Data Anonymization that Leads to the Most Accurate Estimates of Statistical Characteristics

Gang Xiang
Applied Biomathematics
100 North Country Rd.
Setauket, NY 11733, USA
Email: gxiang@sigmaxi.net

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
Email: vladik@utep.edu

***Abstract*—To preserve privacy, we divide the data space into boxes, and instead of original data points, only store the corresponding boxes. In accordance with the current practice, the desired level of privacy is established by having at least k different records in each box, for a given value k (the larger the value k, the higher the privacy level).**

When we process the data, then the use of boxes instead of the original exact values leads to uncertainty. In this paper, we find the (asymptotically) optimal subdivision of data into boxes, a subdivision that provides, for a given statistical characteristic like variance, covariance, or correlation, the smallest uncertainty within the given level of privacy.

In areas where the empirical data density is small, boxes containing k points are large in size, which results in large uncertainty. To avoid this, we propose, when computing the corresponding characteristic, to only use data from boxes with a sufficiently large density. This deletion of data points increases the statistical uncertainty, but decreases the uncertainty caused by introducing the privacy-related boxes. We explain how to compute an optimal threshold for which the overall uncertainty is the smallest.

I. Formulation of the Problem

Need to preserve privacy. One of the main objectives of engineering is to help people: civil engineering designs houses in which we live and roads along which we travel, electrical engineering designs appliances – and designs electric networks that make it possible to use these appliances, etc. In all these applications, it is important to know as much as possible about the potential customers and their preferences. For example, it is desirable to know the customer's age, income level, etc., so that by gathering statistics about preferences of different customers we will be able to tailor engineering designs to these customers.

Many customers are reluctant to share too much of this information, since they are concerned that this detailed information can be potentially used against them. For example, information about the customer's eating habits – which, for most people, are not always maximally healthy – can be used by insurance companies to raise the individual insurance rates. Even such seemingly innocent information as age, if leaked, can be used by unscrupulous companies to unlawfully discriminate against older job applicants.

How to preserve privacy: general idea. To avoid such concerns, we should be able to distort the original information, so that individual data is no longer available, but it is still possible to make statistical conclusions about the whole population sample. One way to do this is as follows:

- instead of storing, for each individual, the values $x = (x_1, \ldots, x_n)$ of all the corresponding numerical characteristics,
- we divide the n-dimensional space $\mathbb{R}^n$ into *boxes* $[\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_n, \overline{x}_n]$, so that for each individual, instead of storing the exact vector x, we only store the label of the box that contains this vector x.

In this case, even if an undesirable agent gets access of the data, this agent will not get full information about the person, only ranges of values $[\underline{x}_i, \overline{x}_i]$ for each of the corresponding quantities x_i.

The notion of k-anonymity. An additional problem with individual records is that if we store the exact data, then, even if we do not store the names, it may still be possible to find out who these records refer to. For example, if we store the exact birthdate and the name of the county where a person lives, then in many cases, based on this information, we will able to uniquely identify the person (see, e.g., [4]) and thus, use the presumably anonymized database to learn all about this person's income, habits, etc. To avoid such a situation, it is desirable to select boxes in such a way that each box contains records of at least two different people.

If we only have records of two people in each box, then an agent who happens to have some information about one of these two people will thus get a lot of information about the second person as well. To avoid such situations, it is therefore reasonable to select an integer $k > 2$, and to make sure that in each box, there are at least k different records – corresponding to at least k different customers. This idea is known as k-*anonymity*; see, e.g., [4].

The corresponding integer k can be viewed as characterizing the privacy level: the larger the value k, the higher the privacy level.

The notion of ℓ-diversity. The notion of k-anonymity does not help much if it turns out that all the persons whose records are contained in a box have the exact same value of a certain characteristic. In such as situation, if we learn that a person's record belongs to the corresponding box, it does not matter

that there may be hundreds of record in this box: for example, if this box corresponds to a shift in which all the workers receive the same salary, then discovering the salary of one of these folks will enable the malicious agent to learn the salaries of all of them.

To avoid this situation, it is reasonable to require that within each box, there should be at least ℓ different values of each parameter i. This requirement is known as *ℓ-diversity*; see, e.g., [1].

Statistical data processing. One of the main objectives of storing the data is to be able to perform statistical processing on this data. For example, when we are designing a transportation system, it is desirable to know the driving habits of the city's inhabitants. Thus, in addition to information about driving habits, we gather all possible numerical characteristics of the inhabitants, so that we will be able to check which of these characteristics are important, i.e., which are correlated with the driving characteristics. For that, it is important to know the *correlation* between different statistical characteristics.

The correlation ρ_{ij} between the two quantities x_i and x_j is usually estimated as follows (see, e.g., [3]):

$$\rho_{ij} = \frac{C_{ij}}{\sigma_i \cdot \sigma_j},$$

where the correlation C_{ij} is estimated as

$$C_{ij} = \frac{1}{N} \cdot \sum_{p=1}^{N} \left(x_i^{(p)} - E_i\right) \cdot \left(x_j^{(p)} - E_j\right),$$

the means E_i and E_j are estimated as

$$E_i = \frac{1}{N} \cdot \sum_{p=1}^{N} x_i^{(p)}, \quad E_j = \frac{1}{N} \cdot \sum_{p=1}^{N} x_j^{(p)},$$

each standard deviation σ_i is estimated as $\sigma_i = \sqrt{V_i}$, and the variance V_i is estimated as

$$V_i = \frac{1}{N} \cdot \sum_{p=1}^{N} \left(x_i^{(p)} - E_i\right)^2.$$

Once we figure out which characteristics $x_{i_1}, \ldots, x_{i_m}$ are important, we would like to be able to predict – as accurately as possible – the desired characteristic x_{i_0}. In most cases, a linear regression is used for such a prediction. In this case, our objective is to find the coefficients c_q of the corresponding linear dependence formula

$$x_{i_0} \approx c_0 + \sum_{q=1}^{m} c_q \cdot x_{i_q}$$

is the most accurate. To describe the accuracy of this approximation, for each of N persons p, we can calculate the approximation error

$$e_p \stackrel{\text{def}}{=} c_0 + \sum_{q=1}^{m} c_q \cdot x_{i_q}^{(p)} - x_{i_0}^{(p)}.$$

We want the resulting N-dimensional vector $e = (e_1, \ldots, e_N)$ to be as close to 0 as possible, i.e., we want to minimize the distance between this vector and the ideal (no approximation error) vector $(0, \ldots, 0)$. By Pythagoras Theorem, this distance is equal to the square root of the sum $\sum\limits_{p=1}^{N} e_p^2$, so minimizing this distance is equivalent to minimizing this sum

$$\sum_{p=1}^{N} e_p^2 = \sum_{p=1}^{N} \left(c_0 + \sum_{q=1}^{m} c_q \cdot x_{i_q}^{(p)} - x_{i_0}^{(p)}\right)^2.$$

Differentiating this sum with respect to each of the unknown c_0, $c_1, \ldots, c_m$ and equating this derivative to 0 leads to the known Least-Squares system of linear equations:

$$c_0 + \sum_{q=1}^{N} c_q \cdot E_{i_q} - E_{i_0} = 0;$$

$$c_0 \cdot E_{i_q} + \sum_{r=1}^{m} c_r \cdot \left(\frac{1}{N} \cdot \sum_{p=1}^{N} x_{i_q}^{(p)} \cdot x_{i_r}^{(p)}\right) -$$

$$\frac{1}{N} \cdot \sum_{p=1}^{N} x_{i_0}^{(p)} \cdot x_{i_q}^{(p)} = 0.$$

We can simplify this system if we subtract, from the q-the equation, the 0-th equation multiplied by E_{i_q} and take into account that

$$C_{i_q i_r} = \frac{1}{N} \cdot \sum_{p=1}^{N} x_{i_q}^{(p)} \cdot x_{i_r}^{(p)} - E_{i_q} \cdot E_{i_r};$$

then, the system takes the following simplified form:

$$c_0 + \sum_{q=1}^{N} c_q \cdot E_{i_q} - E_{i_0} = 0;$$

$$\sum_{r=1}^{m} c_r \cdot C_{i_q i_r} - C_{i_0 i_q} = 0.$$

To solve this system, we can first find the coefficients $c_1, \ldots, c_m$ from the linear system

$$\sum_{r=1}^{m} c_r \cdot C_{i_q i_r} = C_{i_0 i_q},$$

and then compute the remaining coefficient c_0 as

$$c_0 = E_{i_0} - \sum_{q=1}^{N} c_q \cdot E_{i_q}.$$

In all these tasks, we need to estimate such statistical characteristics as the averages E_i, variances V_i, covariances C_{ij}, and correlations ρ_{ij}.

In statistical data processing, privacy leads to uncertainty. To maintain privacy, we replace each exact vector $x^{(p)}$ with a box. In other words, we replace each numerical value $x_i^{(p)}$ with the corresponding interval. If we combine different values from these intervals, then, in general, we get different values

of the resulting statistical characteristics. Hence, for each of these characteristics, instead of a single value, we have a whole interval of possible values.

Formulation of the problem. If this interval is too wide, the resulting range is useless. For example, since the correlation is always between -1 and 1, we do not learn anything new if it turns out that the range of possible values of the correlation is the interval $[-1,1]$. It is therefore desirable to select the boxes in such a way that the corresponding intervals are as narrow as possible.

In other words, among all possible subdivisions into boxes which preserve k-anonymity (and ℓ-diversity), we need to select the one which leads to the narrower intervals for the desired statistical characteristic.

What we do in this paper. In this paper, we find the (asymptotically) optimal subdivision of data into boxes, a subdivision that provides, for a given statistical characteristic like variance, covariance, or correlation, the smallest uncertainty within the given level of privacy. In Section 2, we describe this optimal subdivision for the case when we only require k-anonymity; in Section 4, we also take into account the requirement of ℓ-diversity.

II. Optimal Privacy-Enhancing Subdivision into Boxes: Case of k-Anonymity

First conclusion: (almost) every box should contain exactly k records. To minimize the uncertainty caused by the boxes, we must make these boxes as narrow as possible (as long as the privacy requirement is satisfied). As a result, each box should contain exactly k records – because if we have boxes with more than k records, we can rearrange them into smaller boxes.

Comment. Of course, if the total number of records N cannot be divided by k, we must have at least one box with $> k$ records. However, when the number N of records is high, the overwhelming majority of boxes has exactly k records in them. So, from the asymptotic viewpoint, we can safely assume that all boxes have this property.

Notations. For each box $[\underline{x}_1,\overline{x}_1]\times\ldots\times[\underline{x}_n,\overline{x}_n]$ and for each quantity i, let us denote the midpoint of the corresponding interval $[\underline{x}_i,\overline{x}_i]$ by $\widetilde{x}_i \stackrel{\text{def}}{=} \dfrac{\underline{x}_i+\overline{x}_i}{2}$, and this interval's half-width by $\Delta_i \stackrel{\text{def}}{=} \dfrac{\overline{x}_i-\underline{x}_i}{2}$. In these terms, the original interval takes the form $[\widetilde{x}_i-\Delta_i,\widetilde{x}_i+\Delta_i]$.

In these terms, each point $x=(x_1,\ldots,x_n)$ from this box can be represented as $x_i=\widetilde{x}_i+\Delta x_i$, where the difference $\Delta x_i \stackrel{\text{def}}{=} x_i-\widetilde{x}_i$ satisfies the inequality $|\Delta x_i|\le\Delta_i$. Vice versa, when we have n values Δx_i which satisfy this inequality, then the point x with coordinates $x_i=\widetilde{x}_i+\Delta x_i$ belongs to our box.

When we compute the value of the desired statistical characteristic

$$C(x_1^{(1)},\ldots,x_n^{(1)},\ldots,x_1^{(N)},\ldots,x_n^{(N)}),$$

then all we know about each of the values $x_i^{(p)}$ is that this value belongs to the corresponding interval, i.e., has the form

$$C(\widetilde{x}_1^{(1)}+\Delta x_1^{(1)},\ldots,\widetilde{x}_n^{(1)}+\Delta x_n^{(1)},\ldots,\widetilde{x}_n^{(N)}+\Delta x_n^{(N)}),$$

where $\left|\Delta x_i^{(p)}\right|\le\Delta_i^{(p)}$; here, the index p indicates the midpoints $\widetilde{x}_i^{(p)}$ and the half-widths $\Delta_i^{(p)}$ of the box that contains the p-th record.

We have a sufficient amount of data: a natural assumption and its consequences. When the number of records N is high, the records are close enough. So, if we combine k closest records into a single box, we expect that the corresponding values of the parameters x_i are close to each other.

Under this assumption, the differences Δ_i are reasonably small and thus, if we expand the dependence of C on the differences $\Delta_i^{(p)}$ in Taylor series, we can safely ignore terms which are quadratic and of higher order in terms of these differences and only keep linear terms. As a result, we get the following expression:

$$C=\widetilde{C}+\sum_{p=1}^{N}\sum_{i=1}^{n}\frac{\partial C}{\partial x_i}\cdot\Delta x_i^{(p)},$$

where

$$\widetilde{C}\stackrel{\text{def}}{=}C(\widetilde{x}_1^{(1)},\ldots,\widetilde{x}_n^{(1)},\ldots,\widetilde{x}_1^{(N)},\ldots,\widetilde{x}_n^{(N)}),$$

and the partial derivatives are taken at the point

$$\widetilde{x}\stackrel{\text{def}}{=}(\widetilde{x}_1^{(1)},\ldots,\widetilde{x}_n^{(1)},\ldots,\widetilde{x}_1^{(N)},\ldots,\widetilde{x}_n^{(N)}).$$

Thus, to find the range of possible values of this characteristic, we must finding the maximum and minimum of this linear expression when each of the variables takes values from the interval $\left[-\Delta_i^{(p)},\Delta_i^{(p)}\right]$.

In general, a linear function $A+c_i\cdot\Delta x_i$ on the interval $\Delta x_i\in[-\Delta_i,\Delta_i]$, (where A denotes the sum of all the terms which do not depend on Δx_i) is either increasing (when $c_i\ge 0$) or decreasing (when $c_i\le 0$):

- If the function is increasing, then it attains its maximum at the largest possible value $\Delta x_i=\Delta_i$, and its largest value is thus equal to $A+c_i\cdot\Delta_i$.
- If the function is decreasing, then it attains its maximum at the smallest possible value $\Delta x_i=-\Delta_i$, and its largest value is thus equal to $A-c_i\cdot\Delta_i$.

In both cases, the largest possible value of the term $c_i\cdot\Delta x_i$ is $|c_i|\cdot\Delta_i$. Similarly, the smallest possible value of this term is $-|c_i|\cdot\Delta_i$.

Thus, the largest possible value of the above linear expression is $\widetilde{C}+\Delta$, where

$$\Delta\stackrel{\text{def}}{=}\sum_{p=1}^{N}\sum_{i=1}^{n}\left|\frac{\partial C}{\partial x_i}\right|\cdot\Delta_i^{(p)},$$

and the smallest possible value of this linear expression is $\widetilde{C}-\Delta$. Hence, the width of the resulting interval is 2Δ, so minimizing this width is equivalent to minimizing Δ.

The expression Δ is the sum of terms corresponding to different points x. For k points within a box, the sum of their contribution to Δ is equal to

$$k \cdot \sum_{i=1}^{n} \left| \frac{\partial C}{\partial x_i} \right| \cdot \Delta_i,$$

where Δ_i are half-widths of this box.

Expressions for the corresponding partial derivatives. The estimate for the accuracy Δ is described in terms of partial derivatives $\dfrac{\partial C}{\partial x_i}$ of the statistical characteristic C. We have already listed all the statistical characteristics in which we are interested. Let us describe the explicit expressions for the partial derivatives for these characteristics.

For the mean $E_i = \dfrac{1}{N} \cdot \sum_{p=1}^{N} x_i^{(p)}$, the derivative is equal to

$$\frac{\partial E_i}{\partial x_i} = \frac{1}{N}.$$

For the variance $V_i = \dfrac{1}{N} \cdot \sum_{p=1}^{N} \left(x_i^{(p)}\right)^2 - E_i^2$, the derivative is equal to

$$\frac{\partial V_i}{\partial x_i} = \frac{1}{N} \cdot (2 \cdot x_i) - 2E \cdot \frac{\partial E_i}{\partial x_i} = \frac{2 \cdot (x_i - E_i)}{N}.$$

Therefore, for $\sigma_i = \sqrt{V_i}$, we get

$$\frac{\partial \sigma_i}{\partial x_i} = \frac{\partial \sqrt{V_i}}{\partial x_i} = \frac{1}{2} \cdot \frac{1}{\sqrt{V_i}} \cdot \frac{\partial V_i}{\partial x_i} = \frac{x_i - E_x}{\sigma_x}.$$

For the covariance $C_{ij} = \dfrac{1}{N} \cdot \sum_{p=1}^{N} x_i^{(p)} \cdot x_j^{(p)} - E_i \cdot E_j$, the derivative is equal to

$$\frac{\partial C_{ij}}{\partial x_i} = \frac{1}{N} \cdot x_j - \frac{\partial E_i}{\partial x_i} \cdot E_j = \frac{x_j - E_j}{N}.$$

Based on these formulas, we can find the expression for the derivatives of the correlation $\rho_{ij} = \dfrac{C_{ij}}{\sigma_i \cdot \sigma_j}$:

$$\frac{\partial \rho_{ij}}{\partial x_i} = \frac{1}{N} \cdot \frac{(x_j - E_j) - \dfrac{C_{ij}}{\sigma_i^2} \cdot (x_i - E_i)}{\sigma_i \cdot \sigma_j}.$$

Towards an optimal subdivision into boxes. The overall expression for Δ is a sum of terms corresponding to different points. So, to minimize Δ, we must, for each point, minimize the corresponding term $\sum_{i=1}^{n} a_i \cdot \Delta_i$, where $a_i \stackrel{\text{def}}{=} \left| \dfrac{\partial C}{\partial x_i} \right|$. The only constraint on the values Δ_i is that the corresponding box should contain exactly k different points. The number of points can be obtained by multiplying the data density $\rho(x)$ (which can be estimated based on the data) by the volume of the box. Each side of this box has width $2\Delta_i$, so the volume is equal to the product of these sides, i.e., to $2^n \cdot \prod_{i=1}^{n} \Delta_i$.

Thus, we arrive at the following optimization problem: find the values $\Delta_1, \ldots, \Delta_n$ that minimize the sum $\sum_{i=1}^{n} a_i \cdot \Delta_i$ under the constraint $\rho(x) \cdot 2^n \cdot \prod_{i=1}^{n} \Delta_i = k$.

Applying the Lagrange multiplier method to this constraint optimization problem, we get an equivalent unconstraint optimization problem of maximizing the expression

$$\sum_{i=1}^{n} a_i \cdot \Delta_i + \lambda \cdot \left(\rho(x) \cdot 2^n \cdot \prod_{i=1}^{n} \Delta_i - k \right).$$

Differentiating this expression with respect to Δ_i and equating the derivative to 0, we conclude that

$$a_i + \lambda \cdot \rho(x) \cdot 2^n \cdot \prod_{j \neq i} \cdot \Delta_j = 0.$$

Multiplying both sides of this equality by Δ_i, we get

$$a_i \cdot \Delta_i + \lambda \cdot \rho(x) \cdot 2^n \cdot \prod_{j=1}^{n} \Delta_j = 0.$$

Therefore, $\Delta_i = \dfrac{c}{a_i}$, where

$$c \stackrel{\text{def}}{=} -\lambda \cdot \rho(x) \cdot 2^n \cdot \prod_{j=1}^{n} \Delta_j.$$

The constant c depends on the Lagrange multiplier λ whose value is not known a priori. As usual in the Lagrange multiplier method, the value of this constant can be found from the constraint

$$\rho(x) \cdot 2^n \cdot \prod_{j=1}^{n} \Delta_j = k;$$

substituting the expression $\Delta_i = \dfrac{c}{a_i}$ into this constraint, we conclude that

$$\rho(x) \cdot 2^n \cdot \frac{c^n}{\prod_{j=1}^{n} a_j} = k,$$

hence

$$c^n = \frac{1}{2^n} \cdot \frac{k}{\rho(x)} \cdot \prod_{j=1}^{n} a_j,$$

and

$$c = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)} \cdot \prod_{i=1}^{n} a_i}.$$

For this c, the above expression $\Delta_i = \dfrac{c}{a_i}$ leads to the following expression for selecting an (asymptotically) optimal box:

Main result of this section: the expression for the (asymptotically) optimal subdivision into boxes. Around each point x, we need to select the box with half-widths

$$\Delta_i = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod_{j=1}^{n} a_j}}{a_i},$$

where $a_i = \left|\dfrac{\partial C}{\partial x_i}\right|$.

The resulting accuracy. Since the optimal value of Δ_i is equal to $\dfrac{c}{a_i}$, the resulting contribution $\sum\limits_{i=1}^{n} a_i \cdot \Delta_i$ to accuracy Δ has the form $n \cdot c(x)$, i.e., the form

$$n \cdot \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)} \cdot \prod_{i=1}^{n} a_i(x)}.$$

Thus, the overall uncertainty is equal to

$$\Delta = n \cdot \sum_x c(x) = n \cdot \sum_x \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)} \cdot \prod_{i=1}^{n} a_i(x)},$$

where the sum is taken over all N data points x.

III. To Improve Accuracy, We Need to Dismiss Rare Points

Formulation of the problem. In many practical situations, we have outlier points. For such outlier points, the smallest box which contains k of them may be huge, and this big-size box will contribute a large amount of uncertainty to Δ. To be more precise, according to our formula, the contribution is decreasing with the data density $\rho(x)$, so when this density tends to 0, the contribution of the corresponding points tends to infinity.

Idea about how to solve this problem. To avoid this problem, we propose to only use data from boxes with a sufficiently large density when computing the corresponding characteristic. If we select a subset $S \subset \{1, 2, \ldots, N\}$ of the set of N original points, then the resulting privacy-related uncertainty reduces to $n \cdot \sum\limits_{x \in S} c(x)$.

Towards a formal implementation of this idea. On the one hand, when we delete rare points, we decreases the uncertainty caused by introducing the privacy-related boxes. On the other hand, when we delete some data points, the resulting statistical estimates become based on a smaller sample and thus, less accurate.

In other words, if we dismiss too few points (e.g., none at all), the resulting uncertainty in estimating the desired statistical characteristic C is too high – due to the introduction of privacy-related boxes. If we dismiss too many points (e.g., leave only k points), then the resulting uncertainty in estimating C is still too high – this time due to inaccuracy of a statistical estimate based on a small sample.

It is therefore necessary to find an optimal dismissal for which the overall uncertainty is the smallest.

Formulation of the problem in precise terms. For statistical estimates, the accuracy of their estimation based on a sample of size $M = \#(S)$ is inverse proportional to $\sqrt{M}$; see, e.g., [3]. This fact is known for estimating mean, where the standard deviation of the estimate is equal to $\dfrac{\sigma}{\sqrt{M}}$; a similar asymptotic dependence on M holds for all the above characteristics. In general, the corresponding accuracy is equal to $\dfrac{A}{\sqrt{M}}$ for an appropriate constant M.

Thus, the above optimization problem takes means that we must select a subset $S \subseteq \{1, 2, \ldots, N\}$ in the set of the original points for which the overall estimation error

$$n \cdot \sum_{x \in S} c(x) + \frac{A}{\sqrt{\#(S)}}$$

attains the smallest possible value.

Analysis of the problem. Our main idea was to dismiss points x for which the value $c(x)$ – describing this point's contribution to overall uncertainty – is too large. From this viewpoint, a reasonable idea is to select a threshold c_0 and to dismiss all the points for which $c(x) > c_0$. Let us prove that this idea indeed leads to the optimal solution. Indeed, let S be an optimal set, and let c_0 be the largest value of $c(x)$ for all points from this set. By definition, this means that all the points with $c(x) > c_0$ are dismissed from the set S. So, to complete our proof, we need to show that all the points with $c(x) < c_0$ are included in the optimal set S.

We will prove this by contradiction. Let us assume that a point x^- with $c(x^-) < c_0$ is not in the set S, and let us derive a contradiction from this assumption. Indeed, by definition of the value c_0, there exists a point $x^+ \in S$ for which $c(x^+) = c_0$. Then, if we swap x^- and x^+, i.e., replace x^+ with x^- in the set S, we keep the number of S-points intact and thus, we keep the statistical term $\dfrac{A}{\sqrt{\#(S)}}$ in the formula for the total accuracy intact. On the other hand, in the first sum $\sum\limits_{x \in S} c(x)$, we replace a larger term $c(x^+) = c_0$ with a smaller term $c(x^-) < c_0$. Thus, after this swap, the sum $\sum\limits_{x \in S} c(x)$ decreases and hence, the overall estimation error decreases. This decrease contradicts to the fact that S is the optimal set, i.e., the set for which the overall error is the smallest possible. This contradiction shows that the optimal set S can indeed be obtained by selecting an appropriate threshold c_0.

Thus, we arrive to the following solution.

Which points to dismiss: the resulting optimal solution. For each point x, we should estimate

$$c(x) = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)} \cdot \prod_{i=1}^{n} a_i(x)},$$

where $a_i = \left|\dfrac{\partial C}{\partial x_i}\right|$. Then, we should find a value c_0 for which the sum

$$n \cdot \sum_{x : c(x) \le c_0} c(x) + \frac{A}{\sqrt{\#\{x : c(x) \le c_0\}}}$$

attains the smallest possible value. We then dismiss all the points for which $c(x) < c_0$, and estimate the desired characteristic based only on the remaining points.

Examples. For estimating the mean E_i, we have $a_i = \text{const}$ and thus, $c(x) = \text{const} \cdot \dfrac{1}{\sqrt[n]{\rho(x)}}$. In this case, $c(x)$ is a decreasing function of density, so dismissing all the points with sufficiently large $c(x)$ is equivalent to dismissing all the points for which density $\rho(x)$ is blow a certain threshold.

For computing covariance C_{ij}, the derivative a_i is proportional to $x_i - E_i$, so the upper threshold c_0 on $c(x)$ is equivalent to the lower threshold on the ratio $\dfrac{\rho(x)}{|x_i - E_i| \cdot |x_j - E_j|}$. In other words, points x at which the density $\rho(x)$ is small may be OK – if one of the values x_i or x_j is close to the corresponding mean. This enables us to add more points than if we simply went by a lower bound on density and thus, get a slightly better accuracy.

IV. Optimal Privacy-Enhancing Subdivision into Boxes: Case of k-Anonymity and ℓ-Diversity

How to take into account ℓ-diversity. In the previous sections, we only took into account the k-anonymity requirement. We also need to take into account the ℓ-diversity requirement, i.e., the requirement that within each box, for each variable i, there are at least ℓ different values of this variable.

To formalize this requirement, we first need to describe what "different" means. Usually, very small differences between the two values do not count as difference. So, for each variable i, there should be some threshold ε_i such that if the difference between the two values is at least ε_i, this means that the corresponding two values are indeed different.

In our analysis, we have assumed that the data is randomly distributed according to some probability density $\rho(x)$. As we have mentioned, boxes are small. So, within each box, the density $\rho(x)$ practically does not change, so within each box, we have, in effect, a uniform distribution. This means, in particular, that within the box, the values of the i-th variable x_i are uniformly distributed; thus, these values uniformly fill the side $[\underline{x}_i, \overline{x}_i]$ of width $2\Delta_i$. Thus, we must require that $2\Delta_i \ge \ell \cdot \varepsilon_i$.

With this additional requirement, we arrive at the following modified formulation of the optimization problem.

Formal description of the corresponding optimization problem. We need to find the values $\Delta_1, \ldots, \Delta_n$ for which the sum $\sum\limits_{i=1}^{n} a_i \cdot \Delta_i$ is the smallest possible under the constraints $\prod\limits_{i=1}^{n} \Delta_i \ge \dfrac{k}{2^n \cdot \rho(x)}$ and $2\Delta_i \ge \ell \cdot \varepsilon_i$ for all i.

Analysis of the problem. If these additional constraints are automatically satisfied for the above solution Δ_i to the k-anonymity problem, i.e., if the constraints $2\Delta_i \ge \ell \cdot \varepsilon_i$ are satisfied for all i, then we simply take these optimal values Δ_i as desired half-widths.

Let us analyze the problem for the case when the original optimal solution does not satisfy at least one of these constraints. For each i, we have either $2\Delta_i = \ell \cdot \varepsilon_i$ or $2\Delta_i > \ell \cdot \varepsilon_i$.

Let us show that if $\varepsilon_{i'} \cdot a_{i'} < \varepsilon_{i''} \cdot a_{i''}$, then in the optimal arrangement of the values Δ_i, we cannot have $2\Delta_{i'} = \ell \cdot \varepsilon_{i'}$ and $2\Delta_{i''} > \ell \cdot \varepsilon_{i''}$. Let us prove this by contradiction. Let us assume that in the optimal solution, there exist such indices i' and i''. Then, let us slightly modify the values $\Delta_{i'}$ and $\Delta_{i''}$, to $\Delta'_{i'} = \Delta_{i'} \cdot (1 + \varepsilon)$ for some small $\varepsilon > 0$, and $\Delta'_{i''} = \dfrac{\Delta_{i''}}{1 + \varepsilon}$, and keep all other values Δ_i intact. This change does not change the product of these two values: by our construction,

$$\Delta'_{i'} \cdot \Delta'_{i''} = \Delta_{i'} \cdot \Delta_{i''};$$

thus, the product $\prod\limits_{i=1}^{n} \Delta_i$ remains unchanged and so, the corresponding is still satisfied.

Here, from $\Delta_{i'} = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_{i'}$ and $\Delta'_{i'} > \Delta_{i'}$, we conclude that $\Delta'_{i'} > \dfrac{1}{2} \cdot \ell \cdot \varepsilon_{i'}$. Similarly, from $\Delta_{i''} > \dfrac{1}{2} \cdot \ell \cdot \varepsilon_{i''}$, for sufficiently small ε, we will still get

$$\Delta'_{i''} > \frac{1}{2} \cdot \ell \cdot \varepsilon_{i''}.$$

Thus, for the new values $\Delta'_{i'}$ and $\Delta'_{i''}$, the ℓ-constraints are satisfied as well.

Let us analyze how the change in $\Delta_{i'}$ and $\Delta_{i''}$ affect the value of the sum $s \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} a_i \cdot \Delta_i$. The new values of Δ_i are equal to

$$\Delta'_{i'} = \Delta_{i'} + \varepsilon \cdot \Delta_{i'},$$

and

$$\Delta'_{i''} = \Delta_{i''} \cdot (1 - \varepsilon + O(\varepsilon^2)) = \Delta_{i''} - \varepsilon \cdot \Delta_{i''} + O(\varepsilon^2).$$

Thus, the difference $\Delta s = s' - s$ between the new value s' of the sum and its original value s is equal to

$$\Delta s = a_{i'} \cdot (\Delta'_{i'} - \Delta_{i'}) + a_{i''} \cdot (\Delta'_{i''} - \Delta_{i''}) =$$

$$\varepsilon \cdot (a_{i'} \cdot \Delta_{i'} - a_{i''} \cdot \Delta_{i''}) + O(\varepsilon^2).$$

Since $\Delta_{i'} = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_{i'}$ and $\Delta_{i''} > \dfrac{1}{2} \cdot \ell \cdot \varepsilon_{i''}$, we conclude that

$$a_{i'} \cdot \Delta_{i'} - a_{i''} \cdot \Delta_{i''} < a_{i'} \cdot \frac{1}{2} \cdot \ell \cdot \varepsilon_{i'} - a_{i''} \cdot \frac{1}{2} \cdot \ell \cdot \varepsilon_{i''} =$$

$$\frac{1}{2} \cdot \ell \cdot (a_{i'} \cdot \varepsilon_{i'} - a_{i''} \cdot \varepsilon_{i''}).$$

We assumed that $\varepsilon_{i'} \cdot a_{i'} < \varepsilon_{i''} \cdot a_{i''}$, so the difference in parentheses is negative and thus, $\Delta s = s' - s < 0$ and $s' < s$. This inequality contradicts to the fact that s is the smallest possible value of the sum under given constraints. This contradiction shows that when $\varepsilon_{i'} \cdot a_{i'} < \varepsilon_{i''} \cdot a_{i''}$, we indeed cannot have $2\Delta_{i'} = \ell \cdot \varepsilon_{i'}$ and $2\Delta_{i''} > \ell \cdot \varepsilon_{i''}$.

To describe the resulting optimal solution, let us sort the variables 1 through n in the decreasing order of the product $a_i \cdot \varepsilon_i$:

$$a_1 \cdot \varepsilon_1 \ge a_2 \cdot \varepsilon_2 \ge \ldots \ge a_n \cdot \varepsilon_n.$$

In this order, if $2\Delta_{i'} = \ell \cdot \varepsilon_{i'}$ for some i', then for all i for which $\varepsilon_{i'} \cdot a_{i'} < \varepsilon_i \cdot a_i$, we cannot have $2\Delta_i > \ell \cdot \varepsilon_i$ and thus, we must have $2\Delta_i = \ell \cdot \varepsilon_i$. In other words, there exists a threshold t such that

- when $i \le t$, we have $\Delta_i = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_i$, and
- when $i > t$, we have $\Delta_i > \dfrac{1}{2} \cdot \ell \cdot \varepsilon_i$.

Minimizing the sum $s = \sum\limits_{i=1}^{n} a_i \cdot \Delta_i$ over all the variables Δ_i, $i = t+1, \ldots, n$, we – similarly to the case of k-anonymity optimization – conclude that $\Delta_i = \dfrac{c_t}{a_i}$ for some constant c_t. The constant c_t can be determined from the condition that

$$\rho(x) \cdot 2^n \cdot \prod_{i=1}^{n} \Delta_i = k.$$

Substituting, into this formula, $\Delta_i = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_i$ for $i = 1, \ldots, t$, and $\Delta_i = \dfrac{c_t}{a_i}$ for $i = t+1, \ldots, n$, we conclude that

$$\rho(x) \cdot 2^n \cdot \frac{1}{2^t} \cdot \ell^t \cdot \prod_{i=1}^{t} \varepsilon_i \cdot \frac{c_t^{n-t}}{\prod\limits_{i=t+1}^{n} a_i} = k,$$

hence

$$c_t^{n-t} = \frac{k \cdot \prod\limits_{i=t+1}^{n} a_i}{\rho(x) \cdot 2^{n-t} \cdot \ell^t \cdot \prod\limits_{i=1}^{t} \varepsilon_i},$$

so

$$c_t = \frac{1}{2} \cdot \left(\frac{k \cdot \prod\limits_{i=t+1}^{n} a_i}{\rho(x) \cdot \ell^t \cdot \prod\limits_{i=1}^{t} \varepsilon_i} \right)^{1/(n-t)}.$$

We need to make sure that for all $i > t$, the resulting values $\Delta_i = \dfrac{c_t}{a_i}$ satisfy the constraint $2\Delta_i \ge \ell \cdot \varepsilon_i$. In other words, we needs to make sure that $2 \cdot \dfrac{c_t}{a_i} \ge \ell \cdot \varepsilon_i$ or, equivalently, that $\dfrac{2c_t}{\ell} \ge a_i \cdot \varepsilon_i$ for all $i = t+1, \ldots, n$. Since the sequence $a_i \cdot \varepsilon_i$ is decreasing, it is sufficient to check that $\dfrac{2c_t}{\ell} \ge a_{t+1} \cdot \varepsilon_{t+1}$; if this inequality is satisfied, then $\dfrac{2c_t}{\ell}$ is automatically larger than or equal to all the following values $a_i \cdot \varepsilon_i$ for $i = t+2, \ldots, n$.

The above formulas describes the solution provided that we know the threshold t. Of all possible values of t, we must select the value that minimizes the sum

$$\Delta(t) = \sum_{i=1}^{n} a_i \cdot \Delta_i = \sum_{i=1}^{t} a_i \cdot \Delta_i + \sum_{i=t+1}^{n} a_i \cdot \Delta_i.$$

Substituting, into this formula, the expressions $\Delta_i = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_i$ for $i \le t$ and $\Delta_i = \dfrac{c_t}{a_i}$ for $i > t$, we get

$$\Delta(t) = \frac{1}{2} \cdot \ell \cdot \sum_{i=1}^{t} a_i \cdot \varepsilon_i + (n-t) \cdot c_t.$$

So, we arrive at the following algorithm.

Resulting algorithm. Around each point x, we first compute the values

$$\Delta_i = \frac{1}{2} \cdot \sqrt[n]{\frac{k}{\rho(x)}} \cdot \frac{\sqrt[n]{\prod\limits_{j=1}^{n} a_j}}{a_i},$$

where $a_i = \left| \dfrac{\partial C}{\partial x_i} \right|$. If each of these values satisfies the inequality $2\Delta_i \ge \ell \cdot \varepsilon_i$, then Δ_i are the widths that we select.

If at least one of the inequalities $2\Delta_i \ge \ell \cdot \varepsilon_i$ is not satisfied, then we first sort the n quantities in the decreasing order of the product $a_i \cdot \varepsilon_i$:

$$a_1 \cdot \varepsilon_1 \ge a_2 \cdot \varepsilon_2 \ge \ldots \ge a_n \cdot \varepsilon_n.$$

Then, for each t from 1 to n, we do the following:

- we compute the value

$$c_t = \frac{1}{2} \cdot \left(\frac{k \cdot \prod\limits_{i=t+1}^{n} a_i}{\rho(x) \cdot \ell^t \cdot \prod\limits_{i=1}^{t} \varepsilon_i} \right)^{1/(n-t)};$$

- we check whether this value c_t satisfies the inequality

$$\frac{2c_t}{\ell} \ge a_{t+1} \cdot \varepsilon_{t+1};$$

  if this inequality is not satisfied, we dismiss this value t and go to the next one;
- if the inequality $\dfrac{2c_t}{\ell} \ge a_{t+1} \cdot \varepsilon_{t+1}$ is satisfied, then we compute the value

$$\Delta(t) = \frac{1}{2} \cdot \ell \cdot \sum_{i=1}^{t} a_i \cdot \varepsilon_i + (n-t) \cdot c_t.$$

We then select the threshold t for which the value $\Delta(t)$ is the smallest. Once this t is selected, we take the following half-widths:

- we take $\Delta_i = \dfrac{1}{2} \cdot \ell \cdot \varepsilon_i$ for $i \le t$, and
- we take $\Delta_i = \dfrac{c_t}{a_i}$ for $i > t$.

Comment. The computation time of this algorithm is quadratic in n. This is OK, since the number n of different characteristics is usually reasonably small. What is important is that the algorithm is still linear-time in terms of the number of records N (which can be large), since all computations are performed point-by-point.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, by Grant 1 T36 GM078000-01 and grant "Balancing disclosure risk with inferential power: software for intervalized data" from the National Institutes of Health, and by a grant on F-transforms from the Office of Naval Research.

The authors are thankful to Scott Ferson, Lev Ginzburg, and Luc Longpré for valuable discussions.

REFERENCES

[1] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints", *ACM Transactions on Database Systems*, 2009, Vol. 34, No. 2, Article 9.

[2] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, 2012.

[3] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[4] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based System*, 2002, Vol. 10, No. 5, pp. 557–570.