Interval Computations Related to Privacy in Statistical Databases

Luc Longpré and Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
{longpre,vladik}@cs.utep.edu

Abstract

We show that the need to maintain privacy in statistical databases naturally leads to interval computations, and provide feasible algorithms for the corresponding interval computation problems.

1 Privacy in Statistical Databases: A Problem

1.1 What is a Statistical Database

Privacy is an important issue in the statistical analysis of human-related data. For example, to check whether in a certain geographic area, there is a gender-based discrimination, we can use the census data to check, e.g., whether for all people from this area who have the same same level of education, there is a correlation between salary and gender. One can think of numerous possible questions of this type related to different sociological, political, medical, economic, and other questions. From this viewpoint, it is desirable to give researches *ability to perform* whatever *statistical analysis* of this data that is reasonable for their specific research.

On the other hand, we do not want to give them direct access to the raw census data, because a large part of the census data is *confidential*. For example, for most people (those who work in private sector) salary information is confidential. Suppose that a corporation is deciding where to built a new plant and has not yet decided between two possible areas. This corporation would benefit from knowing the average salary of people of needed education level in these two areas, because this information would help them estimate how much it will cost to bring local people on board. However, since salary information is confidential, the company should not be able to know the exact salaries of different potential workers.

The need for privacy is also extremely important for *medical* experiments, where we should be able to make statistical conclusions about, e.g., the efficiency of a new medicine without disclosing any potentially embarrassing details from the individual medical records.

Such databases in which the outside users have cannot access individual records but can solicit statistical information are often called *statistical databases*.

1.2 Maintaining Privacy Is Not Easy

Maintaining privacy in statistical databases is not easy. Clerks who set up policies on access to statistical databases sometimes erroneously assume that once the records are made anonymous, we have achieved perfect privacy. Alas, the situation is not so easy: even when we keep all the records anonymous, we can still extract confidential information by asking appropriate questions. For example, suppose that we are interested in the salary of Dr. X who works for a local company. Dr. X's mailing address can be usually taken from the phone book; from the company's webpage, we can often get his photo and thus find out his race and approximate age. Then, to determine Dr. X's salary, all we need is to ask what is the average salary of all people with a Ph.D. of certain age brackets who live in a small geographical area around his actual home address – if the area is small enough, then Dr. X will be the only person falling under all these categories.

Even if only allow statistical information about salaries $s_1, \ldots, s_q$ when there are several people within a requested range, we will still be able to reconstruct the exact salaries of all these people if we ask for an average salary

$$\frac{s_1 + \ldots + s_q}{q},$$

and for several moments of salary (variance, third moment, etc): if we know the values v_j at least q different functions $f_j(s_1, \ldots, s_q)$ of s_i, then we can, in general, reconstruct all these values from the corresponding system of q equations with q unknowns: $f_1(s_1 \ldots, s_q) = v_1, \ldots, f_q(s_1, \ldots, s_q) = v_q$.

At first glance, moments are natural and legitimate statistical characteristics, so researchers would be able to request them, but on the other hand, we do not want them to be able to extract the exact up-to-cent salaries of all the folks leaving in a certain geographical area. What restriction should we impose on possible statistical queries that would guarantee privacy but restrict research in the least possible way?

A similar problem occurs in *security*: for example, we want to be able to publish statistical data about testing of a new plane without enabling potential competitors (or, even worse, potential enemy) to extract the original data and thus, gain a detailed insight into the design of this new plane. This issue was a big concern in the Soviet Union, and even there – as one of the authors (VK)

has personally witnessed – there have been spectacular security lapses. Two anecdotal examples:

- The first example shows that by asking the values of sufficiently many functions, we can potentially reconstruct all the – supposedly unavailable – data. VK was working at the Special Astrophysical Observatory, Pulkovo, Russia, on Very Large Baseline Interferometry (VLBI). The main idea of VLBI is that by observing the same distant source (e.g., quasar) from different telescopes, we can determine, very accurately, the relation positions of these telescopes, the time delay between the clocks, the exact location of the observed sources, etc. Although in principle, it is possible to determine the telescope's coordinates α and δ from these measurements without any prior information, the more we know about these coordinates, the fewer computations we need, and with GHz data (billions of bits per second), computation time was a big concern. Good quality coordinates of all geographic objects were top secret – so that the "enemy" would not be able to aim precisely. So, we could not get the actual values of α and δ. However, it turned out that while we could not be legally provided with the actual values of α and δ, there was no legal restriction against providing with us with necessary *combinations* of α and δ. So, by asking for the values of $\alpha \cdot \cos(\delta)$ and $\alpha \cdot \sin(\delta)$, we could easily reconstruct both "top secret" values.

- Another example is that sometimes, the very fact that some information is concealed discloses some supposedly secret information. In the Ukrainian city of Dnepropetrovsk, like in many other Soviet cities, there were quite a few military plants. It was known that one of these plants produces the most secret devices – intercontinental missiles. We common folks were not supposed to know where these plants are located, especially we were not supposed to know where the missiles were produced. To ensure this protection, all military plants had no signs outside – which, of course, achieved the opposite effect of telling us immediately where military plants are located. But, to ensure that the missile plant is doubly protected, the authorities not only made sure that this plant had no sign outside – they also deleted it from the freely available city maps. The result, of course, was the opposite to what was intended: by investing 20 kopecks or so into a local city map, one could easily detect a missile plant as the only plant that was not on the published city map.

1.3 What Is Known, and What We Are Planning to Do

These are anecdotal example of poorly designed privacy and security, but, as we have mentioned, the problem is indeed difficult: many seemingly well-designed privacy schemes later turn out out to have unexpected privacy and security problem. For different aspects of the problem of privacy in statistical databases,

and for different proposed solution to this problem and their drawbacks, the readers is referred to [1, 2, 3, 4, 5, 8, 9, 10, 11, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]; an extended bibliography of pre-1980s papers appears in Chapter 6 of [2].

That the privacy problem is really difficult was confirmed by the fact that several formalizations of this general privacy problem turned out to be, in their *general* formulations, NP-hard [2].

In this paper, we consider several *simple* cases of this problem. We start with the simplest possible of the privacy problem, the case when from each individual, we collect a single real number: the value of a real-valued characteristic. We show that for this simple problem, privacy naturally leads to interval computations. Then, we show how this result can be extended to the case of several characteristics.

2 Privacy in Statistical Databases Naturally Leads to Interval Computations

A natural way to fully describe a single real-valued random variable η is to provide the values of its cumulative density function (CDF) $F(x) = \text{Prob}(\eta \le x)$ for all possible real numbers x. Once we know $F(x)$, we can determine the values of all possible statistical characteristics of this random variable – e.g., its first moment, second moment, variance, etc. Thus, it is natural to allow the users to solicit the values of $F(x)$ for different x; from this information, the users will be able to reconstruct all other statistical characteristics.

For discrete data $x_1, \ldots, x_n$, the corresponding sample distribution – in which each value x_i occurs with probability $1/n$ – is described by the CDF $F(x)$ for which

$$F(x) = \frac{\#i : x_i \le x}{n}. \tag{1}$$

To get the full information about the data, we should allow the user to ask for the values $F(x)$ for all possible real numbers x. However, since we are worried about privacy, we may want to restrict possible queries to only some real values x. Let us formulate privacy in more precise terms. Suppose that we already have a partial information about one of the values x_i, e.g., that we know bounds for x_i: $x_i \in \mathbf{a} \stackrel{\text{def}}{=} [\underline{a}, \overline{a}]$. Perfect privacy means that after the user asks all allowed queries about different values of $F(x)$, the user will not be able to get any additional information about x_i – i.e., the user will not be able to deduce a narrower interval containing x_i. Let us describe this requirement in precise terms.

Definition 1.

- *By a 1D statistical database, we mean a finite sequence $x_1, \ldots, x_n$ of real numbers.*
- *By a query policy, we mean a subset $X \subseteq R$ of the set of all real numbers.*
- *For a given query policy X and a given 1D statistical database, by query results, we mean the values $F(x)$ (described by formula (1)) for all $x \in X$.*

If we allow all possible queries, i.e., if the query policy is $X = R$, then, from the query results, we can reconstruct all n values $x_1, \ldots, x_n$: indeed, they are exactly the values x at which the function $F(x)$ is discontinuous ("makes a jump"). Since we are only allowing some queries ($X \neq R$), from the query results, we may not be able to reconstruct the original 1D statistical database; in other words, it may happen that two different 1D statistical databases lead to the same query results.

Definition 2.

- *We say that a formula can be deduced from the query results $F(x)$, $x \in X$, if this formula holds for all 1D statistical databases that lead to given query results.*
- *Let X be a query policy, and let $\{x_1, \ldots, x_n\}$ be a statistical database. By a privacy violation, we mean a triple consisting of an integer i and of two intervals $\mathbf{b} \subset \mathbf{a}$ ($\mathbf{b} \neq \mathbf{a}$) for which the formula*

  $$(x_i \in \mathbf{a}) \to (x_i \in \mathbf{b})$$

  can be deduced from the corresponding query results.
- *We say that a given query policy X maintains perfect privacy for a given 1D statistical database $\{x_1, \ldots, x_n\}$ if no privacy violations are possible for the corresponding query results.*

It turns out that if a query policy maintains perfect privacy, then we can make the following conclusion about it:

Proposition 1. *If a query policy X maintains perfect privacy for a statistical database $\{x_i\}$, then between every two values $x, x' \in X$, there is at least one value x_i.*

Proof. The proof is by reduction to a contradiction. Let us assume that there exist values $x' < x''$ from X for which there is no x_i in the interval $(x', x'']$. Let us show that in this case, we have a privacy violation.

Indeed, since there are no values x_i inside the interval $(x', x'']$, all the values x_i are either smaller than or equal to x' or larger than x''. Without losing

generality, let us assume that there are values $x_i \le x'$. Let x_j be the largest of such values, i.e., $x_j \le x'$ and there are no other values x_i inside the interval $[x_j, x']$. Let $z \le x_j$ be any real number which does not exceed x_j. Let us show that in this case, the triple

$$(j, \mathbf{a} = [z, x''], \mathbf{b} = [z, x'])$$

is a privacy violation.

Indeed, let us assume that we know that $x_j \in \mathbf{a}$. Query results include the values $F(x')$ and $F(x'')$. Since there are no values x_i between x' and x'', we have $F(x') = F(x'')$. Vice versa, from the query results, we conclude that $F(x') = F(x'')$ and thus, we conclude that there are no values x_i in the interval $(x', x'']$. Thus, once we know that $x_j \in \mathbf{a} = [z, x'']$, we can use the fact that x_j cannot be inside $(x', x'']$ and thus, conclude that $x_j \in [z, x']$.

The existence of a privacy violation contradicts to our assumption that the query policy X maintains perfect privacy. This contradiction shows that our initial assumption was false and therefore, that between every two values $x, x' \in X$, there is at least one value x_i. The proposition is proven.

Due to Proposition 1, if we sort n values into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, we can conclude that in each of $n+1$ interval $(-\infty, x_{(1)}]$, $(x_{(1)}, x_{2)}]$, $\ldots$, $(x_{(n-1)}, x_{(n)}]$, $(x_{(n)}, \infty)$, there is at most one element $x \in X$. Thus, the set X of allowed queries consists of no more than $n+1$ values.

Let us sort these allowed values into an increasing sequence $X = \{x^{(1)}, \ldots, x^{(m)}\}$, where $x^{(1)} < x^{(2)} < \ldots < x^{(m)}$. After getting answers to the corresponding queries, all we know is the values $F(x^{(i)})$, i.e., how many of x_i are $\le x^{(i)}$. In other words, the only information that we get about x_i is how many of these values are in $(-\infty, x^{(1)}]$, how many are within the interval $(x^{(1)}, x^{(2)}]$, how many are within the interval $(x^{(2)}, x^{(3)}]$, etc. – but we do not know where exactly on these intervals are there values located.

In other words, instead of the actual values x_i, all we know is *intervals* that contain these values. In practice, it is convenient to consider closed intervals, so we have intervals $(-\infty, x^{(1)}]$, $[x^{(1)}, x^{(2)}]$, etc. It is important to mention that these intervals are *almost disjoint* in the sense that every two of them either coincide, or have no intersection at all, or their intersection is a single point.

Based on the values $F(x)$, we must determine other statistical characteristics. If we knew all the values $F(x)$, $x \in R$, then we would be able to reconstruct the actual values x_i, and thus, we would be able to compute the sample average E, the sample variance V, etc. Since we are only allowed to ask queries $F(x)$ for $x \in X$, we can only deduce the intervals that contain x_i; based on these intervals of possible values of x_i, we must therefore find the intervals of possible values of E, V, etc. The problem of computing such intervals is a particular case of the general problem of interval computations – where we have a function $f(x_1, \ldots, x_n)$, we know the intervals $\mathbf{x}_i$ of possible values of the inputs x_i, and

we must find the range $\mathbf{y}$ of possible values of $f(x_1, \ldots, x_n)$ when $x_i \in \mathbf{x}_i$. Thus, privacy in statistical databases naturally leads to interval computations.

3 Specific Interval Computation Problems Related to Privacy in Statistical Databases, and How to Solve Them

Before we start describing algorithms for solving the corresponding interval computation problems, let us first mention that for some values x_i, we have infinite intervals – with the possibility of arbitrarily small or arbitrarily large values. Thus, even for the simplest statistical characteristic – sample average

$$E = \frac{x_1 + \ldots + x_n}{n}, \tag{2}$$

the set of possible values of E simply coincides with the entire real line R. To get non-trivial estimates, we must ignore the smallest and the largest values – for which no containing interval is known – and only consider values which are bounded by intervals.

For the average (2), once we know the intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ that contain x_i, we can easily find the range $\mathbf{E} = [\underline{E}, \overline{E}]$ of possible values of E:

$$\underline{E} = \frac{\underline{x}_1 + \ldots + \underline{x}_n}{n}; \quad \overline{E} = \frac{\overline{x}_1 + \ldots + \overline{x}_n}{n}.$$

The problem of estimating the range $\mathbf{V} = [\underline{V}, \overline{V}]$ of the sample variance

$$V = \frac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n - 1} \tag{3}$$

when $x_i \in [\underline{x}_i, \overline{x}_i]$ is not so easy: while there is a feasible algorithm for computing the lower bound $\underline{V}$ of this range, the general problem of computing the upper endpoint $\overline{V}$ of this range for x_i is known to be NP-hard [6, 7, 12, 14]. It turns out, however, that for specific case of intervals coming from privacy, a special feasible algorithm is possible that computes $\overline{V}$:

Proposition 2. *There exists a quadratic-time algorithm that computes the exact range* $\mathbf{V}$ *of the variance* V *for the case when intervals* $\mathbf{x}_i$ *of possible values of* x_i *are pairwise almost disjoint.*

Proof. Since there exists an algorithm that computes $\underline{V}$ in feasible time (see, e.g., [6, 7]), it is sufficient to produce a feasible algorithm for computing $\overline{V}$.

According to the proof of Theorems 4.1 and 4.2 from [6], the values $x_i \in \mathbf{x}_i$ that lead to the largest possible value of V satisfy the following property:

- if $E \leq \underline{x}_i$, then $x_i = \overline{x}_i$;

- if $E \geq \overline{x}_i$, then $x_i = \underline{x}_i$;
- if $E \in (\underline{x}_i, \overline{x}_i)$, then $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$.

In order to use this property to compute $\overline{V}$, we test all possible locations of E in relation to the intervals $\mathbf{x}_i$: $E = \underline{x}_i$, $E = \overline{x}_i$, and $E \in (\underline{x}_i, \overline{x}_i)$ for different $i = 1, 2, \ldots, n$.

Let us first consider the cases when $E = \underline{x}_i$ (the case when $E = \overline{x}_i$ is treated similarly). In these cases, since the intervals $\mathbf{x}_i$ are almost disjoint, the above property uniquely determines the values x_i; thus, we can compute E, check whether it indeed satisfies the corresponding condition, and if yes, compute the corresponding value V.

Let us now consider the cases when $E \in (\underline{x}_i, \overline{x}_i)$. Let k denote the number of different intervals of such type, and let n_j, $j = 1, \ldots, k$ denote the number of intervals $\mathbf{x}_i$ that coincide with j-th interval. Then, $n = n_1 + \ldots + n_k$. For each of these k intervals $\mathbf{x}_j$, the values of x_i are uniquely determined when $\overline{x}_j \leq \underline{x}_i$ or $\overline{x}_i \leq \underline{x}_j$; for the remaining n_j values x_i, we have $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$. Modulo transposition, the resulting set of values $\{x_1, \ldots, x_n\}$ is uniquely determined by how many of these n_j x_i's are equal to $\overline{x}_i$. The number of such x_i's can be 0, 1, 2,..., $n_j + 1$. Thus, the total number of such combinations is equal to $n_j + 1$. Overall, for all j from 1 to k, we have

$$\sum_{j=1}^{k} (n_j + 1) = \sum_{j=1}^{k} n_j + k = n + k \leq 2n$$

resulting sets $\{x_1, \ldots, x_n\}$. For each of these sets, we compute E, check that the resulting E is indeed inside the corresponding interval $\mathbf{x}_i$, and if it is, we compute V.

Thus, we have $\leq 2n + n = 3n$ cases, for each of which we need $O(n)$ computations to compute V. The largest of these V is the desired $\overline{V}$, and we compute it in time $\leq 3n \cdot O(n) = O(n^2)$. The proposition is proven.

A similar algorithm can be proposed for computing covariance

$$C = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}), \tag{4}$$

where

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}; \quad \bar{y} = \frac{y_1 + \ldots + y_n}{n}.$$

In general, the problem of computing the range $\mathbf{C}$ of covariance C over intervals $\mathbf{x}_i$ and $\mathbf{y}_i$ is NP-hard [6, 7, 12, 14]. It turns out, however, that for specific case of intervals coming from privacy, a special feasible algorithm is possible that computes $\mathbf{C}$:

Proposition 3. *There exists a cubic-time algorithm that computes the exact range* $\mathbf{C}$ *of the covariance* C *for the case when intervals* $\mathbf{x}_i$ *of possible values of* x_i *are pairwise almost disjoint, and the intervals* $\mathbf{y}_i$ *of possible values of* y_i *are pairwise almost disjoint.*

Proof. Let us describe the algorithm for computing the largest possible value $\overline{C}$ of the covariance C (the problem of computing minimum can be reduced to computing maximum if we take $y_i' \stackrel{\text{def}}{=} y_i$ with intervals $\mathbf{y}_i' = -\mathbf{y}_i$).

Let $x_1, \ldots, x_n, y_1, \ldots, y_n$ be values for which which the covariance C attains its maximum. The covariance function is linear with respect to each of its $2n$ variables; therefore, this maximum is attained either for $x_i = \overline{x}_i$ or for $x_i = \underline{x}_i$ – depending on whether the function C increases or decreases as a functions of x_i (same for y_i).

The derivative of C with respect to x_i can be easily computed as follows:

$$\frac{\partial C}{\partial x_i} = \frac{1}{n} \cdot (y_i - \bar{y}) - \frac{1}{n^2} \sum_{j=1}^{n} (y_j - \bar{y}).$$

Since $\sum_{j=1}^{n} (y_j - \bar{y}) = 0$, we conclude that

$$\frac{\partial C}{\partial x_i} = \frac{1}{n} \cdot (y_i - \bar{y}).$$

Thus:

- when $y_i \geq \bar{y}$, we have $\dfrac{\partial C}{\partial x_i} \geq 0$ hence $x_i = \overline{x}_i$;
- when $y_i \leq \bar{y}$, we have $\dfrac{\partial C}{\partial x_i} \leq 0$ hence $x_i = \underline{x}_i$.

Hence:

- when $\underline{y}_i \geq \bar{y}$, we have $y_i \geq \underline{y}_i \geq \bar{y}$ hence $x_i = \overline{x}_i$;
- when $\overline{y}_i \leq \bar{y}$, we have $y_i \leq \overline{y}_i \leq \bar{y}$ hence $x_i = \underline{x}_i$.

Similarly:

- when $\underline{x}_i \geq \bar{x}$, we have $y_i = \overline{y}_i$;
- when $\overline{x}_i \leq \bar{x}$, we have $y_i = \underline{y}_i$.

Due to these implications, to find the largest possible value $\overline{C}$ of the covariance C, it is sufficient to consider all $k_x \cdot k_y$ possible cases when $\bar{x} \in \mathbf{x}_i$ and when $\bar{y} \in \mathbf{y}_i$, where k_x is the number of different intervals $\mathbf{x}_i$, and k_y is the number of

different intervals $\mathbf{y}_i$. For each of these cases, similarly to the proof of Proposition 2, there are $n_j(y)+1$ possible situations with x and $n_k(x)+1$ psosible situations with y: the total of $\leq (n_k(x)+1)\cdot(n_j(y)+1)$. The overall number of resulting combinations $x_1,\ldots,x_n,y_1,\ldots,y_n$ is therefore bounded by

$$\sum_j \sum_k (n_k(x)+1)\cdot(n_j(y)+1) = \left(\sum_k (n_k(x)+1)\right)\cdot\left(\sum_j (n_j(y)+1)\right) \leq$$

$$2n\cdot 2n = O(n^2).$$

For each of these cases, we need $O(n)$ arithmetic operations ot check whether the resulting values $\bar{x}$ and $\bar{y}$ are indeed within the corresponding intervals and, if they are, to compute C. Thus, we need $O(n^2)\cdot O(n) = O(n^3)$ computational steps to compute $\overline{C}$ as the largest of these $O(n^2)$ values. The proposition is proven.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209 and grant NCC2-1232, by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-00-1-0365, by NSF grants CDA-9522207, ERA-0112968 and 9710940 Mexico/Conacyt, by Small Business Innovation Research grant 9R44CA81741 to Applied Biomathematics from the National Cancer Institute (NCI), a component of the National Institutes of Health (NIH), and by a research grant from Sandia National Laboratories as part of the Department of Energy Accelerated Strategic Computing Initiative (ASCI).

This work was partly done while one of the authors (LL) participated in the NSF-sponsored Carnegie Mellon Educating Information-Security Experts Program and another author (VK) was a visiting researcher at the Euler International Mathematical Institute, St. Petersburg, Russia.

The authors are thankful to Daniel Berleant (Iowa University), Andrew Bernat (Computing Research Association), Scott Ferson and Lev Ginzburg (Applied Biomathematics), and David Novick (University of Texas El Paso) for valuable discussions.

References

[1] T. Dalenius, "Finding a needle in a haystack – or identifying anonymous census record", *Journal of Official Statistics*, 1986, Vol. 2, No. 3, pp. 329–336.

[2] D. E. R. D. Denning, *Cryptography and data security*, Addison-Wesley, Reading, MA, 1982.

[3] G. Duncan and D. Lambert, "The risk of disclosure for microdata", *Proc. of the Bureau of the Census Third Annual Research Conference*, Bureau of the Census, Washington, DC, 1987, pp. 263–274.

[4] G. Duncan and S. Mukherjee, "Microdata disclosure limitation in statistical databases: query size and random sample query control", *Prof. 1991 IEEE Symposium on Research in Security and Privacy*, Oakland, CA, May 20–22, 1991.

[5] I. Fellegi, "On the question of statistical confidentiality", *Journal of the American Statistical Association*, 1972, pp. 7–18.

[6] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Computing Variance for Interval Data is NP-Hard", *ACM SIGACT News*, 2002, Vol. 33, No. 2, pp. 108–118.

[7] S. Ferson, L. Ginzburg, V. Kreinovich, and M. Aviles, "Exact Bounds on Sample Variance of Interval Data", *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, Toronto, Canada, May 23–25, 2002, pp. 67–69.

[8] J. Kim, "A method for limiting disclosure of microdata based on random noise and transformation", *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 1986, pp. 370–374.

[9] N. Kirkendall et al., *Report on Statistical Disclosure Limitations Methodology*, Office of Management and Budget, Washington, DC, Statistical Policy Working Paper No. 22, 1994.

[10] M. Morgenstern, "Security and inference in multilevel database and knowledge base systems", *Proc. of the ACM SIGMOD Conference*, 1987, pp. 357–373.

[11] Office of Technology Assessment, *Protecting privacy in computerized medical information*, US Government Printing Office, Washington, DC, 1993.

[12] R. Osegueda, V. Kreinovich, L. Potluri, and R. Aló, "Non-Destructive Testing of Aerospace Structures: Granularity and Data Mining Approach", *Proceedings of FUZZ-IEEE'2002*, Honolulu, Hawaii, May 12–17, 2002, Vol. 1, pp. 685–689.

[13] M. Palley and J. Siminoff, "Regression methodology based disclosure of a statistical database", *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 1986, pp. 382–387.

[14] L. Potluri, *Correlation techniques for information fusion*, University of Texas at El Paso, Department of Computer Science, Master's Thesis, 2002.

[15] T. Su and G. Ozsoyoglu, "Controlling FD and MVD inference in multilevel relational database systems", *IEEE Transactions on Knowledge and Data Engineering*, 1991, Vol. 3, pp. 474–485.

[16] L. Sweeney, "Replacing personally-identifying information in medical records, the scrub system", *Journal of the American Medical Informatics Association*, 1996, pp. 333–337.

[17] L. Sweeney, "Weaving technology and policy together to maintain confidentiality", *Journal of Law, Medicine and Ethics*, 1997, Vol. 25, pp. 98–110.

[18] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the datafly system", *Journal of the American Medical Informatics Association*, 1997, pp. 51–55.

[19] L. Sweeney, "Computational disclosure control for medical microdata", *Proceedings of the Record Linkage Workshop*, Bureau of the Census. Washington, DC, 1997.

[20] L. Sweeney, "Commentary: researchers need not rely on consent or not", *New England Journal of Medicine*, 1998, Vol. 338, No. 15.

[21] L. Sweeney, "Towards the optimal suppression of details when disclosing medical data, the use of sub-combination analysis", *Proceedings of MEDINFO'98, International Medical Informatics Association, Seoul, Korea*, North-Holland, 1998, p. 1157.

[22] L. Sweeney, "Three computational systems for disclosing medical data in the year 1999", *Proceedings of MEDINFO'98, International Medical Informatics Association, Seoul, Korea*, North-Holland, 1998, pp. 1124–1129.

[23] L. Sweeney, "Datafly: a system for providing anonymity in medical data", In: T. Y. Lin and S. Qian (eds.), *Database Security XI: Status and Prospects*, Elsevier, Amsterdam, 1998.

[24] L. Willenborg and T. De Waal, *Statistical disclosure control in practice*, Springer Verlag, New York, 1996.