Measuring privacy loss in statistical databases

Vinod Chirayath
University of Texas at El Paso
vinodpchirayath@yahoo.com

Luc Longpré
University of Texas at El Paso
longpre@utep.edu

Vladik Kreinovich*
University of Texas at El Paso
vladik@utep.edu

Abstract

Protection of privacy in databases has become of increasing importance. While a number of techniques have been proposed to query databases while preserving privacy of individual records in the database, very little is done to define a measure on how much privacy is lost after statistical releases. We suggest a definition based on information theory. Intuitively, the privacy loss is proportional to how much the descriptional complexity of a record decreases relative to the statistical release. There are some problems with this basic definition and we suggest ways to address these problems.

1 Introduction

The earliest article we know that refers to privacy is an 1890 Harvard Law Review article entitled "The Right to Privacy", by Louis Brandeis and Samuel Warren [9]. The authors warned that technological advances threatened our privacy. They were referring to 'instantaneous photographs' and other 'mechanical devices'. Over 100 years later, this warning is even more relevant, although technological advances revolve around computers, sharing of information, and the internet. Data collection through surveys, registration pages, user forms have resulted in more personal information being available than before. Organizations like the Census Bureau, insurance companies, hospitals, universities keep databases that contain valuable information about an individual.

A large body of literature in computer security describes research in access control, which is a model that describes who has access to what. Access control is built upon appropriate authentication. While it is important to protect our sensitive database from improper access, we are interested in a related problem. Database managers who have appropriate authorization may inadvertantly release some statistical information which is thought to be *sanitized*, or harmless to privacy but from which one can recover some sensitive information. The problem of privacy in such databases is to protect information specific to an individual while releasing aggregate data for research purposes.

*This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, and Texas Department of Transportation grant No. 0-5453

1.1 Statistical databases

A statistical database is a database where the data is collected with the purpose of releasing aggregate statistical information, or allowing certain types of queries to retrieve relevant statistical information. The major difference between a statistical database and a typical database is its limited querying interface. Querying is typically limited to operations such as count, sum and mean. The central issue in statistical databases is to provide accurate and reliable aggregates of data while preventing the disclosure of individual information.

While researchers with good intentions should be able to get the statistical information about the database, a malicious user may be able to craft a series of queries that allows the recovery of specific field values. In [7] the authors show that a mischievous researcher can calculate the value of any query, so long as the statistical database's minimum query size is set no higher than $k = \frac{N}{6}$ where k is the minimum query size and N is the size of the database.

1.2 Privacy preserving approaches

The following approaches are used to maintain privacy in statistical databases.

Query restriction Here, we only allow a restricted set of queries. This set is carefully selected so that private information cannot be recovered. Typical restrictions include query set size control, where answers to queries are not provided if the set on which the statistics is based is smaller than some prespecified threshold. Another way is to only allow queries with predetermined range. For example, statistics on population age groups are only provided for ages 0-10, 10-20 and so on.

Data perturbation Some noise is added to the database. The challenge is to preserve statistical properties of the original database. For example, if we add to a field noise that follows a normal distribution with zero mean, then the mean is not expected to change and the variance will increase in a predictable way.

Output perturbation Some noise is added to answers to queries. It is important that the noise depends on the query only, so that the result is the same whenever the same query is answered.

Query monitoring In this method, queries are monitored to ensure that it is not possible to recover sensitive information even with answers to previous queries. This method is usually computationally intensive, as the general problem of deciding if some sensitive information can be recovered from a set of answers is NP-complete. In addition, there is a potential for a denial of service attack, because the more queries are answered, the fewer new queries can be allowed to preserve privacy.

Cell Suppression In this approach, data is published in tables where each table entry is called a cell. All cells that might cause confidential information to be disclosed (for example when it is derived from only a few individuals) are suppressed. Other cells of non confidential information that might lead to a disclosure of confidential information also have to suppressed (complementary suppression). In [6] a thorough study on cell suppression is presented and shows that cell suppression becomes impractical if an arbitrary complex syntax for queries is allowed.

2 Background

Denning's book [6] contains a chapter on protecting privacy. The treatment is quite extensive, but follow-up research on the topics covered has been meager. More recently (about year 2000) the database community started investigating privacy-preserving data mining. Publications in this area define privacy in many different ways, and many don't even provide a formal definition. Definitions typically describe whether privacy is preserved or not. In our research, we want to measure privacy loss, not just state if privacy loss occurred or not.

In [2] the authors use a measure of privacy as follows. If the original value of a field can be estimated with $c\%$ confidence to lie in the interval $[\alpha_1, \alpha_2]$, then the interval width $(\alpha_2 - \alpha_1)$ defines the amount of privacy at $c\%$ confidence level. For example, with data perturbation, if the added noise is uniformly distributed in an interval of width 2α, then α is the amount of privacy at confidence level 50% and 2α is the amount of privacy at confidence level 100%. This can be explained by the following example.

Consider an attribute X with the density function

$$f_X(x) = \begin{cases} 0.5 & 0 \leq x \leq 1 \\ 0.5 & 4 \leq x \leq 5 \\ 0 & \text{otherwise} \end{cases}$$

Assume that the added noise Y is distributed uniformly between $[-1, 1]$. Then according to the measure proposed, the amount of privacy is 2 at confidence level 100%. Let a large amount of data be available, so that the distribution function is revealed to a high degree of accuracy. Since the perturbing additive is publicly known, the two pieces of information can be combined to determine that if $Z \in [-1, 2]$, then $X \in [0, 1]$; whereas if $Z \in [3, 6]$ then $X \in [4, 5]$.

Thus, in each case the value of X can be localized to an interval of length 1. This means that the actual amount of privacy offered by the perturbing additive Y is at most 1 at confidence level 100%. This method suffers from the fact that a reconstruction algorithm provides a certain amount of knowledge that can be used to guess a data value to a higher level of accuracy.

In [1] the authors present a measure to quantify information loss. Given the perturbed values $z_1, z_2, ..., z_n$, it is not possible to reconstruct the original density function $f_X(x)$ with an arbitrary precision. The greater the variance of the perturbation, the lower the precision in estimating $f_X(x)$. The lack of precision in estimating $f_X(x)$ is referred to as *information loss*. Let $\hat{f}_X(x)$ denote the density function of X as estimated by a reconstruction algorithm. The following metric has been proposed to measure the information loss incurred by a reconstruction algorithm in estimating $f_X(x)$:

$$I(f_X, \hat{f}_X) = \frac{1}{2} E\left[\int_{\omega_X} \left| f_X(x) - \hat{f}_X(x) \right| dx\right]$$

The proposed metric equals half the expected value of $L_1 - norm$ between the original distribution $f_X(x)$ and its estimate $\hat{f}_X(x)$. The information loss $I(f_X, \hat{f}_X)$ lies between 0 and 1; $I(f_X, \hat{f}_X) = 0$ implies perfect reconstruction of $f_X(x)$ and $I(f_X, \hat{f}_X) = 1$ implies that there is no overlap between $f_X(x)$ and its estimate $\hat{f}_X(x)$. The proposed metric

is *universal* in the sense that it can be applied to any reconstruction algorithm since it depends only on the original density $f_X(x)$ and its estimate $\hat{f}_X(x)$.

In [8] a notion of *non-privacy* is presented – a situation which should not be allowed in any reasonable database setting. The authors call a database *non-private* if a computationally bounded adversary can expose a $1-\varepsilon$ fraction of the database entries for all $\varepsilon > 0$. In other words, non-privacy excludes even very weak notions of privacy. The authors then proceed to give a definition of privacy with respect to a bounded adversary with no prior knowledge. They denote the content of a statistical database by $(d_1, \ldots, d_n) \in \{0,1\}^n$. A query $q \subseteq [n]$ is answered by $A(q)$. Their definition of privacy is based on the fact that an adversary should not be able to predict the ith bit, regardless of the content of the rest of the database. The database-adversary is modelled as a game that consists of two phases. In the first phase, the adversary queries the database (adaptively). At the end of this phase, the adversary outputs an index i, of the private function $\pi_i(d_1, \ldots, d_n) = d_i$ it intends to guess. In the second phase, the adversary is given the query-response transcript of the first phase plus all but the ith database entries, and outputs a guess. Privacy is preserved if the adversary fails to guess d_i with high probability.

We like a recent definition ([4]) of privacy loss based on the intuition that privacy is protected when one blends in the crowd. In a nutshell, if an adversary succeeds to produce an approximation of a database record much better than without the statistical release, then privacy is compromised. The authors compare different sanitization techniques using their definition.

Sometimes, compromising privacy is acceptable, provided only a little bit of privacy is lost. We would like to provide a measure of privacy loss. With such a measure, one would be able to not only state that privacy loss occurred, but also state to what extent privacy has been lost. For example, we would like to be able to say the 10 bits of privacy was lost, or that 10% privacy was lost.

Our measure definitions are theoretical and based on information theory. Although our definitions provide a formal measure of privacy loss, it is a stretch to imagine these definitions can be used in practice. Because the measure depends on on the probability distribution for the database and on the outside available knowledge, it is impractical to compute the privacy loss, given a database and a potential statistical release. The main purpose is to provides a uniform setting in which to compare different privacy preserving methods and algorithms.

Some recent work of two of the authors [11] uses a different approach for measuring privacy loss. The motivating example is as follows. If someone's blood pressure can be approximated as a result of a statistical release, a health insurance company may decide to increase the insurance rates for that person. The privacy loss is based on how much potential financial loss could occur to an individual as a result of the statistical release.

This paper is an overview of the results from the Master's thesis [5] of the first author, under the supervision of the second and third author. We have included here only the main definition and a few results from the thesis. For a more elaborate coverage and a number of examples and figures, look for the thesis at *http://digitalcommons.utep.edu/*. A number of our results were also announced at the SCISS05 workshop [12].

3 Defining privacy

Let's consider the case where all the information in the database is sensitive. The other case, when some data is more sensitive than other, is open for further investigations. Since we want to base our measure on information theory, a natural way is to use Kolmogorov complexity.

3.1 Kolmogorov complexity as a measure of information

The Kolmogorov complexity, $K(x)$ of a string x is the length of the shortest binary program to compute x on a universal computer. Intuitively, $K(x)$ represents the minimal amount of information required to generate x by any effective process [10] [14] [3]. The Kolmogorov complexity of a string can be viewed as an absolute and objective quantification of the amount of information in it. The conditional Kolmogorov complexity $K(x|y)$ of x relative to y is defined similarly as the length of the shortest program to compute x if y is furnished as an auxiliary input to the computation.

Kolmogorov complexity is a measure of quantity of information. Suppose the Kolmogorov complexity of a record in the database is 2000 bits. A statistical release that would provide some information about this record would allow us to recover the record with fewer than 2000 bits. This means that the Kolmogorov complexity decreases when we consider the conditional Kolmogorov complexity relative to the statistical release. Intuitively, the privacy loss is proportional to how much the complexity of the record decreases relative to the statistical release. However, consider a database containing the salaries of employees for a company. Suppose someone has a salary which is exactly the average salary of all employees. In this case, according to the naive definition, the statistical release will compromise the privacy of this record almost totally. This is against our intuitive notion of privacy loss. There may be a small privacy loss, but this is really acceptable. Since Kolmogorov complexity does not segregate information, if a released salary average happens to be someone's social security number, the Kolmogorov complexity based definition would consider it a total loss of privacy.

3.2 Defining privacy using entropy

The problem with Kolmogorov complexity is that the information contained in the release may be coincidentally the same as some of the records, but the uncertainty of this fact makes it irrelevant. This means we need to base our definition on probability distributions.

An information source is a mathematical model for a physical entity that produces a succession of symbols. Shannon [13] associated information with the unpredictability of a symbol. The information measure $I(p)$ should satisfy the following properties

- Information is a non-negative quantity: $I(p) \geq 0$.
- If an event has probability 1, we get no information from the occurrence of the event: $I(1) = 0$.
- If two independent events occur, then the information we get from observing the events is the sum of the two information: $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$.

- The information measure should be a continuous and monotonic function of the probability p.

This leads us to the following measure of information

$$I(p) = -\log_b(p) \tag{1}$$

for some positive constant b, where p is the probability associated with the occurrence of an event. As we can see the formula mentioned above satisfies the basic notions of information which are non-negativity, additivity of information when two independent events occur, continuity and zero information gain when the probability of an event occurring is one. The different bases for the logarithm used in 1 lead to different units for our information measure. When $b = 2$, the information measure is *bits* while $b = 10$ results in *digits*. We use *bits* as the unit for measuring information as it seems more intuitive in our approach.

Now, let us assume we have a set of events $E = (e_1, e_2, \ldots, e_n)$ and a set of probabilities (probability distribution) associated with each event $P = (p_1, p_2, \ldots, p_n)$. Shannon defined the entropy of the probability distribution by

$$H(P) = \sum_{i=1}^{n} p_i . \log(1/p_i) \tag{2}$$

3.3 Defining privacy loss

In order to define privacy, we must fix a suitable model to represent our database. The data in many application domains, for example, medical records, financial transactions or employee data, can be represented as data tables. A data table can be seen as a simplification of a relational database, since the latter in general consists of a number of data tables. A formal definition of a database is given below

Definition 1 *A* database *is a triple* $D = (R, A, V_a | a \in A)$ *such that*

- R *is a nonempty finite set of records*
- A *is a nonempty finite set of attributes, and*
- *every attribute* $a \in A$ *is a total function* $a : R \to V_a$*, where* V_a *is the set of values of* a*, called the domain of* a*.*

For example, consider a database consisting of 100 records with the following attributes age, sex and salary. Then, for the database D we have

- $R = \{r_1, r_2, \ldots, r_n\}$.
- $A = \{a_1, a_2, a_3\}$, where a_1, a_2, a_3 imply age, sex and salary repectively.
- $V_{a_1} = \{0, 1, \ldots, 150\}$.
  $V_{a_2} = \{M, F\}$.
  $V_{a_3} = \{0, 100, \ldots, 10000\}$.

Consider the scenario where the database is generated using a random source. The database is assumed to have a fixed number of records n. Let $\mathcal{R} = \{r_1, r_2, \ldots r_n\}$ be the set of records associated with a database and $\mathcal{V} = \{v_1, v_2, \ldots v_k\}$ be the values associated with a record for an attribute $a \in A$. Let $\mathcal{D} = \{d_1, d_2, \ldots d_q\}$ be the set of possible databases with each $d_i \in \mathcal{D}$ having a probability q_i.

We associate successive observations of possible values of a record r_i with a random variable X_i. Let the probability mass associated with the value v_x is p_i i.e

$$P(X_i = v_x) = p_{ix}, \text{ where } \sum_{x=1}^{k} p_{ix} = 1$$

Therefore, the information associated with the outcome

$$X_i = v_x \text{ is } -\lg(p_{ix}).$$

We calculate the entropy of an individual record as

$$H(X_i) = -\sum_{x=1}^{k} p_{ix} \lg(p_{ix}). \tag{3}$$

where $H(X_i)$ is the entropy of the ith record in the database for an attribute $a \in A$.

Let S be any statistics released about the database. After observing S, we can rule out all the possible databases d_i that are inconsistent with S. This reduction in the set of possible databases results in a change in q_i's associated with a database. To maintain consistency, the q_i's are adjusted accordingly such that $\sum q_i = 1$. Let the new possible set of databases be $\mathcal{D}'$. This reduction in the set of possible databases corresponds to a change in the probability mass associated with the value v_x for a record r_i. We get,

$$P(X_i = v_x) = q_i \cdot |t_x|, = p'_{ix}$$

where t_x is the set of records r_i in $\mathcal{D}'$ that has value v_x.

Hence, the entropy of a record after the statistical release is given by

$$H'(X_i) = -\sum_{x=1}^{k} p'_{ix} \lg(p'_{ix}).$$

$H(X_i) - H'(X_i)$ gives the loss of information for record r_i in the database. Similarly we can compute the loss of information for every record in the database. We now formally define privacy loss as follows:

Definition 2 *Let $H(X_1), H(X_2), \cdots, H(X_n)$ be the entropies associated with each record in the database and $H'(X_1), H'(X_2), \cdots, H'(X_n)$ be the entropies associated with the records after a statistical release. We define* privacy loss *as the maximum loss of information associated with a record in the database given by,*

$$\mathit{PrivacyLoss}(\mathcal{P}) = \max_{i \leq n} \left(H(X_i) - H'(X_i)\right)$$

For a given statistical release, our definition relates privacy loss to the maximum loss of information associated with a record in the database. We further illustrate this concept using the example below

Consider a one dimensional database consisting of 3 salaries. A salary in the database may take values from the following set $V = \{100, 200, 300\}$. *We assume that the records in the database are randomly generated resulting in a uniform distribution of the salaries. The user queries the database for its average. We consider two cases:*

Case a: Let the result of the query be 200. We need to determine the privacy loss associated with this release.

Prior to the statistical release, the entropy of a record in the database is given by

$$\begin{aligned} H(X_i) &= -\sum_{i=1}^{3} p_i \cdot \lg(p_i), \text{ where } p_i = \frac{1}{3} \\ &= 1.585 bits \end{aligned}$$

The possible set of databases prior to releasing the average was k^n where k is the number of values that a record can take and n is the number of records. Therefore, the possible set of databases $D = 27$.

After releasing the average, we can ignore all databases where the average is not 200. This narrows the set of possible databases to only 7 of them.

Calculations for this new set of possible databases lead to an entropy of 1.439 *bits*. Since, the entropies for all the records are equal, the loss of information associated with each record is the same. From 2 the privacy loss after releasing the average is $1.585 - 1.439 = 0.146$ *bits*.

The reason for a small loss of information in this example is because the average returned by the query reflects the statistical mean for the three salaries in the database following a uniform distribution.

Case b - Let the result of the query be 100.

From the previous case, we know that the entropy of a record prior to the statistical release is 1.585 *bits*. Because for this case there is only one possible database with an average of 100, the entropy after the statistical release is 0. The loss of information is $1.585 - 0 = 1.585$ *bits* indicating total privacy loss which supports our intuition of exact compromise.

3.4 Application to binary databases

The simplest case to analyse is that of a binary database, a database where every attribute has a value in $\{0, 1\}$. The following results apply to binary databases. Proofs and illustrative graphs are found in [5].

Theorem 1 *Let* D *be a 1-D binary database consisting of* n *records. We assume that each record in the database is generated independantly and randomly. Each value* 0 *or* 1 *occurs with equal probability. Let* S *be a statistical release that releases the average of the data in the database. Then, the privacy loss is given by*

$$\text{Privacy Loss } \mathcal{P}(X_i) = \max_{i \leq n} \left[1 - \left(\frac{k}{n} \lg \left(\frac{n}{k} \right) + \left(\frac{n-k}{n} \right) \lg \left(\frac{n}{n-k} \right) \right) \right] \tag{4}$$

where, k is the sum of the data in the database.

Theorem 2 Let D be a 1-D database consisting of n records. We assume that the database is generated randomly, where each value 0 or 1 occurs with equal probability. Let S be a statistical release that releases the average of the data in the database and $\mathcal{P}(X)$ be the privacy loss due to S. Then, the expected privacy loss is given by

$$\text{Expected privacy loss} = \sum_{k=0}^{n} \frac{1}{2^n} \binom{n}{k} \left[1 - \left(\frac{k}{n} \lg\left(\frac{n}{k}\right) - \left(\frac{n-k}{n}\right) \lg\left(\frac{n}{n-k}\right)\right)\right] \qquad (5)$$

where k is the sum of the salaries in a database.

4 Remaining problems and future work

There are still some problems with our definition. Consider a release which publishes the whole database after it has been encrypted with a public key. Because our definition is based in classical entropy, it does not consider computation time in the encoding of information. In a secure cryptosystem, computing the secret key from the public key is intractable. But theoretically, the secret key is computable from the public key. So, according to our definition, there is a total loss of privacy in this case. It is however generally accepted that encrypted data is useless without the decrypting key, so in this case we should conclude that the privacy has been protected.

To solve this problem, we need to consider effective entropy. This is a version of entropy that takes into consideration the time to compute the coding and decoding of the information. The first definition and use of this concept was by Yao [15].

Another avenue of research is to consider the concept that not all bits of equally sensitive. For example, discovering the first two bits of someone's salary or blood pressure may be considered much worse than discovering the last two bits. The different approach described in [11] is a way to address this problem.

References

[1] D. Agrawal and C. Agrawal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principle of database systems*, pages 247–255, 2001.

[2] R. Agrawal and R. Srikant. Privacy preserving data mining. In *In Proceedings of the ACM SIGMOD*, pages 439–450, 2000.

[3] G.J. Chaitin. On the length of programs for computing finite binary sequences. In *J. Assoc. Comp. Mach.*, volume 13, pages 547–569, 1966.

[4] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In *Theory of Cryptography Conference*, 2005.

[5] V. Chirayath. Using entropy as a measure of privacy loss in statistical databases. Master's thesis, University of Texas at El Paso, 2004.

[6] D. Denning. *Cryptography and data security.* Addison-WesleyPrentice-Hall, 1982.

[7] D. Denning and J. Schlorer. A fast procedure for finding a tracker in a statistical database. In *ACM Transactions on Database Systems*, volume 5, pages 88–102, 1980.

[8] Irit Dinur and Kobbi Nissim. Revealing Information while Preserving Privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principle of database systems*, pages 202–210, 2003.

[9] Klosek Jacqueline. *Data Privacy in the information age.* Greenwood Pub Group, 2000.

[10] A.N. Kolmogorov. Three approaches to the definition of the concept 'quantity of information'. In *Probl. Inform. Transm.*, volume 1, pages 1–7, 1965.

[11] L. Longpré and V. Kreinovich. How to measure loss of privacy. Technical report, University of Texas at El Paso, 2006. http://www.cs.utep.edu/vladik/2006/tr06-24.pdf.

[12] Luc Longpré, Vladik Kreinovich, Eric Freudenthal andMartine Ceberio, François Modave, Neelabh Baijal, Wei Chen, Vinod Chirayath, Gang Xiang, and J. Ivan Vargas. Privacy: Protecting, processing, and measuring loss. In *Abstracts of the 2005 South Central Information Security Symposium SCISS'05*, page 2, Austin, TX, April 2005.

[13] Claude E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, July 1948.

[14] R. J. Solomonoff. A formal of inductive inference. *Information and Control*, 1964.

[15] A. Yao. Theory and applications of trapdoor functions. In *Proc. 23rd IEEE Symposium on Foundations of Computer Science*, pages 80–91, 1982.