

On the Design and Quantification of Privacy Preserving Data Mining Algorithms

Dakshi Agrawal
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
agrawal@us.ibm.com

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
charu@us.ibm.com

ABSTRACT

The increasing ability to track and collect large amounts of data with the use of current hardware technology has led to an interest in the development of data mining algorithms which preserve user privacy. A recently proposed technique addresses the issue of privacy preservation by perturbing the data and reconstructing distributions at an aggregate level in order to perform the mining. This method is able to retain privacy while accessing the information implicit in the original attributes. The distribution reconstruction process naturally leads to some loss of information which is acceptable in many practical situations. This paper discusses an Expectation Maximization (EM) algorithm for distribution reconstruction which is more effective than the currently available method in terms of the level of information loss. Specifically, we prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. We show that when a large amount of data is available, the EM algorithm provides robust estimates of the original distribution. We propose metrics for quantification and measurement of privacy-preserving data mining algorithms. Thus, this paper provides the foundations for measurement of the effectiveness of privacy preserving data mining algorithms. Our privacy metrics illustrate some interesting results on the relative effectiveness of different perturbing distributions.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications—*Data Mining*

General Terms

Algorithms, Experimentation, Theory

1. INTRODUCTION

In recent years, the progress of hardware technology has made it easy to store and process large amounts of transactional information. For example, even simple transactions

of everyday life such as using the phone or credit-cards are recorded today in an automated way. A large amount of such information is often specific to individual users. Depending upon the nature of the information, users may not be willing to divulge the individual values of records. In particular, data mining techniques are considered a challenge to privacy preservation due to their natural tendency to use sensitive information about individuals. Some interesting discourses on the nature of privacy in the context of recent trends in information technology may be found in [3, 5, 10, 11, 12, 13]. This has led to a considerable amount of focus on privacy preserving data collection and mining methods [1, 6, 7, 8, 15]. An innovative approach for privacy preserving data mining was recently proposed in [1]. This technique relies on two facts:

- Users are not equally protective of all values in their records. Thus, users may be willing to provide modified values of certain fields by the use of a (publicly known) perturbing random distribution. This modified value may be generated using a custom code or a browser plug-in.
- Data mining problems do not necessarily require individual records, but only distributions. Since the perturbing distribution is known, it can be used to reconstruct *aggregate* distributions, i.e. the probability distribution of the data set. In many cases, data mining algorithms can be developed which use the probability distributions rather than individual records. An example of a classification algorithm which uses such aggregate information is discussed in [1].

Specifically, let us consider a set of n original data values $x_1 \dots x_n$. These are modeled in [1] as n independent values, each drawn from the same data distribution as the random variable X . In order to create the perturbation, we generate n independent values $y_1 \dots y_n$, each with the same distribution as the random variable Y . Thus, the perturbed values of the data are given by $z_1 = x_1 + y_1, \dots, z_n = x_n + y_n$. In order to protect privacy, only the perturbed values are provided rather than the original data. Given these values, and the (publicly known) density function $f_Y(y)$ for Y , an iterative algorithm was proposed in [1] to estimate the density function $f_X(x)$ for X . (We shall henceforth refer to this technique as the AS algorithm.)

It may be noted that the exact distribution for X is impossible to reconstruct for a given data set. The greater the level of perturbation, the less likely we are to be able to estimate the data distributions effectively. On the other hand, larger perturbations also lead to a greater amount of privacy. Thus, there is a trade-off between loss of information and privacy. Furthermore, the exact accuracy level in estimating the data distribution is sensitive to the reconstruction algorithm. A given reconstruction algorithm may not always converge; and even when it converges, there is no guarantee that it provides a reasonable estimate of the original distribution. For example, the convergence behavior of the AS algorithm has not been discussed in [1]. In this paper, we develop an effective reconstruction algorithm which provably converges to the maximum-likelihood estimate of the data distribution. We discuss the quantification of the information-privacy tradeoff. Furthermore, we propose theoretically sound metrics to measure information loss and privacy, thus providing a foundation to quantify the performance of privacy preserving data mining algorithms.

This paper is organized as follows. In the sections 2 and 3, we discuss the techniques for quantification of privacy and information loss for reconstruction algorithms. In section 4, we will derive an expectation maximization algorithm for distribution reconstruction and show some of its nice convergence properties. The empirical results are presented in section 5. In section 6, we present the conclusions and summary.

1.1 Contributions of this paper

In this paper, we develop optimal algorithms and models based on the interesting perturbation approach proposed in [1]. We propose a reconstruction algorithm for privacy preserving data mining, which not only converges but does so to the maximum likelihood estimate of the original distribution. This is the theoretical best that any reconstruction algorithm can achieve. This effectively means that when a large amount of data is available, the expectation maximization algorithm can reconstruct the distribution with little or almost no information loss.

We examine the problem of quantifying privacy and information loss. For example, the method in [1] quantifies privacy without taking into account the additional knowledge that a user may obtain from the reconstructed (aggregate) distribution. We propose a privacy metric which takes into account the fact that *both* the perturbed individual record and the aggregate distribution are available to the user to make more accurate guesses about the possible values of the record. This privacy metric is based on the concept of *mutual information* between the original and perturbed records. Thus, the metrics proposed by this paper also provide a foundation for testing the effectiveness of privacy-preserving data mining algorithms in the future.

We use these proposed metrics to quantify the effects of data and perturbation parameters. Our empirical results show some simple trends of privacy-preserving data mining algorithms: (1) With increasing perturbation, the privacy level increases, but the effectiveness of reconstruction algorithms decreases. This leads to a privacy-information loss trade-off curve. (2) With increasing amount of data available,

the EM-reconstruction algorithm is able to approximate the original distribution to a very high degree of precision (3) Our metrics also provides somewhat different results to those presented in [1] about the relative effectiveness of different perturbing distributions.

2. QUANTIFICATION OF PRIVACY

The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated. The work in [1] uses a measure that defines privacy as follows: If the original value 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, if the perturbing additive 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%. However, this simple method of determining privacy can be subtly incomplete in some situations. This can be best explained by the following example.

Example 1. Consider an attribute X with the density function $f_X(x)$ given by:

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

Assume that the perturbing additive Y is distributed uniformly between $[-1, 1]$. Then according to the measure proposed in [1], the amount of privacy is 2 at confidence level 100%.

However, after performing the perturbation and subsequent reconstruction, the density function $f_X(x)$ will be approximately revealed. Let us assume for a moment that a large amount of data is available, so that the distribution function is revealed to a high degree of accuracy. Since the (distribution of the) perturbing additive is publically 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%. We use the qualifier ‘at most’ since X can often be localized to an interval of length less than one. For example, if the value of Z happens to be -0.5 , then the value of X can be localized to an even smaller interval of $[0, 0.5]$. ■

This example illustrates that the method suggested in [1] does not take into account the distribution of original data. In other words, the (aggregate) reconstruction of the attribute value also provides a certain level of knowledge which can be used to guess a data value to a higher level of accuracy. To accurately quantify privacy, we need a method which takes such side-information into account.

A search for the correct measure to quantify privacy turns out to be quite elusive. The concept of privacy is easiest to grasp in the context of uniformly distributed random variables. Intuitively, a random variable distributed uniformly between $[0, 1]$ has half as much privacy as a random variable distributed uniformly between $[0, 2]$. In general, we expect that if $f_B(x) = 2f_A(2x)$, then B offers half as much privacy as A . Furthermore, we expect that if a sequence of random variables A_n , $n = 1, 2, \dots$ converges to another random variable B , then privacy inherent in A_n should also converge to the privacy inherent in B . Ideally, we would like to find a privacy measure that satisfies such intuitive notions.

In this paper, we propose a privacy measure based on the *differential entropy* of a random variable. The differential entropy $h(A)$ of a random variable A is defined as follows:

$$h(A) = - \int_{\Omega_A} f_A(a) \log_2 f_A(a) da \quad (2)$$

where Ω_A is the domain of A . It is well-known that $h(A)$ is a measure of uncertainty inherent in the value of A [9]. It can be easily seen that for a random variable U distributed uniformly between 0 and a , $h(U) = \log_2(a)$. For $a = 1$, $h(U) = 0$. Thus, random variables with less uncertainty than a uniform distribution in $[0, 1]$ have negative differential entropy, while random variables with more uncertainty have positive differential entropy.

We propose $2^{h(A)}$ as a measure of privacy inherent in the random variable A and denote it by $\Pi(A)$. Thus, a random variable U distributed uniformly between 0 and a has privacy $\Pi(U) = 2^{\log_2(a)} = a$. For a general random variable A , $\Pi(A)$ denote the length of the interval, over which a uniformly distributed random variable has the same uncertainty as A . Thus if $\Pi(A) = 2$, then A has as much privacy as a random variable distributed uniformly in an interval of length 2. This measure also satisfies all intuitive notions described above.

Now, we will introduce the notion of *conditional* privacy which takes into account the additional information available in the perturbed values. Given a random variable B , the *conditional* differential entropy of A is defined as follows:

$$h(A|B) = - \int_{\Omega_{A,B}} f_{A,B}(a, b) \log_2 f_{A|B=b}(a) da db \quad (3)$$

Thus, the average conditional privacy of A given B is $\Pi(A|B) = 2^{h(A|B)}$. This motivates the following metric $\mathcal{P}(A|B)$ for the conditional privacy loss of A , given B :

$$\begin{aligned} \mathcal{P}(A|B) &= 1 - \Pi(A|B)/\Pi(A) = 1 - 2^{h(A|B)}/2^{h(A)} \\ &= 1 - 2^{-I(A;B)}, \end{aligned} \quad (4)$$

where $I(A; B) = h(A) - h(A|B) = h(B) - h(B|A)$. $I(A; B)$ is also known as the *mutual information* between the random variables A and B . Since $2^{h(A|B)}/2^{h(A)}$ is the ratio of privacy of A after and before revealing B , $\mathcal{P}(A|B)$ is the fraction of privacy of A which is lost by revealing B .

As an illustration, let us reconsider Example 1 given above. In this case, the differential entropy of X is given by:

$$\begin{aligned} h(X) &= - \int_{\Omega_X} f_X(x) \log_2 f_X(x) dx \\ &= - \int_0^1 0.5 \log_2 0.5 dx - \int_4^5 0.5 \log_2 0.5 dx \\ &= 1 \end{aligned} \quad (5)$$

Thus the privacy of X , $\Pi(X) = 2^1 = 2$. In other words, X has as much privacy as a random variable distributed uniformly in an interval of length 2. The density function of the perturbed value Z is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(\nu) f_Y(z - \nu) d\nu.$$

Using $f_Z(z)$, we can compute the differential entropy $h(Z)$ of Z . It turns out that $h(Z) = 9/4$. Therefore, we have:

$$I(X; Z) = h(Z) - h(Z|X) = 9/4 - h(Y) = 9/4 - 1 = 5/4$$

Here, the substitution $h(Z|X) = h(Y)$ in the second equality follows from the fact that X and Y are independent and $Z = X + Y$. In the third equality, we have substituted $h(Y) = \log_2 2 = 1$. Thus the fraction of privacy loss in this case is $\mathcal{P}(X|Z) = 1 - 2^{-5/4} = 0.5796$. As a result after revealing Z , X has privacy $\Pi(X|Z) = \Pi(X) \times (1 - \mathcal{P}(X|Z)) = 2 \times (1.0 - 0.5796) = 0.8408$. This value is less than 1, since X can be localized to an interval of length less than one for many values of Z .

3. QUANTIFICATION OF INFORMATION LOSS

Given the perturbed values z_1, z_2, \dots, z_N , it is (in general) 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 $f_X(x)$.

We refer the lack of precision in estimating $f_X(x)$ as *information loss*. In this section, we will consider how to quantify information loss. Note that the work in [1] uses an application dependent approach to measure the information loss. For example, for a classification problem, the inaccuracy in distribution reconstruction is measured by examining the effects on the misclassification rate.

Let $\hat{f}_X(x)$ denote the density function of X as estimated by a reconstruction algorithm. We propose the metric $\mathcal{I}(f_X, \hat{f}_X)$ to measure the information loss incurred by a reconstruction algorithm in estimating $f_X(x)$:

$$\mathcal{I}(f_X, \hat{f}_X) = \frac{1}{2} E \left[\int_{\Omega_X} |f_X(x) - \hat{f}_X(x)| dx \right] \quad (6)$$

Thus the proposed metric equals half the expected value of L_1 -norm between the original distribution $f_X(x)$ and its estimate $\hat{f}_X(x)$. Note that information loss $\mathcal{I}(f_X, \hat{f}_X)$ lies between 0 and 1; $\mathcal{I}(f_X, \hat{f}_X) = 0$ implies perfect reconstruction of $f_X(x)$ and $\mathcal{I}(f_X, \hat{f}_X) = 1$ implies that there is no overlap between $f_X(x)$ and its estimate $\hat{f}_X(x)$ (see Figure 3). The proposed metric is *universal* in the sense that it can

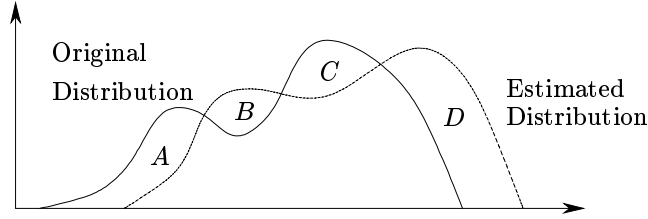


Figure 1: Illustration of the information loss metric: In this case, the estimated distribution is somewhat shifted from the original distribution. The proposed information loss metric is related to the amount of mismatch between two distribution in terms of area. Specifically, it equals half the sum of the areas denoted above by A , B , C , and D . It is also equal to $1 - \alpha$, where α is the area shared by both distributions.

be applied to any reconstruction algorithm since it depends only on the original density $f_X(x)$, and its estimate $\hat{f}_X(x)$. We advocate the use of a universal metric since it is independent of the particular data mining task at hand, and therefore facilitates absolute comparisons between disparate reconstruction algorithms.

4. AN EM ALGORITHM FOR EFFECTIVE DISTRIBUTION RECONSTRUCTION

Assume that $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ are realizations of N independent and identically distributed random variables $\mathbf{X} = \{X_1, X_2, \dots, X_N\}$, each with the density function $f_X(x)$. These realizations constitute the original data. Further assume that $\mathbf{y} = \{y_1, y_2, \dots, y_N\}$ are realizations of N independent and identically distributed random variables $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_N\}$, each with the density function $f_Y(y)$. These realizations constitute the perturbations to the original data. Given the N perturbed values $\mathbf{z} = \{z_1, z_2, \dots, z_N\}$, $z_i = x_i + y_i$, and the density function $f_Y(y)$, we would like to reconstruct $f_X(x)$. We denote the perturbed random variables by $\mathbf{Z} = \{Z_1, \dots, Z_N\}$.

Note that since the function $f_X(x)$ is defined over a continuous domain, we need to parameterize and discretize it for the purpose of any numerical estimation method. We assume that the data domain Ω_X can be discretized into K intervals $\Omega_1, \dots, \Omega_K$, where $\cup_{i=1}^K \Omega_i = \Omega_X$. Let $m_i = m(\Omega_i)$ be the length of the interval Ω_i . We assume that $f_X(x)$ is constant over Ω_i and the corresponding density function value is equal to θ_i . Such a form will restrict $f_X(x)$ to a class parameterized by the finite set of parameters $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$. In order to explicitly denote the parametric dependence of the density function on Θ we will use the notation $f_{X;\Theta}(x)$ for the density function of X . Under these assumptions, we have

$$f_{X;\Theta}(x) = \sum_{i=1}^K \theta_i I_{\Omega_i}(x),$$

where $I_{\Omega_i}(x) = 1$ if $x \in \Omega_i$ and 0 otherwise. Since $f_{X;\Theta}(x)$ is a density, it follows that $\sum_{i=1}^K \theta_i m(\Omega_i) = 1$. By choosing K large enough, density functions of the form discussed above can approximate any density function with arbitrary precision.

After this parameterization, the algorithm will proceed to

estimate Θ , and thereby determine $\hat{f}_{X;\Theta}(x)$. Let

$$\hat{\Theta} = \{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_K\}$$

be the estimate of these parameters produced by the reconstruction algorithm.

Given a set of observations $\mathbf{Z} = \mathbf{z}$, we would ideally like to find the *maximum-likelihood estimate* (MLE)

$$\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}} \ln f_{\mathbf{Z};\Theta}(\mathbf{z}).$$

The MLE has many attractive properties such as consistency, asymptotic unbiasedness, and asymptotic minimum variance among unbiased estimates [14]. However, it is not always possible to find $\hat{\Theta}_{ML}$ directly, and this turns out to be the case with the $f_{\mathbf{Z};\Theta}(\mathbf{z})$.

In order to achieve this goal, we will derive a reconstruction algorithm which fits into the broad framework of Expectation Maximization (EM) algorithms. The algorithm proceeds as if a more comprehensive set of data, say $\mathbf{D} = \mathbf{d}$ is observable and maximizes $\ln f_{\mathbf{D};\Theta}(\mathbf{d})$ over all values of Θ (M-step). Since \mathbf{d} is in fact unavailable, it replaces $\ln f_{\mathbf{D};\Theta}(\mathbf{d})$ by its conditional expected value given $\mathbf{Z} = \mathbf{z}$ and the current estimate of Θ (E-Step). The \mathbf{D} is chosen to make E-step and M-step easy to compute.

In this paper, we propose the use of $\mathbf{X} = \mathbf{x}$ as the more comprehensive set of data. As shown in the next section, this choice results in a computationally efficient algorithm. More formally, we define a Q function as follows:

$$Q(\Theta, \hat{\Theta}) = E \left[\ln f_{\mathbf{X};\Theta}(\mathbf{X}) \mid \mathbf{Z} = \mathbf{z}; \hat{\Theta} \right] \quad (7)$$

Thus, $Q(\Theta, \hat{\Theta})$ is the expected value of $\ln f_{\mathbf{X};\Theta}(\mathbf{X})$ computed with respect $f_{\mathbf{X}|\mathbf{Z}=\mathbf{z};\hat{\Theta}}$, the density of \mathbf{X} given $\mathbf{Z} = \mathbf{z}$ and parameter vector $\hat{\Theta}$. After the initialization of Θ to a nominal value Θ^0 , the EM algorithm will iterate over the following two steps:

1. E-step: Compute $Q(\Theta, \Theta^k)$.
2. M-step: Update $\Theta^{k+1} = \operatorname{argmax}_{\Theta} Q(\Theta, \Theta^k)$.

The above discussion provides the general framework of EM algorithms; the actual details of the E-step and M-steps

require a derivation which is problem specific. Similarly, the precise convergence properties of an EM algorithm are rather sensitive to the problem and its corresponding derivation. In the next subsection, we will derive the EM algorithm for the reconstruction problem and show that the resulting EM-algorithm has desirable convergence properties.

4.1 Derivation of EM Reconstruction Algorithm

THEOREM 4.1. *The value of $Q(\Theta, \hat{\Theta})$ during the E-step of the reconstruction algorithm is given by:*

$$Q(\Theta, \hat{\Theta}) = \sum_{i=1}^K \psi_i(\mathbf{z}; \hat{\Theta}) \ln \theta_i,$$

where

$$\psi_i(\mathbf{z}; \hat{\Theta}) = \hat{\theta}_i \sum_{j=1}^N \frac{\Pr(Y \in z_j - \Omega_i)}{f_{Z; \hat{\Theta}}(z_j)},$$

and $\nu \in z_j - \Omega_i$ if $z_j - \nu \in \Omega_i$.

PROOF. See appendix. ■

In the next proposition, we calculate the value of Θ that maximizes $Q(\Theta, \hat{\Theta})$.

THEOREM 4.2. *The value of Θ which maximizes $Q(\Theta, \hat{\Theta})$ during the M-step of the reconstruction algorithm is given by:*

$$\theta_i = \frac{\psi_i(\mathbf{z}; \hat{\Theta})}{m_i N},$$

where

$$\psi_i(\mathbf{z}; \hat{\Theta}) = \hat{\theta}_i \sum_{j=1}^N \frac{\Pr(Y \in z_j - \Omega_i)}{f_{Z; \hat{\Theta}}(z_j)}.$$

PROOF. Note that $\sum_{i=1}^K m_i \theta_i = 1$. Hence, we use the Lagrange multiplier method to find the maxima. We write down the Lagrange multiplier function as follows:

$$L(\Theta, \lambda) = \sum_{i=1}^K \psi_i(\mathbf{z}; \hat{\Theta}) \ln \theta_i + \lambda \left(\sum_{j=1}^K m_j \theta_j - 1.0 \right) \quad (8)$$

The Lagrange constraints for the above function are $\partial L / \partial \theta_i = 0$ and $\partial L / \partial \lambda = 0$. The corresponding conditions are $\theta_i = -\frac{\psi_i(\mathbf{z}; \hat{\Theta})}{\lambda m_i}$ and $\sum_{j=1}^K m_j \theta_j = 1.0$ respectively. Eliminating the Lagrange multiplier λ from these conditions, we get $\theta_i = \frac{\psi_i(\mathbf{z}; \hat{\Theta})}{m_i \sum_{i=1}^K \psi_i(\mathbf{z}; \hat{\Theta})}$.

At this stage, we only need to evaluate the denominator of

the above value for θ_i . To this effect, we note that:

$$\begin{aligned} \sum_{l=1}^K \psi_l(\mathbf{z}; \hat{\Theta}) &= \sum_{l=1}^K \hat{\theta}_l \sum_{j=1}^N \frac{\Pr(Y \in z_j - \Omega_l)}{f_{Z; \hat{\Theta}}(z_j)} \\ &= \sum_{j=1}^N \frac{\sum_{l=1}^K \hat{\theta}_l \Pr(Y \in z_j - \Omega_l)}{f_{Z; \hat{\Theta}}(z_j)} \\ &= \sum_{j=1}^N \frac{f_{Z; \hat{\Theta}}(z_j)}{f_{Z; \hat{\Theta}}(z_j)} = N \end{aligned} \quad (9)$$

In order to derive (9), we note that the density function of Z can be obtained by observing that $Z = X + Y$, and X and Y are independent. This implies that:

$$\begin{aligned} f_{Z; \hat{\Theta}}(z) &= \int f_X(\nu) f_Y(z - \nu) d\nu \\ &= \sum_{i=1}^K \int_{\Omega_i} \hat{\theta}_i f_Y(z - \nu) d\nu \\ &= \sum_{i=1}^K \hat{\theta}_i \Pr(Y \in z - \Omega_i) \end{aligned} \quad (10)$$

The result follows. ■

Now, we are in a position to describe the EM algorithm for the reconstruction problem.

4.2 EM Reconstruction Algorithm

1. Initialize $\theta_i^0 = \frac{1}{K}$, $i = 1, 2, \dots, K$; $k = 0$;
2. Update Θ as follows: $\theta_i^{(k+1)} = \frac{\psi_i(\mathbf{z}; \Theta^k)}{m_i N}$;
3. $k = k + 1$;
4. If *not termination-criterion* then return to Step 2.

The termination criterion for this method is based on how much Θ^k has changed since the last iteration. The exact threshold at which the two distributions are deemed to be almost the same for the purpose of convergence could either be heuristically set or be based on a more rigorous conditions such as the one indicated in [4].

4.3 Convergence Properties of EM Reconstruction Algorithm

Let us denote the log-likelihood function as

$$\mathcal{L}_\nu(\Theta) = \log f_{Z; \Theta}(\nu) = \sum_{j=1}^K \log f_{z_j, \Theta}(\nu_j).$$

Ideally, we would like the sequence generated by the EM algorithm $\Theta_0, \Theta_1, \Theta_2, \dots$, to converge to the MLE $\hat{\Theta}_{ML} = \operatorname{argmax}_{\Theta} \mathcal{L}_\nu(\Theta)$. Let \mathcal{C} be the set $\{\Theta : 0 \leq \theta_i, \sum_{i=1}^K m_i \theta_i = 1\}$.

PROPOSITION 4.1. *$\mathcal{L}_\nu(\Theta)$ is strictly concave and has a unique global maxima over \mathcal{C} .*

PROOF. It is easy to see that the set \mathcal{C} is convex, that is, if $\Theta_1, \Theta_2 \in \mathcal{C}$, then, $\lambda \Theta_1 + (1 - \lambda) \Theta_2 \in \mathcal{C}$, for any $\lambda \in [0, 1]$.

Since $f_{Z_j; \Theta}(\nu_j)$ is a linear function of Θ (see (10)). Since \log is strictly concave, it follows that $\log f_{Z_j; \Theta}(\nu_j)$ is also strictly concave. It is easy to verify that the sum of strictly concave functions is again strictly concave, and therefore $\mathcal{L}_\nu(\Theta) = \sum_{i=1}^K \log f_{Z_j; \Theta}(\nu_j)$ is strictly concave. Strictly concave functions can have at most one global maxima over a convex set [2]. Since \mathcal{C} is closed and $\mathcal{L}_\nu(\Theta)$ is continuous, $\mathcal{L}_\nu(\Theta)$ has at least one global maxima over \mathcal{C} . It follows that $\mathcal{L}_\nu(\Theta)$ has a unique global maxima over \mathcal{C} . ■

In order to prove the convergence of the EM algorithm, we use a corollary due to Wu [16]. The following theorem paraphrases this corollary:

THEOREM 4.3. *Suppose that $\mathcal{L}_\nu(\Theta)$ is unimodal in \mathcal{C} with $\hat{\Theta}_{ML}$ being the only stationary point and that $\partial Q(\Theta, \hat{\Theta})/\partial \Theta$ is continuous in Θ and $\hat{\Theta}$. Then, any EM sequence $\{\Theta^{(k)}\}$ converges to the unique MLE $\hat{\Theta}_{ML}$.*

Using this theorem in conjunction with Proposition 4.1, we can immediately derive the desirable convergence property of the EM algorithm.

THEOREM 4.4. *The EM sequence $\{\Theta^{(k)}\}$ for the reconstruction algorithm converges to the unique MLE $\hat{\Theta}_{ML}$.*

PROOF. The unimodality of $\mathcal{L}_\nu(\Theta)$ is implied by its strict concavity. It is easy to see from Theorem 4.1 that $Q(\Theta, \hat{\Theta})$ has derivatives continuous in both arguments. It follows that the EM algorithm described above will converge to the maximum-likelihood estimate. ■

The above results lead to the following desirable property of the EM algorithm.

OBSERVATION 4.1. *When there is a very large number of data observations, then the EM algorithm provides zero information loss.*

This is because as the number of observations increases, $\hat{\Theta}_{ML} \Rightarrow \Theta$. Therefore, the original and estimated distribution become the same (subject to the discretization needed for any numerical estimation algorithm), resulting in zero information loss. We will show that for data sets with as few as 20000 points the EM algorithm is able to provide less than 0.5% information loss for reasonably large perturbations. In the next section, we will also illustrate the qualitative advantages of the EM algorithm over the AS algorithm.

5. EMPIRICAL RESULTS

In this section, we present some interesting trends of the privacy-preserving reconstruction algorithms. It turns out that the AS algorithm is quite robust, and in an average case the performance of the AS algorithm is almost competitive to the performance of the EM algorithm. However, in the worst case scenario, the EM reconstruction algorithm

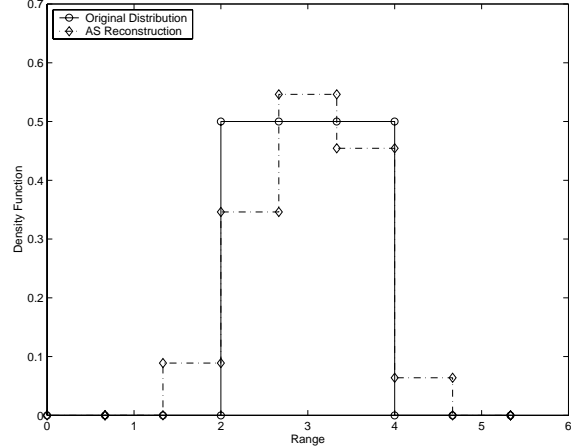


Figure 2: Reconstructed Uniform Distribution (AS Algorithm)

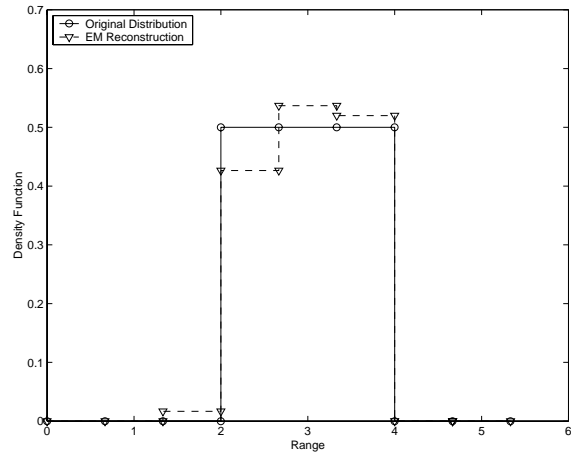


Figure 3: Reconstructed Uniform Distribution (EM Algorithm)

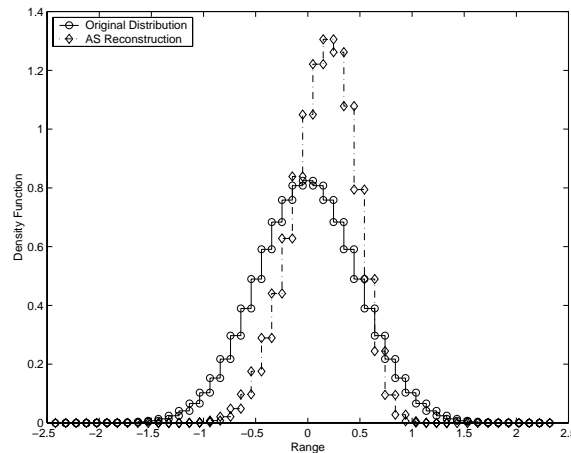


Figure 4: Reconstructed Gaussian Distribution (AS Algorithm)

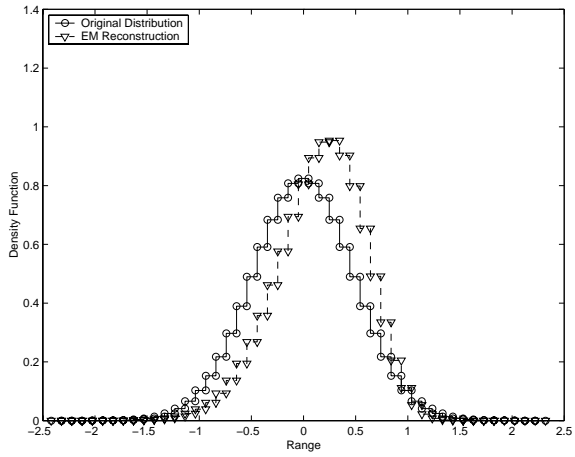


Figure 5: Reconstructed Gaussian Distribution (EM Algorithm)

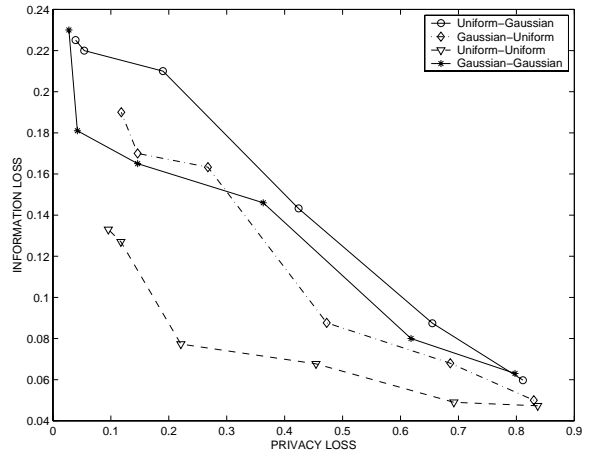


Figure 8: The Tradeoff between Information Loss and Privacy

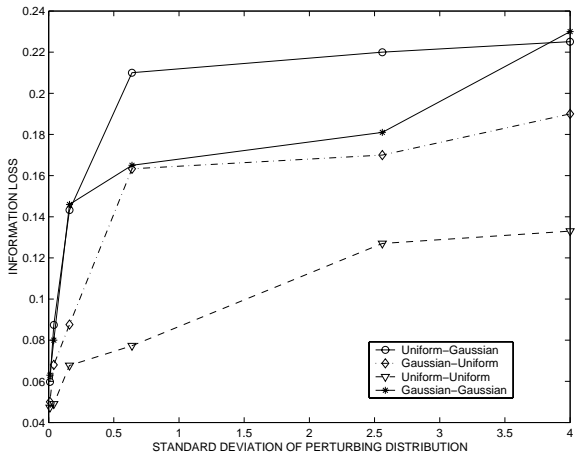


Figure 6: Information Loss with Standard Deviation of Perturbing Distribution

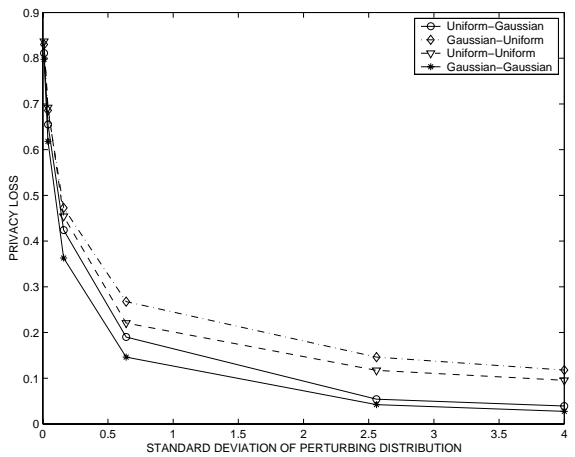


Figure 7: Privacy Loss with Standard Deviation of Perturbing Distribution

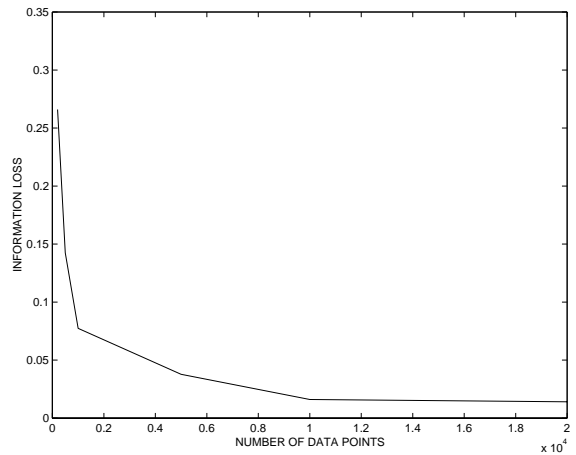


Figure 9: Information Loss with Number of Data Points (Constant Perturbation)

is able to reconstruct the data distribution more effectively than the AS algorithm¹. In Figures 2 and 3, we have illustrated one such case in which we reconstructed a uniform distribution with the use of AS algorithm and EM algorithm respectively. In this case, the original data contains 500 points which are uniformly distributed in the range $[2, 4]$. We added uniformly distributed noise (in the range $[-1, 1]$) in order to perturb the original data. In Figure 2, we have shown the reconstructed distribution obtained by the AS algorithm. The corresponding level of information loss is 13.29%. On the other hand, with the use of the EM algorithm, the information loss level is only 4.9%. The corresponding distribution is illustrated in Figure 3. It is clear that the EM algorithm outperforms the AS method.

In Figures 4 and 5, we have shown a case in which we reconstructed a gaussian distribution with the help of the AS algorithm and the EM algorithm respectively. The original gaussian distribution had the same amount of inherent privacy as the uniformly distributed data in the previous example. This corresponds to a variance of $2/\pi e$. The perturbing distribution is gaussian with variance 1. We generated a total of $N = 500$ data points. In the case of the EM algorithm, we found that the level of information loss was 17.9%, whereas for the AS algorithm, the level of information loss was as high as 26.5%. This difference shows up in the Figure 4 as a higher amount of mismatch with the original distribution as compared to the mismatch produced by the EM algorithm.

In Figure 6, we plot the information loss with increasing standard deviation of the perturbing distribution. These results are presented for four different combinations of the original and perturbing distribution using the uniform and gaussian distributions in each case. The variance of the original gaussian distribution was $2/\pi e$, whereas the range of the original distribution was $[-1, 1]$. This choice of parameters ensured that the uniform and gaussian distributions had the same amount of inherent privacy. In each case, there were $N = 500$ data points. As expected, the amount of information loss grows with the level of perturbation. However, in order to accurately determine which combination of distributions result in the greatest loss of information at a given level of privacy loss, we also need to characterize the dependence of privacy loss on the standard deviation of the perturbing distribution.

The greater amount of loss of information with increased perturbation (as illustrated by Figure 6) comes at the advantage of losing less privacy. In Figure 7, we have illustrated the decrease in privacy loss with increasing standard deviation in perturbing distributions. It is clear that the Figures 6 and 7 can be combined in order to eliminate the standard deviation dimension. This results in a trade-off curve between privacy and information loss. Such curves are shown in Figure 8. One advantage of such curves is that we can use them to compare the relative effectiveness of different combinations of original distributions and perturbations in a uniform way, irrespective of the shape and size of the distributions. One of the interesting observations from the result of Figure 8 is that the best perturbing dis-

¹It turns out that the AS algorithm is actually an approximation to the EM algorithm.

tribution is sensitive to the original distribution. For example, when the original distribution is gaussian, the gaussian and uniform perturbations are almost equally effective. On the other hand, for uniformly distributed data, the uniform perturbation is significantly more effective than the gaussian perturbation. These results qualify the heuristic arguments of [1] which advocate that the gaussian perturbations are superior for providing the best privacy preservation, since the exact behavior of the perturbation turns out to be dependent on the original distribution. Our results differ from [1] because of a more careful quantification of privacy in our paper.

In Figure 9, we have shown the behavior of the EM reconstruction algorithm with increasing number of points in the data. This curve corresponds to the case when the original and perturbing distribution was gaussian. In this case, the variance of the original and perturbing distribution were $2/\pi e$ and 0.8 respectively. We found that when there are a large number of data points, the amount of information loss was negligible. This trend was consistent for all possible combinations of original and perturbing distributions. Recall that we demonstrated earlier that the EM Algorithm produces zero information loss in the asymptotic case when there are a large number of data points. The graph in Figure 9 is consistent with that result. Given the fact that the privacy curves are independent of the number of data points, it follows that for very large data sets, the EM reconstruction algorithm can provide very high privacy guarantees for almost no information loss.

6. CONCLUSIONS AND SUMMARY

In this paper, we discussed the design and quantification of privacy-preserving data mining algorithms. We proposed an expectation-maximization algorithm which provably converges to the maximum-likelihood estimate of the original distribution. Thus, the algorithm provides a robust estimate of the original distribution. We laid the foundations for quantification of privacy gain and information-loss in a theoretically accurate and method independent way. We qualified the relative effectiveness of different perturbing distributions using these metrics. Our tests also demonstrate that when the data is large then the expectation maximization algorithm can reconstruct the data distribution with almost zero information loss.

7. ACKNOWLEDGEMENTS

We would like to thank Arvind Krishna for suggesting the problem.

8. REFERENCES

- [1] R. Agrawal and R. Srikant. Privacy preserving data mining. In *Proceedings of the ACM SIGMOD*, pages 439–450, 2000.
- [2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, Massachusetts, 1995.
- [3] C. Clifton and D. Marks. Security and privacy implications of data mining. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 15–19, May 1996.

- [4] H. Cramer. *Mathematical Models of Statistics*. Princeton University press, 1946.
- [5] L. F. Cranor. Special issue on internet privacy. *Communications of the ACM*, 42(2), 1999.
- [6] V. Estivill-Castro and L. Brankovic. Data swapping: Balancing privacy against precision in mining for logic rule. In *DaWak99*, pages 389–398, 1999.
- [7] T. Lau, O. Etzioni, and D. S. Weld. Privacy interfaces for information management. *CA CM*, 42(10):89–94, 1999.
- [8] C. K. Liew, U. J. Choi, and C. J. Liew. A data distortion by probability distribution. *ACM TODS*, 10(3):395–411, 1985.
- [9] C. E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.
- [10] The Economist. The end of privacy, May 1999.
- [11] The World Wide Web Consortium. The platform for privacy preference (P3P). Available from <http://www.w3.org/P3P/P3FAQ.html>.
- [12] K. Thearling. Data mining and privacy: A conflict in making. *DS**, Mar. 1998.
- [13] Time. The death of privacy, Aug. 1997.
- [14] H. L. V. Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley & Sons, New York, 1968.
- [15] B. P. Truste. An online privacy seal program. *Communications of the ACM*, 42(2):56–59, 1999.
- [16] J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.

APPENDIX

A.

In this section, we will provide a detailed proof of Theorem 4.1.

THEOREM 4.1. *The value of $Q(\Theta, \hat{\Theta})$ during the E-step of the reconstruction algorithm is given by:*

$$Q(\Theta, \hat{\Theta}) = \sum_{i=1}^K \psi_i(\mathbf{z}; \hat{\Theta}) \ln \theta_i,$$

where

$$\psi_i(\mathbf{z}; \hat{\Theta}) = \hat{\theta}_i \sum_{j=1}^N \frac{\Pr(Y \in z_j - \Omega_i)}{f_{Z; \hat{\Theta}}(z_j)}.$$

PROOF. Since $\mathbf{X} = X_1, X_2, \dots, X_N$ are independent, it follows that:

$$f_{\mathbf{X}; \Theta}(\nu) = \prod_{i=1}^N f_{X_i; \Theta}(\nu_i). \quad (11)$$

Using (11), we have:

$$Q(\Theta, \hat{\Theta}) = E \left[\ln f_{\mathbf{X}; \Theta}(\mathbf{X}) | \mathbf{Z} = \mathbf{z}; \hat{\Theta} \right] \quad (12)$$

$$= \sum_{j=1}^N E \left[\ln f_{X_j; \Theta}(X_j) | \mathbf{Z} = \mathbf{z}; \hat{\Theta} \right] \quad (13)$$

$$= \sum_{j=1}^N E \left[\ln f_{X_j; \Theta}(X_j) | Z_j = z_j; \hat{\Theta} \right]. \quad (14)$$

In (14), the expected value is computed only with respect to $f_{X_j | Z_j = z_j; \hat{\Theta}}$. This simplification results by the independence of X_1, \dots, X_N and Y_1, \dots, Y_N .

Furthermore, since

$$f_{X_j | Z_j = z_j; \hat{\Theta}}(\nu) = \frac{f_{X, Z; \hat{\Theta}}(\nu, z_j)}{f_{Z; \hat{\Theta}}(z_j)}, \quad (15)$$

it follows that,

$$\begin{aligned} & E \left[\ln f_{X_j; \Theta}(X_j) | Z_j = z_j; \hat{\Theta} \right] \\ &= \int_{\Omega} \ln f_{X_i; \Theta}(\nu) \frac{f_{X, Z; \hat{\Theta}}(\nu, z_j)}{f_{Z; \hat{\Theta}}(z_j)} d\nu \\ &= \frac{\int_{\Omega} \ln [f_{X_i; \Theta}(\nu)] f_{X; \hat{\Theta}}(\nu) f_Y(z_j - \nu) d\nu}{f_{Z; \hat{\Theta}}(z_j)} \\ &= \frac{\sum_{i=1}^K \int_{\Omega} \ln [f_{X_i; \Theta}(\nu)] \hat{\theta}_i I_{\Omega_i}(\nu) f_Y(z_j - \nu) d\nu}{f_{Z; \hat{\Theta}}(z_j)} \end{aligned} \quad (16)$$

Here the last equation follows from the assumption that f_X is piecewise constant. Since I_{Ω_i} is 1.0 over Ω_i and zero elsewhere,

$$\begin{aligned} & E \left[\ln f_{X_i; \Theta}(X_i) | Z_i; \hat{\Theta} \right] \\ &= \frac{\sum_{i=1}^K \hat{\theta}_i \int_{\Omega_i} \ln [f_{X_i; \Theta}(\nu)] f_Y(z_j - \nu) d\nu}{f_{Z; \hat{\Theta}}(z_j)} \\ &= \frac{\sum_{i=1}^K \hat{\theta}_i \int_{\Omega_i} \ln [\theta_i] f_Y(z_j - \nu) d\nu}{f_{Z; \hat{\Theta}}(z_j)} \\ &= \frac{\sum_{i=1}^K \hat{\theta}_i \ln \theta_i \int_{\Omega_i} f_Y(z_j - \nu) d\nu}{f_{Z; \hat{\Theta}}(z_j)} \\ &= \frac{\sum_{i=1}^K \hat{\theta}_i \ln \theta_i \Pr(Y \in z_j - \Omega_i)}{f_{Z; \hat{\Theta}}(z_j)} \end{aligned} \quad (17)$$

Combining (14) and (17), we get the following expression for $Q(\Theta, \hat{\Theta})$:

$$Q(\Theta, \hat{\Theta}) = \sum_{j=1}^N \frac{\sum_{i=1}^K \hat{\theta}_i \ln \theta_i \Pr(Y \in z_j - \Omega_i)}{f_{Z; \hat{\Theta}}(z_j)} \quad (18)$$

$$= \sum_{i=1}^K \hat{\theta}_i \ln \theta_i \sum_{j=1}^N \frac{\Pr(Y \in z_j - \Omega_i)}{f_{Z; \hat{\Theta}}(z_j)} \quad (19)$$

$$= \sum_{i=1}^K \psi_i(\mathbf{z}; \hat{\Theta}) \ln \theta_i \quad (20)$$

where $\psi_i(\mathbf{z}; \hat{\Theta}) = \hat{\theta}_i \sum_{j=1}^N \frac{\Pr(Y \in z_j - \Omega_i)}{f_{Z; \hat{\Theta}}(z_j)}$. ■