Uncertainty Can Decrease Privacy: An Observation

Karen Villaverde[1] Olga Kosheleva[2]

1. Department of Computer Science
New Mexico State University
Las Cruces, NM 88003, USA
2. Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
Email: kvillave@cs.nmsu.edu, olgak@utep.edu

Abstract— *In many application areas, e.g., in medical applications, it is important to be able to get statistical information about the data without disclosing individual cases. One of the methods to preserve privacy in such statistical databases is to introduce uncertainty, i.e., to replace the exact values with intervals of possible values. In this paper, we show that while the introduction of uncertainty can often enhance privacy, sometimes, the opposite effect occurs: adding uncertainty can decrease privacy.*

Keywords— Fuzzy uncertainty, interval uncertainty, privacy, statistical databases

1 Statistical Databases – Need to Preserve Privacy

1.1 Need for Collecting Data

In many practical situations, it is very useful to collect large amounts of data.

For example, from the data that we collect during a census, we can extract a lot of information about health, mortality, employment in different regions – for different age ranges, and for people from different genders and of different ethnic groups. By analyzing these statistics, we can reveal troubling spots and allocate (usually limited) resources so that the help goes first to social groups that need it most.

Similarly, by gathering data about people's health in a large medical database, we can extract a lot of useful information on how the geographic location, age, and gender affect a person's health. Then, we can improve the public health by appropriate appropriate public health measures to different portions of the population.

Finally, a large database of purchases can help find out what people are looking for, make shopping easier for customers and at the same time, decrease the stores' expenses related to storing unnecessary items.

1.2 Need for Privacy

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, one can use the census data to check, e.g., whether for all people from this area who have the 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 researchers *ability to perform* whatever *statistical analysis* of this data that is reasonable for their specific research.

On the other hand, we may not want to give the researchers 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 the 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 cannot access individual records but can solicit statistical information are often called *statistical databases*.

1.3 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.

Many examples of such extraction can be found in a book by D. Denning [1]. 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

ISBN: 978-989-95079-6-8

only person falling under all these categories.

Even if we only allow statistical information about salaries $s_1, \ldots, s_q$ when there are at least a certain amount n_0 people within a requested range, we will still be able to reconstruct the exact salaries of all these people. Indeed, for example, we can ask for the number and average salary of all the people who live on Robinson street at houses 1 through 1001, and then we can ask the same question about all the people who live in houses from 1 to 1002. By comparing the two numbers, we get the average salary of the family living at 1002 Robinson – in other words, we gain the private information that we tried to protect.

In general, we can ask for the average

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

and for several moments of salary (variance, third moment, etc):

- if we know the values v_j of 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 these characteristics. On the other hand, we do not want the researchers to be able to extract the exact up-to-cent salaries of all the people 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?

These are anecdotal examples 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 to have unexpected privacy and security problems, and it is known that the problem of finding a privacy-preserving scheme is, in general, NP-hard [1].

Different aspects of the problem of privacy in statistical databases, different proposed solutions to this problem, and their drawbacks, are described in [1, 11, 12] (see also references therein).

2 Known Fact: Uncertainty Can Enhance Privacy

A reasonable way to avoid privacy violations is to store ranges (intervals) of values instead of the actual values. For example, instead of keeping the exact age, we only record whether the age is between 0 and 10, 10 and 20, 20 and 30, etc.

In this case, no matter what statistics we allow, the worst that can happen is that the corresponding ranges will be disclosed. However, in this situation, we do not disclose the original exact values – since these values are not stored in the database in the first place; see, e.g., [5, 6, 7].

3 New Observation: Uncertainty Can Decrease Privacy

3.1 *What We Do in This Section*

The successful use of uncertainty to enhance privacy in statistical databases may lead to an impression that uncertainty *always* enhances privacy. In this paper, we show that uncertainty can actually *decrease* privacy in a statistical database.

3.2 *Simplest Possible Case: 1-D Databases*

We will show that the privacy decrease phenomenon occurs in the simplest 1-dimensional case, when each record in the statistical database consists of a single value x_i. In this case, the database consists of n values $x_1, \ldots, x_n$.

3.3 *1-D Case: Choice of Statistical Characteristics*

Among the most important statistical characteristics are the mean values of certain quantities.

In the 1-D case, each record contains the value x_i of a single quantity x. So, in the 1-dimensional case, we are interested in estimating the mean values of different characteristics $u(x)$, i.e., in estimating the values

$$E[u] \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} u(x_i). \tag{1}$$

3.4 *Smooth Characteristics*

Among different statistical characteristics, an important class is formed by characteristics for which the corresponding function $u(x)$ is smooth.

Usually, most values x_i corresponding to different individuals do not differ much. In other words, most values x_i belong to a small-size region. In this region, we can expand the function $u(x)$ in Taylor series and ignore higher-order terms. In the first approximation, we thus approximate the function $u(x)$ by a linear expression

$$u(x) \approx u_0 + u_1 \cdot x. \tag{2}$$

In the next approximation, we approximate an arbitrary smooth function $u(x)$ by a quadratic expression:

$$u(x) \approx u_0 + u_1 \cdot x + u_2 \cdot x^2. \tag{3}$$

In this paper, we will show that the phenomenon of decreasing privacy occurs already for such quadratic characteristics.

3.5 *For Exactly Known Values, the Computation of Quadratic Characteristics Preserves Privacy*

In applications, we may be interested in the mean values of different statistical characteristics. If we restrict ourselves to quadratic characteristics, this means that we must be able, given an arbitrary quadratic characteristic (3), to estimate the average value (1) of this characteristic.

Substituting the general expression (3) into the formula (1), we can conclude that the desired average value has the form

$$E[u] = u_0 + u_1 \cdot E[x] + u_2 \cdot E[x^2], \tag{4}$$

where

$$E[x] \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{5}$$

and

$$E[x^2] \stackrel{\rm def}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2. \qquad (6)$$

Thus, in effect, even when we know the mean value $E[u]$ for all possible quadratic characteristics, all we know about the values x_i are the two combinations: the mean $E[x]$ and the second moment $E[x^2]$.

Based on the knowledge of these two values, we cannot reconstruct n different values $x_1, \ldots, x_n$ and thus, privacy is preserved.

3.6 *Case of Interval Uncertainty*

Let us now consider what happens if instead of the exact value of x_i, we only know the *interval* $[\underline{x}_i, \overline{x}_i]$ of possible values of x_i.

Often, these intervals come from the fact that we know an approximate value $\widetilde{x}_i$, and we know the upper bound Δ_i on the approximation accuracy $\Delta x_i \stackrel{\rm def}{=} \widetilde{x}_i - x_i$: $|\Delta x_i| \le \Delta_i$. In this case, the only information that we have about the actual (unknown) value x_i is that x_i belongs to the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. So, we get an interval $[\underline{x}_i, \overline{x}_i]$ with $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

In general, every interval $[\underline{x}_i, \overline{x}_i]$ can be represented in the form $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$: it is sufficient to take:

- as $\widetilde{x}_i$, the midpoint of the interval, i.e., the value

$$\widetilde{x}_i = \frac{\underline{x}_i + \overline{x}_i}{2}; \qquad (7)$$

  and

- as Δ_i, the radius (half-width) of the interval, i.e., the value

$$\Delta_i = \frac{\overline{x}_i - \underline{x}_i}{2}. \qquad (8)$$

3.7 *Estimating Statistical Characteristics under Interval Uncertainty: Formulation of the Problem*

Suppose that we are interested in the mean value $E[u]$ of a statistical characteristic $u(x)$. In the ideal case when we know the exact values $x_1, \ldots, x_n$, this mean value is simply equal to the arithmetic average of the corresponding values $u(x_i)$.

Under interval uncertainty, we can have different values $x_i \in [\underline{x}_i, \overline{x}_i]$. In general, different values x_i can lead to different values of the mean. We are interested in computing the *range* $[\underline{u}, \overline{u}]$ of possible values of this mean, i.e., the interval

$$[\underline{u}, \overline{u}] \stackrel{\rm def}{=} \left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} u(x_i) : x_i \in [\underline{x}_i, \overline{x}_i] \right\}. \qquad (9)$$

3.8 *Estimating Quadratic Characteristics under Interval Uncertainty: Algorithm*

How can we compute the range of the mean

$$E[u] = \frac{1}{n} \cdot \sum_{i=1}^{n} u(x_i)?$$

Each of the variables x_i can take any value from the corresponding interval $[\underline{x}_i, \overline{x}_i]$. Thus,

- the mean $E[u]$ takes the *largest* value if and only if each term $u(x_i)$ takes the *largest* possible value on the interval $[\underline{x}_i, \overline{x}_i]$; we will denote this largest value by $\overline{u}_i$;
- similarly, the mean $E[u]$ takes the *smallest* value if and only if each term $u(x_i)$ takes the *smallest* possible value on the interval $[\underline{x}_i, \overline{x}_i]$; we will denote this smallest value by $\underline{u}_i$.

For a given quadratic function $u(x_i)$ on a given interval $[\underline{x}_i, \overline{x}_i]$, computing its smallest value $\underline{u}_i$ and its largest value $\overline{u}_i$ is an easy computational task. Indeed, according to calculus, if a function attains its minimum or maximum at some point inside an interval, then its derivative is 0 at this point. Thus, to find the minimum and the maximum of a given function $u(x_i)$ on a given interval $[\underline{x}_i, \overline{x}_i]$, it is sufficient to compute the values of this function

- at the endpoints $\underline{x}_i$ and $\overline{x}_i$ of this interval and
- at the point(s) (if any) where the derivative $u'(x)$ of the function $u(x)$ is equal to 0.

Then,

- the smallest of these values is the desired minimum $\underline{u}_i$, and
- the largest of these values is the desired maximum $\overline{u}_i$.

For a quadratic function $u(x) = u_0 + u_1 \cdot x + u_2 \cdot x^2$, the derivative is a linear function $u'(x) = u_1 + 2 \cdot u_2 \cdot x$, so computing the point where the derivative is 0 is a straightforward task: $u_1 + 2 \cdot u_2 \cdot x = 0$ implies $x = -u_1/(2u_2)$.

Once we compute the values $\underline{u}_i$ and $\overline{u}_i$, we can now find the desired range $[\underline{u}, \overline{u}]$ as

$$\underline{u} = \frac{1}{n} \cdot \sum_{i=1}^{n} \underline{u}_i, \quad \overline{u} = \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{u}_i. \qquad (10)$$

3.9 *Under Interval Uncertainty, Privacy Is No Longer Preserved: Result*

Let us show that in this case, privacy is no longer preserved. To be more precise, we assume that we are given a 1-D database, i.e., a collection of intervals $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$. We do not have an explicit access to intervals from this collection, but for every quadratic function $u(x)$, we can generate the range of the mean value $E[u]$ over these intervals.

We will show that in almost all cases, based on these ranges, we can actually reconstruct all the original intervals $[\underline{x}_i, \overline{x}_i]$. In other words, adding uncertainty leads to a loss of privacy.

Comment. By "almost all" cases, we mean all cases in which all n midpoints $\widetilde{x}_i$ are different. Situations when two midpoints coincide are indeed degenerate, since a minor modification of the original data leads to $\widetilde{x}_i \ne \widetilde{x}_j$.

3.10 *Under Interval Uncertainty, Privacy Is No Longer Preserved: Proof*

In our proof, for every real number a, we consider the quadratic function $u(x) = (x - a)^2$. For this function, the derivative is equal to 0 at the minimum point $x = a$. Thus, this function attains its largest value on the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ at one of the endpoints $\widetilde{x}_i - \Delta_i$ or $\widetilde{x}_i + \Delta_i$. One can easily check that:

ISBN: 978-989-95079-6-8

- when $a \le \widetilde{x}_i$, then the largest possible value $\overline{u}_i$ of $u(x)$ on the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ is attained when $x_i = \overline{x}_i = \widetilde{x}_i + \Delta_i$ and is equal to $\overline{u}_i = (\overline{x}_i - a)^2$;
- when $a \ge \widetilde{x}_i$, then the largest possible value $\overline{u}_i$ of $u(x)$ on the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ is attained when $x_i = \underline{x}_i = \widetilde{x}_i - \Delta_i$ and is equal to $\overline{u}_i = (\underline{x}_i - a)^2$.

Let us use this fact to describe the dependence of $\overline{u}$ on the parameter a.

When $a \ne \widetilde{x}_i$, the value $\overline{u}$ is the average of n smooth expressions.

At each point $a = \widetilde{x}_i$, all the terms $\overline{u}_j$ in the sum $\overline{u}$ are smooth except for the term $\overline{u}_i$ that turns from $(\overline{x}_i - a)^2$ to $(\underline{x}_i - a)^2$. The derivative of $\overline{u}_i$ with respect to a changes from $2 \cdot (a - \overline{x}_i)$ to $2 \cdot (a - \underline{x}_i)$, i.e., increases by

$$2 \cdot (a - \underline{x}_i) - 2 \cdot (a - \overline{x}_i) = 2 \cdot (\overline{x}_i - \underline{x}_i) = 4 \cdot \Delta_i. \quad (11)$$

Since all the other components $\overline{u}_j$ are smooth at $a = \widetilde{x}_i$, at $a = \widetilde{x}_i$, the derivative of the average $\overline{u}(a)$ increases by $\dfrac{4}{n} \cdot \Delta_i$.

Thus, once we know the value $\overline{u}$ for all a,

- we can find the values $\widetilde{x}_i$ as the values at which the derivative is discontinuous; and
- we can find each value Δ_i as $n/4$ times the increase of the derivative at the corresponding point $\widetilde{x}_i$.

The statement is proven.

3.11 Case of Fuzzy Uncertainty

In many practical situations, the estimates $\widetilde{x}_i$ come from experts. Experts often describe the inaccuracy of their estimates in terms of imprecise words from natural language, such as “approximately 0.1”, etc. A natural way to formalize such words is to use special techniques developed for formalizing this type of estimates – specifically, the technique of fuzzy logic; see, e.g., [4, 10].

In this technique, for each possible value of $x_i \in [\underline{x}_i, \overline{x}_i]$, we describe the degree $\mu_i(x_i)$ to which this value is possible. For each degree of certainty α, we can determine the set of values of x_i that are possible with at least this degree of certainty – the *α-cut* $\mathbf{x}_i(\alpha) = \{x \mid \mu_i(x) \ge \alpha\}$ of the original fuzzy set. Vice versa, if we know α-cuts for every α, then, for each object x, we can determine the degree of possibility that x belongs to the original fuzzy set [2, 4, 8, 9, 10]. A fuzzy set can be thus viewed as a nested family of its (interval) α-cuts.

3.12 From the Computational Viewpoint, Fuzzy Uncertainty Can Be Reduced to the Interval One

Once we know how to propagate interval uncertainty, i.e., how to compute the range

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=} \{f(x_1, \ldots, x_n) : x_i \in \mathbf{x}_i\} \quad (12)$$

of a given function $f(x_1, \ldots, x_n)$ over given intervals $\mathbf{x}_i$, then, to propagate the fuzzy uncertainty, we can consider, for each α, the fuzzy set y with the α-cuts

$$\mathbf{y}(\alpha) = f(\mathbf{x}_1(\alpha), \ldots, \mathbf{x}_n(\alpha)); \quad (13)$$

see, e.g., [2, 4, 8, 9, 10]. This is equivalent to using Zadeh's extension principle.

So, from the computational viewpoint, the problem of propagating fuzzy uncertainty can be reduced to several interval propagation problems.

For example, the fuzzy value of $E[u]$ can be described as follows: for each α, the corresponding α-cut is equal to the range of $E[u]$ when for each i, all values x_i belong to the corresponding α-cuts: $x_i \in \mathbf{x}_i(\alpha)$.

3.13 Under Fuzzy Uncertainty, Privacy Is Also Not Preserved: Result

Let us assume that we are given a 1-D database, i.e., a collection of fuzzy numbers $\mu_1(x_1), \ldots, \mu_n(x_n)$. We do not have an explicit access to fuzzy numbers from this collection, but for every quadratic function $u(x)$, we can generate the fuzzy number $\mu(u)$ corresponding to the mean $E[u]$.

We will show that in almost all cases (i.e., when the midpoints are all different), based on the resulting fuzzy numbers $\mu(u)$, we can actually reconstruct all the original fuzzy numbers $\mu_i(x_i)$.

In other words, adding fuzzy uncertainty also leads to a loss of privacy.

3.14 Under Fuzzy Uncertainty, Privacy Is Also Not Preserved: Proof

For each α, the α-cuts of the resulting fuzzy number $\mu(u)$ is the range of $E[u]$ when $x_i \in \mathbf{x}_i(\alpha)$.

We already know that from the ranges computed for all possible quadratic functions $u(x)$, we can reconstruct all the intervals. Thus, for every α, we can reconstruct all the α-cuts of all the membership functions $\mu_i(x_i)$.

Since we can thus reconstruct all α-cuts for all α, hence, we can uniquely reconstruct all the membership functions.

The statement is proven.

Acknowledgment

The authors are thankful to the anonymous referees for useful suggestions.

References

[1] D. Denning, *Cryptography and Data Security*, Addison-Wesley, Reading, MA, 1982.

[2] D. Dubois and H. Prade, Operations on fuzzy numbers, *International Journal of Systems Science*, 1978, Vol. 9, pp. 613–626.

[3] P. C. Fishburn, *Nonlinear preference and utility theory*, The John Hopkins Press, Baltimore, MD, 1988.

[4] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: theory and applications.* Prentice Hall, Upper Saddle River, New Jersey, 1995.

[5] V. Kreinovich and L. Longpré, Computational complexity and feasibility of data processing and interval computations, with extension to cases when we have partial information about probabilities, In: V. Brattka, M. Schroeder, K. Weihrauch, and N. Zhong, editors, *Proc. Conf. on Computability and Complexity in Analysis CCA'2003*, Cincinnati, Ohio, USA, Aug. 28–30, 2003, pp. 19–54.

[6] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos, "Interval Versions of Statistical Techniques, with Applications to Environmental Analysis, Bioinformatics, and Privacy in Statistical Databases", *Journal of Computational and Applied Mathematics*, 2007, Vol. 199, No. 2, pp. 418–423.

[7] L. Longpré, G. Xiang, V. Kreinovich, and E. Freudenthal, "Interval Approach to Preserving Privacy in Statistical Databases: Related Challenges and Algorithms of Computational Statistics", In: V. Gorodetsky, I. Kotenko, and V. A. Skormin (eds.), *Proceedings of the International Conference "Mathematical Methods, Models and Architectures for Computer Networks Security"*, MMM-ACNS-07, St. Petersburg, Russia, September 13–15, 2007, Springer Lecture Notes in Computer Science, 2007, Vol. CCIS-1, pp. 346–361.

[8] R. E. Moore and W. Lodwick, Interval Analysis and Fuzzy Set Theory, *Fuzzy Sets and Systems*, 2003, Vol. 135, No. 1, pp. 5–9.

[9] H. T. Nguyen and V. Kreinovich, Nested Intervals and Sets: Concepts, Relations to Fuzzy Sets, and Applications, In: R. B. Kearfott and V. Kreinovich, eds., *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290.

[10] H. T. Nguyen and E. A. Walker, *A first course in fuzzy logic*, CRC Press, Boca Raton, Florida, 2005.

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

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