equi-width histogram. The noisy wavelet coefficients are then published, from which range queries can be answered. Essentially, what being published is a series of equi-width histograms for different bin-widths where the noise added to the histograms of larger bin-width are smaller. A range query can then be decomposed and answered from the histograms series different scales.

Isotonic regression has been used to improve a differentially private query result. Hay et al. [11] proposed a method that employs isotonic regression to boost accuracy, but in a way different from our mechanism. They consider publishing *unattributed histogram*, which is the (unordered) multi-set of the frequencies of a histogram. As the frequencies are unattributed (i.e. order of appearance is irrelevant), they proposed publishing the sorted frequencies and later employing isotonic regression to improve accuracy.

Machanavajjhala et al. [16] proposed a 2D dataset publishing method that can handle the sparse data in 2D equi-width histogram. To mitigate the sparse data, their method shrinks the sparse blocks by examining publicly available data such as a previously release of similar data. They demonstrate this idea on the commuting patterns of the population of the United States, which is a real-life sparse 2D map in large domain. As their method partitions the space based on a previously released data, we consider the partition as pre-determined partition and is not adaptive to the publishing dataset.

The median is an important statistic, and a differentially private median finding process can be useful in many constructions, such as in pointset spatial decomposition [4, 23]. However, finding the median differentially privately is not easy due to the large global sensitivity. Nissim et al. [20] introduced the notion of smooth sensitivity and proposed an accurate mechanism with $\Theta(n^2)$ running time.

9 Conclusion

Our mechanism is very simple from the publisher's point of view. The publisher just has to sort the points, group consecutive values, add Laplace noise and publish the noisy data. There is also minimal tuning to be carried out by the publisher. The main design decision is the choice of the group size k, which can be determined using our proposed noise models, and the locality-preserving mapping for which the classic Hilbert curve suffices to attain high accuracy. Through empirical studies, we have shown that the published raw data contain rich information for the public to harvest, and provide high accuracy even for usages like median-finding and range-searching that our mechanism is not initially designed for.

References

1. Twitter census: Twitter users by location. `http://www.infochimps.com/datasets/twitter-census-twitter-users-by-location`.
2. B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *symposium on principles of database systems*, pages 273–282, 2007.
3. A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. pages 128–138, 2005.
4. G. Cormode, M. Procopiuc, E. Shen, D. Srivastava, and T. Yu. Differentially private spatial decompositions. *To be appeared in ICDE*, 2012.
5. C. Dwork. Differential privacy. *Automata, languages and programming*, page 1, 2006.
6. C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, pages 265–284, 2006.
7. D. Feldman, A. Fiat, H. Kaplan, and K. Nissim. Private coresets. page 361, 2009.
8. B. Fung, K. Wang, R. Chen, and P. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, page 14, 2010.

9. C. Gotsman and M. Lindenbaum. On the metric properties of discrete space-filling curves. *IEEE Transactions on Image Processing*, pages 794–797, 1996.
10. S. Grotzinger and C. Witzgall. Projections onto order simplexes. *Applied mathematics and optimization*, pages 247–270, 1984.
11. M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *VLDB Endowment*, page 1021, 2010.
12. B. Kaluža, V. Mirchevska, E. Dovgan, M. Luštrek, and M. Gams. An agent-based approach to care in independent living. *Ambient Intelligence*, pages 177–186, 2010.
13. D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Management of data*, pages 193–204, 2011.
14. C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. pages 123–134, 2010.
15. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *International Conference on Data Engineering*, pages 24–24, 2006.
16. A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *International Conference on Data Engineering*, pages 277–286, 2008.
17. M. C. Meyer. Inference using shape-restricted regression splines. *Annals of Applied Statistics*, pages 1013–1033, 2008.
18. G. Mitchison and R. Durbin. Optimal numberings of an n x n array. *Algebraic Discrete Methods.*, pages 571–582, 1986.
19. R. Niedermeier, K. Reinhardt, and P. Sanders. Towards optimal locality in mesh-indexings. pages 364–375, 1997.
20. K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. pages 75–84, 2007.
21. G. Piatetsky-Shapiro and C. Connell. Accurate estimation of the number of tuples satisfying a condition. pages 256–276, 1984.
22. V. Poosala, P. Haas, Y. Ioannidis, and E. Shekita. Improved histograms for selectivity estimation of range predicates. *ACM SIGMOD Record*, page 294, 1996.
23. W. Qardaji and N. Li. Recursive partitioning and summarization: a practical framework for differentially private data publishing. *To be appeared in ASIACCS*, 2012.
24. Y. Rubner, L. Guibas, and C. Tomasi. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. pages 661–668, 1997.
25. Q. F. Stout. Optimal algorithms for unimodal regression. *Computer Science and Statistics*, pages 109–122, 2000.
26. L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, pages 557–570, 2002.
27. X. Wang and F. Li. Isotonic smoothing spline regression. *Journal Computational and Graphical Statistics*, pages 21–37, 2008.
28. X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on Knowledge and Data Engineering*, pages 1200–1214, 2010.
29. Y. Xiao, L. Xiong, and C. Yuan. Differentially private data release through multidimensional partitioning. *Secure Data Management*, pages 150–168, 2011.