Generic Probability Density Function Reconstruction for Randomization in Privacy-Preserving Data Mining

Vincent Yan Fu Tan[1,⋆] and See-Kiong Ng[2]

[1] Massachusetts Institute of Technology (MIT), Cambridge, MA 02139
vtan@mit.edu
[2] Institute for Infocomm Research (I^{2}R), Singapore 119613
sknq@i2r.a-star.edu.sg

Abstract. Data perturbation with random noise signals has been shown to be useful for data hiding in privacy-preserving data mining. Perturbation methods based on additive randomization allows accurate estimation of the Probability Density Function (PDF) via the Expectation-Maximization (EM) algorithm but it has been shown that noise-filtering techniques can be used to reconstruct the original data in many cases, leading to security breaches. In this paper, we propose a *generic* PDF reconstruction algorithm that can be used on non-additive (and additive) randomization techiques for the purpose of privacy-preserving data mining. This two-step reconstruction algorithm is based on Parzen-Window reconstruction and Quadratic Programming over a convex set – the probability simplex. Our algorithm eliminates the usual need for the iterative EM algorithm and it is generic for most randomization models. The simplicity of our two-step reconstruction algorithm, without iteration, also makes it attractive for use when dealing with large datasets.

Keywords: Randomization, Privacy-preserving data mining, Parzen-Windows, Quadratic Programming, Convex Set.

1 Introduction

Consider the following scenario: There are two hospitals which seek to predict new patients' susceptibility to illnesses based on existing data. It would be useful for the hospitals to pool their data, since data mining tasks can often benefit from a large training dataset. However, by law, the hospitals cannot release private patient data. Instead, some form of sanitized data has to be provided to a centralized server for further analysis. It is thus imperative to discover means to protect private information and be able to perform data mining tasks with a masked version of the raw data. Can privacy and accuracy co-exist? This is the fundamental question in privacy-preserving data mining [2,3].

⋆ Vincent Tan is supported by the Agency for Science, Technology and Research (A*STAR), Singapore. He performed this work at I^{2}R, A*STAR.

P. Perner (Ed.): MLDM 2007, LNAI 4571, pp. 76–90, 2007.

Randomization has been shown to be a useful technique for hiding data in privacy-preserving data mining. The basic concept is to sufficiently mask the actual values of the data by perturbing them with an appropriate level of noise that can still allow the underlying Probability Density Function (PDF) of the original dataset to be adequately estimated from the randomized data. A balance has to be achieved between two conflicting concerns in such approaches. On one hand, the *confidentiality* of the precise information has to be protected *i.e.* to minimize *privacy loss*. On the other hand, the *utility* of the aggregate data statistics has to be maintained *i.e.* to minimize *information loss*.

The use of randomization for preserving privacy was studied extensively in the framework of statistical databases [1]. It typically involves a trusted centralized database in which the data are already fully known before they are randomized and released for publication (e.g. census data). As such, privacy-preserving transformations such as sampling [24] and swapping [24] are more suitable for perturbing the data as they can incorporate knowledge about the aggregate characteristics of the dataset. In privacy preserving data mining (PPDM), we consider both (trusted) centralized database scenarios as well as distributed scenarios in which there is one (untrusted) central server that needs pieces of private information from multiple clients to build a aggregate model for the data, and the clients would each perturb the information before releasing them to the server to preserve privacy.

The early attempts by the pioneering authors in PPDM [2] applied additive white noise (e_i), generated from a pre-determined distribution, to the true data (x_i) and then transmitting the sum ($z_i = x_i + e_i$) instead of the raw data. As it was shown that the distribution of the original data $f_X(x)$ can be reconstructed to a high accuracy *i.e.* low information loss, data mining can then be done satisfactorily using the sum (z_i) instead of the original data values (x_i). The reconstruction process hinges on the use of the *iterative* Expectation-Maximization (EM) algorithm taking the original values x_i as the latent variables.

However, Kargupta *et al.* [15] showed that such methods risk privacy breaches as the additive noise can be filtered off leaving a reasonably good estimation of the original data in many cases. Thus, other randomization models, such as using multiplicative noise, have been suggested [15].

Motivated by this, we develop a novel *non-iterative* PDF reconstruction scheme based on Parzen-Window reconstruction and Quadratic Programming (QP) optimization with only one equality constraint and one inequality constraint. These constraints define the probability simplex, which is a convex set. Convex programming/optimization [5,7] has been widely studied and efficient methods can be employed to estimate the PDF. As far as we know, currently only the mean and the variance in a multiplicative model can be estimated accurately [16]. To the best of our knowledge, for the first time, our method can allow the underlying PDF of the original dataset to be accurately reconstructed from randomized data set perturbed with multiplicative noise, additive noise or other noise models. From the estimated PDF, *all* the statistics can be inferred

for distribution-based data mining purposes. Our approach therefore provides a complete description of the original data without compromising on the privacy.

Other randomization/reconstruction methods based on multiplicative noise have been proposed [22] but the implementation of the reconstruction method was very computationally demanding. As such, we have also made sure that our reconstruction method can be efficiently implemented, avoiding using the iterative Expectation-Maximization algorithm employed in many reconstruction approaches for perturbation-based privacy-preserving data mining [2]. This makes our method attractive for use with the increasingly large datasets that have become commonplace in recent years.

More recently, [18] proposed a data perturbation approach in which the data is multiplied by a randomly generated matrix, hence preserving privacy by effectively projecting the data into a lower dimension subspace. As the transformation is distance-preserving, the authors showed that it is possible to estimate from the perturbed data various distance-related statistical properties of the original data. We consider non-distance-preserving randomization models in this paper because the distance-preserving nature of the randomization scheme in [18] may result in security breaches if some private data is also revealed.

In short, our reconstruction algorithm has two main advantages:

1. Unlike EM, it is non-iterative and can handle large datasets.
2. More importantly, it can be applied to generic (non-additive) randomization models, including multiplicative noise models.

The rest of this paper is organized as follows: We define the generic perturbation model and state some assumptions in Section 2. We describe the Parzen-Window and Quadratic Programming reconstruction algorithm in Section 3. In Section 4 we describe the evaluation metrics. We then present extensive evaluation results on both simulated and real data sets to validate our technique in Section 5. Finally, we conclude in Section 6 and provide some discussions on future work.

2 Problem Definition

The current PPDM framework consists of two processes: a randomization process, followed by a reconstruction process. First, the source data is randomized at possibly multiple client sites. The randomized data are then transmitted to a centralized server which attempts to recover the PDF of the original data for aggregate analyses. In the next two sections, we will first formally define the randomization model for privacy-preserving preservation, followed by the basic assumptions that are necessary for the subsequent reconstruction process.

2.1 Randomization Model

The generic randomization problem can be stated, succintly and generally, using the following mathematical model. Consider a set of N original scalars representing a particular private attribute (e.g. income) $x_1, \ldots, x_N$, which are drawn from

independent and identically distributed (IID) random variables $X_1, \ldots, X_N$. These random variables X_i follow a common PDF $f_X(x)$. To create the perturbation, we consider the generic two-variable randomization model:

$$z_i = \mathcal{Z}(e_i, x_i), \qquad \forall\, i \in \{1, \ldots, N\} \tag{1}$$

where the $e_1, \ldots, e_N$ are realizations of known IID random variables $E_1, \ldots, E_N$. $\mathcal{Z}(\cdot, \cdot)$ is a deterministic, possibly nonlinear, randomization operator. The e_i's are sampled from a specified uniform distribution. Therefore $E_i \sim \mathcal{U}(e; a_E, b_E)$, where $f_E(e) = \mathcal{U}(e; a_E, b_E)$ is the uniform distribution parameterized by lower and upper limits a_E and b_E respectively.

2.2 Reconstruction of PDF and Assumptions

Given the perturbed values $z_1, \ldots, z_N$ and the noise distribution, the reconstruction task is to obtain an estimate for the original PDF, which we denote $\hat{f}_X(x)$[1]. We make the following simple assumptions for recovering the PDF of X, $f_X(x)$:

$\mathcal{A}1$. The random variables X and E are statistically independent (SI) *i.e.* the joint distribution $f_{X,E}(x, e) = f_X(x) f_E(e)$ is equal to the product of the marginals.

$\mathcal{A}2$. The PDFs of X and E are finitely supported by $\mathcal{D}_X$ and $\mathcal{D}_E$ respectively. Outside these domains, $f_X(x) = f_E(e) = 0$.

Assumption $\mathcal{A}1$ is a common assumption in privacy-preserving data mining using randomization. It basically implies that the perturbing and original distributions are SI, which is a reasonable assumption. Assumption $\mathcal{A}2$ simplifies the computation for the reconstruction of the original PDF $\hat{f}_X(x)$ without loss of generality. This will be evident in Section 3, where the reconstruction algorithm is presented.

3 Randomization and Reconstruction Algorithms

Given the original data x_i, we will generate random numbers from a known uniform distribution to obtain the randomized data values z_i (c.f. Section 2.1). Because we are applying the noise e_i element-wise (as in Eq (1)), our randomization and reconstruction algorithm can be applied to both the centralized the distributed scenarios. It has been suggested [15] that the use of multiplicative noise is better than the additive model for minimizing risk of security breaches. In fact, our model goes beyond multiplicative noise. Any noise model of the form $z_i = \mathcal{Z}(e_i, x_i)$ can be used.

The key here, is whether we can effectively reconstruct the PDF of original data from the perturbed data. In this section, we will show how this can be done effectively and efficiently, without the need of the commonly-used iterative EM

[1] In this paper, estimates of functions, vectors and other variables are denoted with a overhead hat. For example, $\hat{a}$ is the estimate for a.

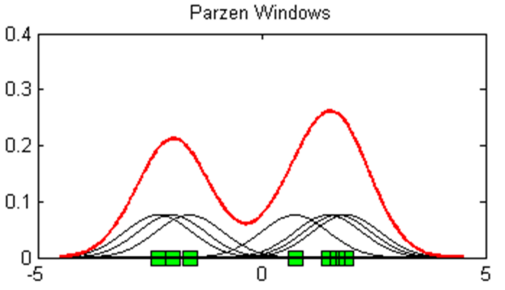

Fig. 1. Illustration of Parzen-Windows for estimation of the multimodal PDF. The *boxes* are the $N = 7$ independent realizations of the multimodal random variable. The individual Gaussian kernels are centered at the realizations. Their *sum*, as detailed in Eq (2) and indicated by the bold line, is the Parzen-Window approximation [20].

reconstruction algorithm. The general idea is as follows. Given the perturbed values z_i, we will first obtain an estimate of $f_Z(z)$ via Parzen-Windows [20]. Following the estimation of $f_Z(z)$, we will use Quadratic Programming (QP) to obtain an estimate of $f_X(x)$.

3.1 Estimate PDF of Perturbed Samples $f_Z(z)$ Via Parzen-Windows

The first step of the reconstruction algorithm is to estimate $\hat{f}_Z(z)$ using Parzen density estimation [20]. In this step, we are given N perturbed samples $z_1,\ldots,z_N$. They follow a common random variable Z, with true PDF $f_Z(z)$.

Parzen-Windows. The Parzen-Window approximation of the PDF of the perturbed samples is

$$\hat{f}_Z(z) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sigma_p\sqrt{2\pi}} \exp\left[-\frac{(z-z_i)^2}{2\sigma_p^2}\right], \tag{2}$$

where σ_p is the standard deviation or 'width' of the kernel. This estimator uses the Gaussian kernel function to smooth the raw sample set, placing more probability mass in regions with many samples, which is intuitively evident.

Example 1. An illustration of how the Parzen-Window method works for $N = 7$ is shown in Fig 1. We show the samples drawn from an arbitrary distribution. The Parzen approximation is the sum of the individual Gaussian kernels of equal standard deviations σ_p.

Remark 1. For Parzen-Window estimation, the quality of the estimate depends on the number of samples N as well as the standard deviation (SD) σ_p. If σ_p is too small, the Parzen approximation suffers from too much statistical variability and if σ_p is too large, the Parzen approximation is over-smoothed. Hence, we will now turn our attention to the selection of the optimal value of σ_p.

Cross-validation scheme for σ_p. In our experiments, we will use a cross-validation scheme that guarantees an optimal value of σ_p [4,21] in the l_2 sense. In this univariate optimization procedure, we seek to minimize the Integrated Squared Error (ISE) between the estimated PDF $\hat{f}_Z(z)$ and the actual PDF $f_Z(z)$:

$$\text{ISE} \triangleq \int_{\mathcal{D}_Z} \left(\hat{f}_Z(z) - f_Z(z) \right)^2 dz. \tag{3}$$

The ISE can be simplified to given the 'leave-one-out' (LOO) cross-validation criterion

$$\sigma_p^* = \underset{\sigma_p}{\text{argmin}}\ E_{LOO}(\sigma_p), \tag{4}$$

with $E_{LOO}(\sigma_p)$ defined as

$$E_{LOO}(\sigma_p) \triangleq \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{N}(z_i; z_j, \sqrt{2}\sigma_p) - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathcal{N}(z_i; z_j, \sigma_p), \tag{5}$$

and $\mathcal{N}(x; \mu, c) = (c\sqrt{2\pi})^{-1} \exp\left[-(x-\mu)^2/2c^2\right]$ is the Gaussian kernel with mean μ and variance c^2. The optimization problem in Eq (4) is one-dimensional and efficient line search methods [17] will yield sufficiently accurate solutions.

3.2 Estimate Original PDF $f_X(x)$ Via Quadratic Programming (QP)

Equipped with an estimate of the perturbed PDF $\hat{f}_Z(z)$, we are ready to estimate the original PDF $f_X(x)$.

Theorem 1. *Let $Z = \mathcal{Z}(X, E)$ be the result of a function of two random variables that can also be expressed as $E = \mathcal{E}(X, Z)$ i.e. given $X = x$, the transformation is one-to-one. Then, if assumptions $\mathcal{A}1$ and $\mathcal{A}2$ (c.f. Section 2.2) are satisfied, the Probability Density Function (PDF) of Z, $\hat{f}_Z(z)$ can be written as*

$$\hat{f}_Z(z) = \int_{\mathcal{D}_X} f_X(x) f_E[\mathcal{E}(x, z)] \left| \frac{\partial \mathcal{E}(x, z)}{\partial x} \right| dx. \tag{6}$$

Proof. See Appendix A. □

The assumption that the transformation from Z to E given $X = x$ is one-to-one is made *without any loss of generality*. This is because we can represent the set $\mathcal{A} = \{(x, e)\}$ of input variables as the union of a finite number, say K, of mutually disjoint subsets $\{\mathcal{A}_k\}_{k=1}^K$ such that the transformation is one-to-one in each of $\mathcal{A}_k$ onto $\mathcal{B} = \{(v, z)\}$. We focus on the one-to-one case for notational simplicity but note that it is straightforward to extend the argument to the case where the transformation is not one-to-one. For example, the randomization model $z_i = \mathcal{Z}(e_i, x_i) = x_i e_i + x_i^2 e_i^4$ is not one-to-one. Nonetheless, it is still possible to apply our reconstruction algorithm, with appropriate modifications to Eq (6). We refer the reader to the excellent treatment of functions of random variables by Hogg and Craig [14, Chapter 4].

QP Formulation. Using Theorem 1, we can formulate the Quadratic Program to estimate the optimal $f_X(x)$. Discretizing[2] the integral in Eq. (6) yields

$$\hat{f}_Z(z) \approx \sum_{n\Delta x \in \mathcal{D}_X} f_X(n\Delta x)\, f_E\left[\mathcal{E}(n\Delta x, z)\right] \left|\frac{\partial \mathcal{E}(x,z)}{\partial x}\right|_{x=n\Delta x} \Delta x, \tag{7}$$

where $\Delta x > 0$ is the step size of $f_X(x)$ and $n\Delta x \in \mathcal{D}_X$ and $\mathcal{D}_X \triangleq \{(n_0+1)\Delta x, \ldots, (n_0+N_x)\Delta x\}$ is the set of discretized points contained in the finitely-supported[3] domain of X. Then for $z \in \{z_1, \ldots, z_{N_z}\}$, Eq. (7) can be written as

$$\hat{\mathbf{f}}_Z = \mathbf{G}_E\, \mathbf{f}_X, \tag{8}$$

where the length N_z vector $\hat{\mathbf{f}}_Z$, length N_x vector $\mathbf{f}_X$ and the N_z by N_x matrix $\mathbf{G}_E$ are defined as

$$[\hat{\mathbf{f}}_Z]_j \triangleq \hat{f}_Z(z_j), \tag{9}$$

$$[\mathbf{f}_X]_i \triangleq f_X((n_0+i)\Delta x), \tag{10}$$

$$[\mathbf{G}_E]_{ij} \triangleq f_E\left[\mathcal{E}((n_0+i)\Delta x, z_j)\right] \left|\frac{\partial \mathcal{E}(x,z_j)}{\partial x}\right|_{x=(n_0+i)\Delta x} \Delta x, \tag{11}$$

and $[\mathbf{v}]_k$ is the k^{th} element of the vector $\mathbf{v}$ and $[\mathbf{M}]_{ij}$ is the element in the i^{th} row and j^{th} column of the matrix $\mathbf{M}$ and $i \in \{1, \ldots N_z\}$ and $j \in \{1, \ldots N_x\}$. Eq. (8) can be converted into the canonical cost function in a Quadratic Program as shown in Appendix B.

Example 2. If as in [3], we use an additive scheme *i.e.* $z_i = \mathcal{Z}_{add}(e_i, x_i) = x_i + e_i$, then Eq (11), together with the convolution formula [19], simplifies to give

$$[\mathbf{G}_E]_{ij} \triangleq f_E\left[z_j - (n_0+i)\Delta x\right]. \tag{12}$$

Example 3. If instead we use a multiplicative scheme [16] *i.e.* $z_i = \mathcal{Z}_{mul}(e_i, x_i) = e_i \times x_i$, then Eq (11) together with the result in [13] simplifies to give

$$[\mathbf{G}_E]_{ij} \triangleq \left|\frac{1}{n_0+i}\right| f_E\left[\frac{z_j}{(n_0+i)\Delta x}\right], \qquad n_0 \neq -i. \tag{13}$$

Constraints. As $f_X(x)$ is a PDF, it has to satisfy the stochastic constraints $f_X(x) \geq 0\,, \forall\, x \in \mathcal{D}_X$ and $\int_{\mathcal{D}_X} f_X(x)\, dx = 1$. This places an inequality and an equality constraint on the vector $\mathbf{f}_X$, which can be easily incorporated into the QP as:

$$\mathbf{f}_X \geq \mathbf{0}_{N_X \times 1}, \qquad \sum_{n\Delta x \in \mathcal{D}_X} f_X(n\Delta x) = \frac{1}{\Delta x}. \tag{14}$$

[2] This is done using the Rectangular rule. We can alternatively use the Trapezoidal, Simpson or Quadrature rules [8] to discretize the integral. Our experimental results, however, show that the performances of these discretization rules are very similar and hence for simplicity, we shall only present Rectangular rule.

[3] By assumption $\mathcal{A}2$.

Sufficient conditions for QP. We now derive sufficient conditions for the optimization problem. Because the constraint set is particularly simple, we can obtain the optimal solution without the use of iterative methods (such as gradient projection or modern interior points methods). Consider our quadratic program:

$$\min_{\mathbf{f}_X} \quad J(\mathbf{f}_X) = \frac{1}{2}\mathbf{f}_X^{\mathrm{T}}\mathbf{H}\mathbf{f}_X + \mathbf{h}^{\mathrm{T}}\mathbf{f}_X, \tag{15}$$

subject to

$$\mathbf{f}_X \in \mathcal{C}, \quad \text{with} \quad \mathcal{C} = \left\{ \mathbf{f}_X \,\middle|\, \mathbf{f}_X \geq \mathbf{0}, \sum_{i=1}^{N_x} [\mathbf{f}_X]_i = 1 \right\}, \tag{16}$$

for appropriately chosen $\mathbf{H}$ and $\mathbf{h}$ (as shown in Appendix B). Then, the necessary condition for $\mathbf{f}_X^*$ to be a local minimum over a convex set [6, Section 2.1] is

$$\sum_{i=1}^{N_x} \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_i} ([\mathbf{f}_X]_i - [\mathbf{f}_X^*]_i) \geq 0, \quad \forall\, \mathbf{f}_X \in \mathcal{C}. \tag{17}$$

Subsequent simplification yields the condition

$$[\mathbf{f}_X^*]_i > 0 \Rightarrow \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_i} < \frac{\partial J(\mathbf{f}_X^*)}{\partial [\mathbf{f}_X]_j} \Leftrightarrow \sum_{k=1}^{N_x} [\mathbf{H}]_{ik}[\mathbf{f}_X]_k + [\mathbf{h}]_i < \sum_{k=1}^{N_x} [\mathbf{H}]_{jk}[\mathbf{f}_X]_k + [\mathbf{h}]_j, \, \forall\, j. \tag{18}$$

Thus, all coordinates which are (strictly) positive at the optimum must have minimal (and equal) partial cost derivates [6]. Since $\mathbf{G}_E$ only contains real entries, the Hessian matrix $\mathbf{H} = \mathbf{G}_E^{\mathrm{T}}\mathbf{G}_E$ of the QP is positive semidefinite. Consequently, the cost function $J(\cdot)$ is convex [7] and any local optimum of Eq (15) is also a global optimum, which implies that the cost value is equal for all local optima. Moreover, the set of local optima is always convex.

We exploit the convexity of the cost function to conclude that Eq (18) is also a sufficient condition for global optimality of $\mathbf{f}_X^* = \hat{\mathbf{f}}_X$.

3.3 Discussion

We have completed the discussion of our non-iterative PDF reconstruction for generic randomization schemes for privacy-preserving data mining. There are two steps: Firstly, we build the Parzen-Window of the perturbed samples $\hat{f}_Z(z)$. Secondly, we perform a QP over the probability simplex to reconstruct an estimate of the original PDF $\hat{f}_X(x)$. Our algorithm is summarised in Fig 2. We conclude this section with two comments on our algorithm.

1. Discretizing the integral in Eq (6) is, in general, intractable if we are reconstructing PDFs of high dimensions as the problem suffers from the 'curse of dimensionality'. We can mitigate the effects of the curse by assuming the dimensions are independent, if possible. Using this naïve approach, we estimate the PDF in each dimension before taking their product to form the joint density. Alternatively, we can project the data onto a lower dimensional subspace and perform the same analysis in that subspace.

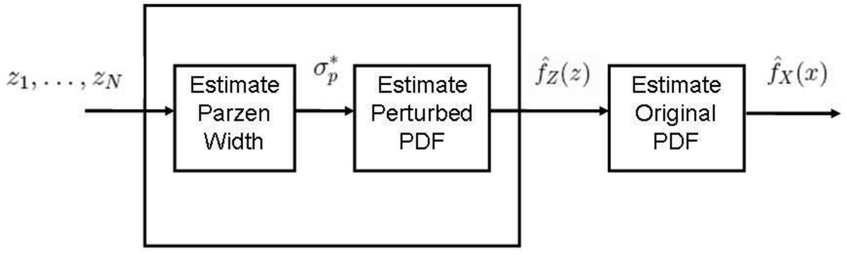

Fig. 2. The PDF reconstruction algorithm. There are two main steps. We reconstruct $\hat{f}_Z$ via Parzen-Windows. Then we estimate of $\hat{f}_X(x)$ using the QP.

2. The reconstruction algorithm can handle large datasets. One approach is to find a *random subset* of the samples from the dataset z_i to build the Parzen-Window and to perform the QP. This is known as the reduced Parzen-Window and is discussed in more detail in [12].

4 Performance Metrics

As mentioned earlier, there are two competing issues. Firstly, we hope to minimize the *privacy loss* so that individual information is not revealed. At the same time, we want to preserve the structure and the aggregate statistics of the underlying data. In other words, we also hope to minimize the *information loss*.

4.1 Privacy Loss

In this section, we will quantify privacy loss using mutual information. It was argued in [2] that the mutual information between two random variables X and Z measures the degree of independence between the random variables and hence, the privacy loss for X when Z is revealed.

The mutual information $I(X;Z)$ tells us how much information one random variable tells about another one. In other words, $I(X;Z)$ is the amount of uncertainty in X, which is removed by knowing Z. When X and Z are independent, $I(X;Z) = 0$. The lower the value of $I(X;Z)$, the better the privacy gain via the given perturbation scheme, the more the privacy is preserved. This leads us to the notion of the privacy loss $\mathcal{P}(X|Z)$ of X when Z is known. It is defined as:

$$\mathcal{P}(X|Z) \stackrel{\triangle}{=} 1 - 2^{-I(X;Z)}. \tag{19}$$

By definition, $0 \leq \mathcal{P}(X|Z) \leq 1$. $\mathcal{P}(X|Z) = 0$ if and only if X and Z are SI.

Remark 2. Privacy breach [9], based on worst-case information loss, has also been suggested as an alternative privacy measure. However, in our work, we consider an *average* disclosure measure – mutual information. Also, the privacy breach [10] measure is typically used in the context of association-rule mining, which is not applicable in our context.

4.2 Information Loss

In this section, we will define information loss, which is a measure of the effectiveness and accuracy of the reconstruction algorithm. It is clear that given the perturbed values $z_1, \ldots, z_N$, it is, in general, not possible to reconstruct the original density $f_X(x)$ with arbitrary precision. The lack of precision in estimating $f_X(x)$ from the perturbed values is referred to as information loss. The closer our estimate $\hat{f}_X(x)$ is to the actual PDF $f_X(x)$, the lower the information loss. We use the following *universal* metric suggested in [2] to quantify the information loss in the reconstruction of $f_X(x)$.

$$\mathcal{I}(f_X, \hat{f}_X) \stackrel{\triangle}{=} \frac{1}{2}\mathbf{E}\left[\int_{\mathcal{D}_X} \left|f_X(x) - \hat{f}_X(x)\right| \, dx\right], \tag{20}$$

where $\hat{f}_X(x)$ is the estimate for the PDF of the random variable X. It is easy to see that $0 \leq \mathcal{I}(f_X, \hat{f}_X) \leq 1$. We will see that our algorithm produces an accurate original PDF that is amendable to various distribution-based data mining tasks.

5 Experiments

We conducted two main experiments to demonstrate the efficiency and accuracy of the PQP reconstruction algorithm.

1. Firstly, we examine the tradeoff between the privacy loss and information loss. In Section 5.1, we show empirically that our generic PDF reconstruction algorithm performs as well as the additive randomization-EM algorithm suggested in [2]. We emphasize that our PDF reconstruction algorithm is applicable to all randomization models that can be expressed in the form Eq (1).
2. Secondly, we applied our algorithm to a *real dataset* and demonstrate that privacy can be preserved and, at the same time, the aggregate statistics can be mined. The results are discussed in Section 5.2.

5.1 Privacy/Accuracy Tradeoff

As mentioned previously, data perturbation based approaches typically face a privacy/accuracy loss tradeoff. In this section, we shall examine this tradeoff and compare it to existing technologies. We used two different randomization models – multiplicative and additive and examine the efficacy of the PDF reconstruction algorithm ('PQP'). The results are summarized in Fig 3.

We observe that our reconstruction algorithm performs as well as EM with the added bonus that it is generic. It can be applied to multiplicative, additive and other randomization models. Besides, it is non-iterative.

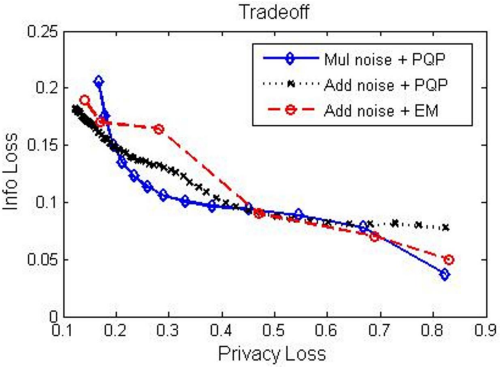

Fig. 3. Plot of the tradeoff between information loss $\mathcal{I}(f_X, \hat{f}_X)$ and privacy loss $\mathcal{P}(X|Z)$. Our PDF reconstruction algorithm ('PQP') performs just as well as EM but has the added bonus of being a generic reconstruction method.

Table 1. Information Losses resulting from the various perturbation/reconstruction methods. Privacy Loss is kept constant at $\mathcal{P}(X|Z) = 0.330$. We observe that the PQP reconstruction algorithm gives a superior (lower) information loss as compared to EM.

Method	Mul + PQP	Add + PQP	Add + EM [2]
$\mathcal{I}(f_X, \hat{f}_X)$	0.1174	**0.0957**	0.1208

5.2 Application to Real Data

We applied the Parzen-Window and QP reconstruction ('PQP') algorithm to real data obtained from The U.S. Department of Housing and Urban Development's (USDHUD's) Office of Policy Development and Research (PD&R) [23]. As with the previous experiment, we perturbed the data with multiplicative noise and additive noise. Other randomization techniques are also applicable.

The data in [23] provides us with the median income of all the counties in the 50 states in the U.S in 2005. The length of the dataset is $N = 3195$. This is plotted as a histogram with 75 bins in Figure 4(a). We multiplied each data value with samples drawn from a uniform distribution with domain $1 \leq e \leq 3$ giving a privacy loss value of $\mathcal{P}(X|Z) = 0.330$.

In addition to using the multiplicative randomization and PQP reconstruction algorithm, we also ran the PQP algorithm on the data corrupted by additive noise. The level of noise was adjusted such that the privacy loss is kept constant at $\mathcal{P}(X|Z) = 0.330$. Finally, we implemented the additive noise and EM reconstruction algorithm [2] on the data.

We averaged our results over 500 independent runs and the results are tabulated in Table 1. The results showed that our PDF reconstruction algorithm ('PQP') performed better than additive/EM [2] on the real data. The added

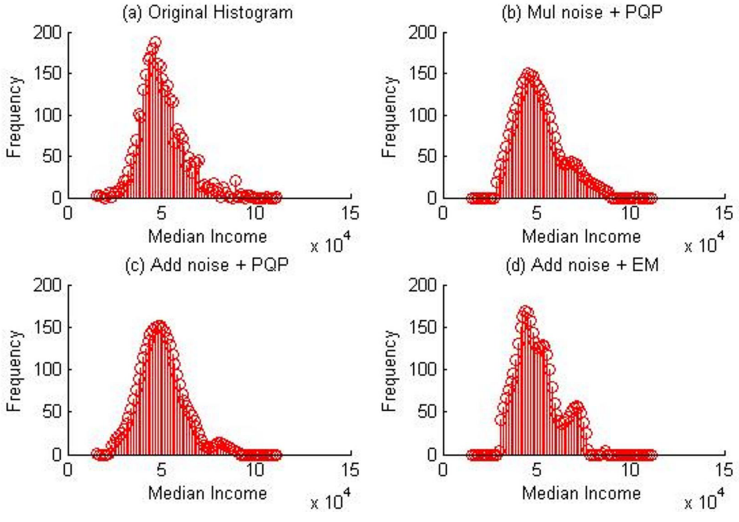

Fig. 4. (a) Original histogram of Median Income of Counties in the U.S. [23] (b) Reconstructed histogram after Multiplicative Randomization and our PDF reconstruction algorithm ('PQP'). (c) Reconstructed histogram after Additive Randomization and 'PQP'. (d) Reconstructed histogram after Additive Randomization and EM [2]. Note the accuracy of our PDF reconstruction algorithm.

advantage here is that our novel non-iterative PDF reconstruction algorithm can be applied to *all* randomization models of the form Eq (1).

6 Conclusions and Further Work

In this paper, we have devised a novel PDF reconstruction scheme for privacy-preserving data mining. This scheme is based on Parzen-Window reconstruction and Quadratic Programming (with a positive semidefinite Hessian) over a convex set. For the first time, the original PDF $f_X(x)$ can be approximated from the samples which have been perturbed by any type of noise (even multiplicative) that follows the generic randomization equation $z_i = \mathcal{Z}(e_i, x_i)$. We performed extensive numerical experiments demonstrating the efficacy of our algoritm. There are two distinct advantages over the existing PDF reconstruction algorithms which are based on the iterative EM algorithm.

1. Firstly, our proposed two-step reconstruction algorithm eliminated the common need for the *iterative* Expectation-Maximization (EM) algorithm. This is essential for problems which involve larger datasets, as it circumvents the need for iteration. It only involves two steps: Parzen-Window reconstruction and Quadratic Programming. The QP is particularly easy to solve because of the nature of the constraints – the (convex) probability simplex.

2. Secondly, our reconstruction method is also *generic*. Theorem 1 shows that the algorithm can be applied to many other randomization models as long as the perturbing random variable E and the underlying random variable X are SI, which is a common assumption for randomization methods in privacy-preserving data mining. We emphasize that although we examined the multiplicative and additive models only in Section 5, our reconstruction algorithm can be applied to all randomization models of the form Eq (1).

A natural extension to this work is to examine even more randomization models and reconstruction algorithms. For instance, we can parameterize Eq (1) as follows: $z_i = \mathcal{Z}(e_i(\psi), x_i; \psi)$ where ψ is an unknown but deterministic/non-random parameter. This adds an additional layer of privacy and the PDF can be estimated using a combination of our PQP reconstruction algorithm and maximum-likelihood methods. Finally, a question of paramount importance that researchers can try to decipher is: Does a fundamental relation between the privacy loss and information loss exist? We believe this needs to be answered precisely in order to unlock the promising future in privacy-preserving data mining.

Acknowledgments. The first author would like to thank Dr. Mafruzzaman Ashrafi of the Institute for Infocomm Research (I^2R) for his helpful comments. An anonymous reviewer made several valuable comments.

References

1. Adam, N.R., Worthmann, J.C.: Security-control methods for statistical databases: A comparative study. ACM Comput. Surv. 21, 515–556 (1989)
2. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy preserving data mining algorithm. In: Symposium on Principles of Database Systems, pp. 247–255 (2001)
3. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proc. of the ACM SIGMOD Conference on Management of Data, pp. 439–450. ACM Press, New York (2000)
4. Assenza, A., Archambeau, C., Valle, M., Verleysen, M.: Assessment of probability density estimation methods: Parzen-Window and Finite Gaussian Mixture ISCAS. In: IEEE International Symposium on Circuits and Systems, Kos (Greece) (2006), pp. 3245–3248 (2006)
5. Bertsekas, D., Nedic, A., Ozdaglar, A.E.: Convex Analysis and Optimization Athena Scientific (2003)
6. Bertsekas, D.: Nonlinear Programming Athena Scientific (2004)
7. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
8. Davis, P.J., Rabinowitz, P.: Methods of Numerical Integration. Academic Press, San Diego (1984)
9. Evfimievski, A., Gehrke, J., Srikant, R.: Limiting Privacy Breaches in Privacy Preserving Data Minin. In: Proc. of ACM SIGMOD/PODS Conference, pp. 211–222 (2003)

10. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy Preserving Mining of Association Rule. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, Edmonton, Alberta, Canada, pp. 217–228 (2002)
11. Fessler, J.A.: On transformations of random vectors. In: Technical Report 314, Comm. and Sign. Proc. Lab. Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122 (1998)
12. Fukunaga, K.: Statistical Pattern Recognition, 2nd edn. California Academic Press, San Diego (1990)
13. Glen, A., Leemis, L., Drew, J.: Computing the distribution of the product of two continuous random variables. Computational statistics & data analysis 44, 451–464 (2004)
14. Hogg, R.V., Craig, A.T.: Introduction to Mathematical Statistics, 5th edn. Prentice-Hall, Englewood Cliffs, NJ (1995)
15. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation technique. In: Proceedings of the 3rd IEEE International Conference on Data Mining, Washington, DC, USA, pp. 99–106 (2003)
16. Kim, J.J., Winkler, W.E.: Multiplicative noise for masking continuous data Technical Report Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census, Washington, DC, USA (2003)
17. Luenberger, D.G.: Linear and Nonlinear Programming. Addison-Wesley, London (1984)
18. Liu, K., Kargupta, H., Ryan, J.: Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining IEEE Transactions on Knowledge and Data Engineering (TKDE), 18, pp. 92–106 (2006)
19. Oppenheim, A.V., Willsky, A.S.: Signals and Systems. Prentice-Hall, Englewood Cliffs (1996)
20. Parzen, E.: On the estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076 (1962)
21. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall, London (1986)
22. Makov, U.E., Trotini, M., Fienberg, S.E., Meyer, M.M.: Additive noise and multiplicative bias as disclosure limitation techniques for continuous microdata: A simulation study. Journal of Computational Methods in Science and Engineering 4, 5–16 (2004)
23. U.S. Department of Housing, Urban Developments (USDHUDs) Office of Policy Development and Research (PD&R). (2005), `http://www.huduser.org/datasets/il/IL_99_05_REV.xls`
24. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y.: State-of-the-art in Privacy Preserving Data Mining. ACM SIGMOD Record 3, 50–57 (2004)

Appendix

A Proof of Theorem 1

Proof. Our proof is adapted from [11] and [13]. Using the transformation technique [14], the transformation $V = X$ and $Z = \mathcal{Z}(X, E)$ constitutes a one-to-one mapping from from $\mathcal{A} = \{(x, e)\}$ to $\mathcal{B} = \{(v, z)\}$. Let u denote the transformation

and w the inverse transformation. The transformation and its inverse can be written as:

$$v = u_1(x, e) = x, \qquad z = u_2(x, e) = \mathcal{Z}(x, e), \tag{A.1}$$

$$x = w_1(v, z) = z, \qquad e = w_2(v, z) = \mathcal{E}(v, z). \tag{A.2}$$

Consequently, the Jacobian determinant can be expressed in the form

$$J = \left| \begin{matrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial v} \\ \frac{\partial e}{\partial z} & \frac{\partial e}{\partial v} \end{matrix} \right| = \left| \frac{\partial \mathcal{E}(v, z)}{\partial v} \right| = \left| \frac{\partial \mathcal{E}(x, z)}{\partial x} \right|. \tag{A.3}$$

The marginal density of Z, which can be obtained through Parzen reconstruction from the samples of z_i can be found by integrating the joint density of V and Z

$$\hat{f}_Z(z) = \int_{\mathcal{D}_V} f_{V,Z}(v, z)\, dv. \tag{A.4}$$

Application of the transformation from $\mathcal{B}$ to $\mathcal{A}$ yields

$$\hat{f}_Z(z) = \int_{\mathcal{D}_V} f_{X,E}(w_1(v, z), w_2(v, z)) \left| \frac{\partial \mathcal{E}(x, z)}{\partial x} \right| dv, \tag{A.5}$$

A further simplification and the use of the statistical independence of X and E (Assumption $\mathcal{A}1$) gives Eq. (6). □

B Detailed Formulation of the Quadratic Program

The canonical QP can be written as

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \boldsymbol{\theta}^{\mathrm{T}} \mathbf{H} \boldsymbol{\theta} + \mathbf{h}^{\mathrm{T}} \boldsymbol{\theta} \right\}, \tag{B.1}$$

subject to

$$\mathbf{A}\boldsymbol{\theta} \leq \mathbf{b}, \quad \mathbf{A_{eq}}\boldsymbol{\theta} = \mathbf{b_{eq}}, \tag{B.2}$$

where $\mathbf{H}$, $\mathbf{A}$ and $\mathbf{A_{eq}}$ are matrices and $\mathbf{h}$, $\mathbf{b}$, $\mathbf{b_{eq}}$ and $\boldsymbol{\theta}$ are vectors, all appropriately sized. To optimize for a solution to Eq (8), we can write it in terms of an cost function

$$J(\mathbf{f}_X) = \frac{1}{2} \left\| \hat{\mathbf{f}}_Z - \mathbf{G}_E \mathbf{f}_X \right\|_2^2, \tag{B.3}$$

where $\| \cdot \|_2$ is the l_2 norm. Eq (B.3) can be can be simplified to give

$$J(\mathbf{f}_X) = \frac{1}{2} \mathbf{f}_X^{\mathrm{T}} \mathbf{G}_E^{\mathrm{T}} \mathbf{G}_E \mathbf{f}_X - \hat{\mathbf{f}}_Z^{\mathrm{T}} \mathbf{G}_E \mathbf{f}_X + c, \tag{B.4}$$

where c is some constant independent of $\mathbf{f}_X$. Hence, by comparing Eq (B.1) and Eq (B.4), we observe that $\boldsymbol{\theta} = \mathbf{f}_X$ is the vector of control variables and

$$\mathbf{H} = \mathbf{G}_E^{\mathrm{T}} \mathbf{G}_E, \quad \mathbf{h} = -\mathbf{G}_E^{\mathrm{T}} \hat{\mathbf{f}}_Z, \tag{B.5}$$

are the matrix (Hessian) and vector that define the cost function. Also, comparing the constraints in Eq (14) to the constraints in the canonical QP, we obtain

$$\mathbf{A} = -\mathbf{I}_{N_X \times N_X}, \quad \mathbf{b} = \mathbf{0}_{N_X \times 1}, \quad \mathbf{A_{eq}} = (\Delta x)\mathbf{1}_{1 \times N_X}, \quad \mathbf{b_{eq}} = 1. \tag{B.6}$$