On Unifying Privacy and Uncertain Data Models

Charu C. Aggarwal

IBM T. J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532, USA
charu@us.ibm.com

Abstract— **The problem of privacy-preserving data mining has been studied extensively in recent years because of the increased amount of personal information which is available to corporations and individuals. Most privacy transformations use some form of data perturbation or representational ambiguity in order to reduce the risk of identification. The final results from privacy transformation methods often require the underlying applications to be modified in order to work with the new representation of the data. Since the end results of privacy-transformation methods have not been standardized, the required modifications may vary with the method used for the privacy transformation. In some cases, it can be an enormous effort to re-design applications to work with the anonymized data.**

While the results of privacy-transformation methods are a natural form of uncertain data, the two problems have generally been studied independently. In this paper, we make a first attempt to unify the two fields, and propose a privacy transformation for which existing uncertain data management tools can be directly used. This is a great advantage, since it means that the wide spectrum of research available for uncertain data management can also be used for privacy-preserving data mining. We propose an uncertain version of the k-anonymity model which is related to the well known deterministic model of k-anonymity. The uncertain version of the k-anonymity model has the additional feature of introducing greater uncertainty for the adversary over an equivalent deterministic model. As specific instantiations of this approach, we test the effectiveness of the privacy transformation on the problems of query estimation and classification, and show that the technique retains greater accuracy than other k-anonymity models.

I. Introduction

In recent years, an increasing amount of personal data is being stored by private corporations and individuals because of advances in storage and collection hardware technology. This has lead to increased privacy concerns about the security of the underlying data. The field of privacy-preserving data mining is designed to develop tools which which can mine such sensitive data without compromising its privacy. A number of models have been proposed [1], [2], [3], [4], [5], [6], each of which leads to a different representation of the underlying data.

A related field to privacy is that of uncertain data management. In uncertain data management, a variety of management and mining tools need to be developed [7], [8], [9], [10] on data which is specified only approximately. Typical models in uncertain data mining [7] assume that a probability distribution function of the data is known. This probability distribution is then used in order to construct and improve the effectiveness of data management models. Although the two fields of privacy modeling and uncertain data mining are closely related, the research in the two fields has been largely independent. While privacy transformations such as randomization [2], [3], [11] use probabilistic models, the end result cannot be effectively leveraged with uncertain data analysis techniques. This is because such methods typically do not calibrate the added noise so that individual records can be used effectively with data management applications. In such cases, data mining techniques such as classification need to be re-designed to work with aggregate distributions [2] rather than individual records. Furthermore, modern techniques for privacy such as k-anonymity use diverse and ad-hoc representations of the data as the end result of the privacy transformation process. Some examples of the different kinds of privacy transformations are as follows:

• Methods such as k-anonymity reduce the granularity of the data using techniques such as generalization and suppression [6]. The final representation of the data may be ad-hoc, and cannot be directly used for an arbitrary data management applications, without some changes to the application itself. For example, the process of generalization may result in partitioning the data into ranges, and the uncertainty information in each range, as well as the ordering among different ranges may be lost, unless an application is specifically designed to take this into account.
• Methods such as perturbation [2], [3] add noise to the data. While the final data is probabilistic in nature, the noise added is *independent* of the behavior of the underlying data. As a result, the noise added is so large, that the individual data records are no long useful for mining purposes. Only aggregate distributions can be used for mining the final data. This reduces the applicability of the technique for general multi-variate applications. Furthermore, since these methods are designed for randomization at data collection time, they do not provide any guarantees on privacy such as k-anonymity definitions.
• Some other techniques for privacy preservation [12] use aggregate noise-based techniques for masking micro-data files. As in the case of [2], the techniques cannot easily be used with uncertain data models, and also do not provide any guarantee on the level of anonymity.

We note that each of the above-mentioned methods have a very different representation of the underlying data after the transformation process. In many cases, existing data management tools cannot be used directly with the output of the privacy model. Therefore, new data mining methods such as those discussed in [2] need to be designed for the underlying data. This can be an enormous effort, and is further complicated by the wide variety of privacy tools available, each of which work with their own representation. Therefore, it is always useful to transform the data to standardized data models which are well studied from a data management and mining perspective.

Many uncertain data models assume that the probability density function of each of the underlying data records is known [7]. This representation seems to be directly related to privacy models. In this paper, we will examine whether a privacy-preserving transform can be constructed so that standardized definitions of uncertainty and privacy can be achieved simultaneously. In order to achieve this goal, we will provide a *probabilistic definition* of k-anonymity and analyze the approach using a couple of standardized distributions. As specific instantiations of this method, we will examine its effectiveness for the problem of query selectivity estimation and classification. We will experimentally demonstrate the advantage of using this approach over the method of condensation based anonymization [1].

This paper is organized as follows. In the next section, we will study the problem of probabilistic transformation, and its application to the problems of classification and query selectivity estimation. The experimental results are presented in section 3. Section 4 contains the conclusions and summary.

II. The Privacy-Preserving Uncertain Model

We assume without loss of generality that the data set $\mathcal{D} = \overline{X_1} \dots \overline{X_N}$ is normalized so that the variance along each dimension is one. Each record $\overline{X_i}$ contains d dimensions. The assumption of unit variance along each dimension can be made because a-priori and a-posteriori scaling can be used in order to achieve the same result for an arbitrary data set. We assume that the data set is transformed to a set of N perturbed data records $\overline{Z_1} \dots \overline{Z_N}$ along with probability density functions $f_1(\cdot) \dots f_N(\cdot)$. We assume that $\overline{Z_i}$ is the mean of each probability density function. In the uncertain data representation, we assume that that the true value $\overline{X_i}$ is probabilistically distributed around $\overline{Z_i}$ using the density function $f_i(\cdot)$. The other parameters of the probability density function will be defined by the statistical behavior of the underlying data points, and should be chosen in such a way that k-anonymity is assured either probabilistically or deterministically. A one-to-one correspondence exists between each data point $\overline{X_i}$ and the *uncertain data point* $(\overline{Z_i}, f_i(\cdot))$. In order to actually generate $\overline{Z_i}$ from $\overline{X_i}$, we use the density function $g_i(\cdot)$, which is the same as $f_i(\cdot)$ except that it is centered around $\overline{X_i}$. We assume that the distributions $f_i(\cdot)$ and $g_i(\cdot)$ are drawn from the family of distributions in which the mean is one of the parameters of the distribution. This makes it possible to construct $f_i(\cdot)$ from $g_i(\cdot)$ and vice-versa. We note that most of the important distributions such as the normal, uniform, and exponential distributions satisfy this property. We note that the above mentioned definition is far more general than the perturbation technique discussed in [2], in that the uncertain representation contains a point-specific density function. This is required in order to carefully calibrate k-anonymity in the uncertain representation. Such calibration is essential in keeping the uncertainty in the data low enough, so that individual records continue to be useful for mining purposes. First, we formally define the uncertain data representation which is the output of the privacy-preserving transformation:

Definition 2.1: Given a set of data records $\overline{X_1} \dots \overline{X_N}$, we would like to construct a new set of records $\overline{Z_1} \dots \overline{Z_N}$, along with density functions $f_1(\cdot) \dots f_N(\cdot)$, such that the pair $(\overline{Z_i}, f_i(\cdot))$ is the (privacy-preserving) uncertain representation of record $\overline{X_i}$. The distribution $f_i(\cdot)$ is centered around $\overline{Z_i}$, and represents the uncertainty about $\overline{Z_i}$ for application purposes. The value of $\overline{Z_i}$ is perturbed from $\overline{X_i}$ using the distribution $g_i(\cdot)$, which is exactly the same as $f_i(\cdot)$, except that it is centered around $\overline{X_i}$ instead of $\overline{Z_i}$.

The relationship between $f_i(\cdot)$ and $g_i(\cdot)$ reflects our level of uncertainty about the perturbed representation $\overline{Z_i}$. For example, if the gaussian distribution is used for transformation to the uncertain data model, then $f_i(\cdot)$ and $g_i(\cdot)$ are defined as follows:

$$f_i(x) = \frac{1}{(\sqrt{2 \cdot \pi} \cdot \sigma_i))^d} e^{-\frac{||x - \overline{Z_i}||^2}{2 \cdot \sigma_i^2}} \tag{1}$$

$$g_i(x) = \frac{1}{(\sqrt{2 \cdot \pi} \cdot \sigma_i))^d} e^{-\frac{||x - \overline{X_i}||^2}{2 \cdot \sigma_i^2}} \tag{2}$$

We note that the level of uncertainty in the data set is governed by the choice of parameters $\sigma_1 \dots \sigma_N$. Only the functions $g_i(\cdot)$ (which approximate $f_i(\cdot)$) are known from the uncertain representation of the data. However, the *true function* $f_i(\cdot)$ used to create the uncertain data cannot be known unless the original data points $\overline{X_i}$ is known. In general, we would like the uncertainty in the function $g_i(\cdot)$ to be sufficient, so that an adversary cannot use the pair $(\overline{Z_i}, f_i(\cdot))$ in order to link it to the original record $\overline{X_i}$. Such linking is typically performed by trying to match records from public databases to the uncertain representation. Specifically, the log-likelihood fit of the uncertain representation $(\overline{Z_i}, f_i(\cdot))$ to any given public database record $\overline{X_i}$ can be used to create a corresponding attack. The log-likelihood fit is a natural choice, because it is directly related to the a-posteriori probability of the uncertain data record $(\overline{Z_i}, f_i(\cdot))$ corresponding to a particular record $\overline{X_i}$.

Let us consider the true record $\overline{X}$. We would like to calculate the likelihood that the uncertain record $(\overline{Z}, f(\cdot))$ corresponds to this true record. In order to do so, the adversary can compute the *potential fit* of the uncertain record to the true record $\overline{X}$. However, we do not even know the function $g_i(\cdot)$ exactly. The uncertain data representation only has *approximations* $(f_1(\cdot) \dots f_N(\cdot))$ of the true functions $(g_1() \dots g_N(\cdot))$ used to add uncertainty to the records, since their means are slightly perturbed in accordance with a k-anonymity requirement. Therefore, we define the *potential perturbation function* $h^{(f(\cdot), \overline{X})}(\cdot)$, which is specific to the true representation $\overline{X}$, which is being tested as a fit to the uncertain data record.

Definition 2.2: The potential perturbation function $h^{(f(\cdot),\overline{X})}(\cdot)$ is the same as the function $f(\cdot)$, except that its mean is replaced by $\overline{X}$.

The potential perturbation function would in fact be the true perturbation function of the data record $(\overline{Z}, f(\cdot))$, if $\overline{Z}$ was indeed generated from $\overline{X}$. Thus, the potential perturbation function can be considered a *conditional density function*, under the assumption that $\overline{Z}$ was generated from $\overline{X}$. As we will see, this will be useful in providing an a-posteriori probability interpretation to the log-likelihood fit. The potential perturbation function is useful in computing the likelihood that the true representation of a record (obtained from a public database or other source) corresponds to the uncertain record $(\overline{Z}, f(\cdot))$. This is achieved by computing the log-likelihood fit of the uncertain record $(\overline{Z}, f(\cdot))$ using the potential perturbation function. We formalize the concept of *potential fit* using the log-likelihood criterion:

Definition 2.3: The potential fit $\mathcal{F}(\overline{Z}, f(\cdot), \overline{X})$ of the uncertain data record $(\overline{Z}, f(\cdot))$ to the record $\overline{X}$ is given by $\log(h^{(f(\cdot),\overline{X}))}(\overline{Z}))$.

The higher the value of the log-likelihood fit, the greater the probability that the true record $\overline{X}$ corresponds to the uncertain record $(\overline{Z}, f(\cdot))$. We observe that the log likelihood fit is an indirect representation of the Bayes a-posteriori probability that the uncertain data record $(\overline{Z}, f(\cdot))$ fits a particular record $\overline{X}$. This is because the potential perturbation function is simply the conditional density function under the assumption that $(\overline{Z}, f(\cdot))$ fits the record $\overline{X}$. We formalize this observation as follows:

Observation 2.1: Consider a database $\mathcal{D}_p$ which is known to contain the true representation of the uncertain record $(\overline{Z}, f(\cdot))$ with equal a-priori probability. Then, the posterior probability $\mathcal{B}(\overline{Z}, \overline{X}, \mathcal{D}_p)$ of a particular record $\overline{X} \in \mathcal{D}_p$ to correspond to $\overline{Z}$ is given by:

$$\mathcal{B}(\overline{Z}, f(\cdot), \overline{X}, \mathcal{D}_p) = \frac{e^{\mathcal{F}(\overline{Z}, f(\cdot), \overline{X})}}{\sum_{\overline{V} \in \mathcal{D}_p} e^{\mathcal{F}(\overline{Z}, f(\cdot), \overline{V})}} \tag{3}$$

The above observation can be verified by applying the Bayes formula in conjunction with equal a-priori probability of each data record corresponding to $(\overline{Z}, f(\cdot))$. Thus, the log likelihood is an indirect representation of the Bayes probability, and the use of this particular representation is chosen for the sake of numerical and algebraic convenience.

The log-likelihood fit can be used to define a k-anonymity analysis for the uncertain data model. Consider an uncertain data record $(\overline{Z}, f(\cdot))$, which corresponds to $\overline{X}$. In general, it may be the case, that other spurious records in the uncertain database $\mathcal{D}_p$ may have a higher log-likelihood fit to $\overline{X}$ than $\overline{Z}$ itself. This reduces the risk of disclosure of the true identity of $\overline{X}$ with the use of adversarial attacks from public databases. We define k-anonymity for the uncertain data model as one in which the expected number of records in $\mathcal{D}$ which have higher (or equal) log likelihood fit to $(\overline{Z}, f(\cdot))$ than $\overline{X}$ is at least k. In such a case, public databases cannot be easily used to distinguish $\overline{X}$, since $\overline{X}$ may be among any of the k best fits to $(\overline{Z}, f(\cdot))$, or may not even be among the k best fits. We will formally define the concept of k-anonymity in the uncertain data model.

Definition 2.4: (**k-anonymity for the uncertain data model**) A record $(\overline{Z}, f(\cdot)) \in \mathcal{D}_p$ with original representation $\overline{X}$ is said to be k-anonymous in expectation, when for some random variable r, there are r records $\{\overline{X_1} \dots \overline{X_r}\} \in \mathcal{D}$ for which the following is true:

$$\mathcal{F}(\overline{Z}, f(\cdot), \overline{X}) \leq \mathcal{F}(\overline{Z}, f(\cdot), \overline{X_i}) \quad \forall i \in \{1, \dots r\} \tag{4}$$

$$E[r] \geq k \tag{5}$$

The uncertain record $(\overline{Z}, f(\cdot))$ cannot be used to distinguish its true representation $\overline{X}$ from the r records $\overline{X_1} \dots \overline{X_r}$ in $\mathcal{D}$. We further note that while the k-anonymity is in expectation, it is actually quite a strong definition, since (unlike deterministic models) the randomization in the data adds an additional level of uncertainty for the adversary. In such cases, an adversary cannot be sure that $(\overline{Z}, f(\cdot))$ corresponds to its true representation $\overline{X}$ (any more than the next $(k-1)$ fits), even in the extreme case when $(\overline{Z}, f(\cdot))$ has a better fit to $\overline{X}$ than any other record in the database. In fact, when the expected anonymity level is k, an adversary may not even be able to distinguish between the $2 \cdot k$ best fits to a given data point, since the true correspondence may occur at the $2 \cdot k$th best fit for many data points. Simply stated, the uncertainty in the rank of the fit helps in better privacy preservation. Next, we define the concept of k-anonymity for the entire database.

Definition 2.5: An uncertain database $\mathcal{D}_p$ is k-anonymized, if every record in it is k-anonymized in expectation.

Next, we will discuss the process of k-anonymization of the data in the uncertain data model. We will discuss two models corresponding to the gaussian and uniform data distributions respectively. We note that these are natural models from a privacy perspective, though the broad approach for anonymization extends to many different models.

A. Gaussian Uncertainty Model

In the gaussian model, we assume that the density function $f_i(\cdot)$ associated with the data point $\overline{Z_i}$ is as follows:

$$f_i(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_i})^d} e^{-\frac{||x - \overline{Z_i}||^2}{2 \cdot \sigma_i^2}} \tag{6}$$

Here σ_i^2 is the variance of the spherically symmetric gaussian distribution in any direction. While we will perform this analysis for spherically symmetric gaussians for simplicity, we will see that the approach and analysis can easily be extended to nonsymmetric distributions. The true representation for $\overline{Z_i}$ is $\overline{X_i}$. We wish to determine the probability that for some $j \neq i$, the fit of $\overline{X_j}$ to $\overline{Z_i}$ is at least equal to that of $\overline{X_i}$.

Lemma 2.1: Let $(\overline{Z_i}, f_i(\cdot))$ be the uncertain representation of $\overline{X_i}$ in the d-dimensional database $\mathcal{D}$. Let $f_i(\cdot)$ be a gaussian distribution with mean $\overline{Z_i}$ and variance σ_i^2. Let δ_{ij} be the euclidian distance between $\overline{X_i}$ and $\overline{X_j}$. Then, for any $j \neq i$, the probability that the fit of $\overline{Z_i}$ to $\overline{X_j}$ is at least equal to that of $\overline{Z_i}$ to $\overline{X_i}$ is given by:

$$P(\mathcal{F}(\overline{Z_i}, f(\cdot), \overline{X_j}) \geq \mathcal{F}(\overline{Z_i}, f(\cdot), \overline{X_i})) = P(M \geq \delta_{ij}/(2 \cdot \sigma_i)) \tag{7}$$

Here M is a normal random variable with zero mean and unit variance.

Proof: The first step is to find the potential perturbation function of $f(\cdot)$ with respect to $\overline{X_i}$ and $\overline{X_j}$, and determine the log-likelihood fit of the data point $\overline{Z_i}$ to $\overline{X_i}$ and $\overline{X_j}$ respectively. From Definition 2.3, these values are $\log(h^{(f_i(\cdot),\overline{X_i})}(\overline{Z_i}))$ and $\log(h^{(f_i(\cdot),\overline{X_j})}(\overline{Z_i}))$ respectively. Therefore, we have:

$$\begin{aligned}
&P(\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_j}) \geq \mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_i})) = \\
&= P(\log(h^{(f_i(\cdot),\overline{X_j})}(\overline{Z_i})) \geq \log(h^{(f_i(\cdot),\overline{X_i})}(\overline{Z_i}))) \\
&= P\left(-\frac{||\overline{Z_i} - \overline{X_j}||^2}{2 \cdot \sigma_i^2} \geq -\frac{||\overline{Z_i} - \overline{X_i}||^2}{2 \cdot \sigma_i^2}\right) \\
&= P(||\overline{Z_i} - \overline{X_i}||^2 \geq ||\overline{Z_i} - \overline{X_j}||^2) \\
&= P(||\overline{Z_i} - \overline{X_i}||^2 \geq ||(\overline{Z_i} - \overline{X_i}) + (\overline{X_i} - \overline{X_j})||^2) \\
&= P((\overline{Z_i} - \overline{X_i}) \cdot (\overline{X_j} - \overline{X_i}) \geq ||X_i - X_j||^2/2) \\
&= P((\overline{Z_i} - \overline{X_i}) \cdot (\overline{X_j} - \overline{X_i}) \geq \delta_{ij}^2/2)
\end{aligned}$$

We note that the expression $(\overline{Z_i} - \overline{X_i}) \cdot (\overline{X_j} - \overline{X_i})$ is the dot product of the spherically symmetric gaussian $(\overline{Z_i} - \overline{X_i})$ with a line segment $(\overline{X_j} - \overline{X_i})$ of length δ_{ij}. This is a random variable drawn from the normal distribution with standard deviation $\sigma_i \cdot \delta_{ij}$ since the projection of a spherically symmetric gaussian on any unit axis creates a normal distribution with variance σ_i^2. This random variable can be represented as $M \cdot \sigma_i \cdot \delta_{ij}$, where M is a normal random variable with zero mean and unit variance. Therefore, continuing the above sequence of equations, by substituting for the dot product on the left hand side of the last equation, we get:

$$\begin{aligned}
&P(\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_j}) \geq \mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_i})) = \\
&= P(M \cdot \sigma_i \cdot \delta_{ij} \geq \delta_{ij}^2/2) \\
&= P(M \geq \delta_{ij}/(2 \cdot \sigma_i))
\end{aligned}$$

This concludes the proof. ■

Theorem 2.1: Let $(\overline{Z_i}, f_i(\cdot))$ be the uncertain representation of $\overline{X_i}$ in the d-dimensional database $\mathcal{D}$ with N data points $\{\overline{X_1} \ldots \overline{X_N}\}$. Let $f_i(\cdot)$ be a gaussian distribution with mean $\overline{Z_i}$ and variance σ_i^2. Let δ_{ij} be the euclidian distance between $\overline{X_i}$ and any other data point $\overline{X_j}$. Then, the expected anonymity level $A(\overline{X_i}, \mathcal{D})$ of the point $\overline{X_i}$ with respect to the data set $\mathcal{D}$ is given by:

$$A(\overline{X_i}, \mathcal{D}) = \sum_{\overline{X_j} \in \mathcal{D}} P(M_j \geq \delta_{ij}/(2 \cdot \sigma_i)) \tag{8}$$

Here $M_1 \ldots M_N$ are each normal random variables with zero mean and unit variance.

Proof: Let $\mathcal{I}_{ij}$ be the indicator variable indicating whether or not $\overline{X_j}$ is at least as good a fit to $\overline{Z_i}$ as $\overline{X_i}$. Specifically, the indicator variable is 1, if $\overline{X_j}$ has at least as high a log-likelihood fit, and zero otherwise. Then, the expected anonymity level $A(\overline{X_i}, \mathcal{D})$ of the point $\overline{X_i}$ with respect to the data set $\mathcal{D}$ is given by:

$$A(\overline{X_i}, \mathcal{D}) = \sum_{j=1}^{N} E[\mathcal{I}_{ij}] \tag{9}$$

Now, we note that the expected value of $E[\mathcal{I}_{ij}]$ is equal to $P(\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_j}) \geq \mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_i}))$. By using the relationship in Lemma 2.1 for this expression, we get the desired result. ■

We note that the key parameter in the gaussian distribution which needs to be determined for each data point $\overline{X_i}$ is the standard deviation σ_i. The result of Theorem 2.1 gives us a natural relationship which can be used to determine the value of σ_i for a desired anonymity level. Specifically, if the target value of $A(\overline{X_i}, \mathcal{D})$ is k_0, then we need to determine σ_i, so that the following relationship is true:

$$\sum_{\overline{X_j} \in \mathcal{D}} P(M_j \geq \delta_{ij}/(2 \cdot \sigma_i)) = k_0 \tag{10}$$

If σ_i were known, then the value of each expression $P(M_j \geq \delta_{ij}/(2 \cdot \sigma_i))$ on the right hand side may be determined with the use of the cumulative normal distribution function. Furthermore, the expression $\sum_{\overline{X_j} \in \mathcal{D}} P(M_j \geq \delta_{ij}/(2 \cdot \sigma_i))$ is monotonic with σ_i. This suggests a natural iterative binary search method in order to determine the value of σ_i.

First, we determine the values of δ_{ij} for different values of i and j. For a given value of i, let δ_{iq} be the largest euclidian distance among all distances from i to other points. The value $10 \cdot \delta_{iq}$ is an overestimate on the value of σ_i, since it results in an anonymity level which is almost equal to N. Next, we need to determine an underestimate on the value of σ_i for binary search purposes. Let δ_{ir} be the distance to its nearest neighbor. Let s be such that $P(M > s) = (k-1)/(N-1)$, where M is a normal distribution with zero mean and unit variance. Then, if we choose σ_i to be $L = \delta_{iq}/(2 \cdot s)$, then L is an under-estimate on the value of σ_i. We formalize this result as follows:

Theorem 2.2: Consider the data point $\overline{X_i}$ for which we wish to find a k-anonymous representation from a database of N points. Let δ_{ir} be the distance to its nearest neighbor $\overline{X_r}$. Let s be such that $P(M > s) = (k-1)/(N-1)$ for the normal random variable M with zero mean and unit variance. Then $L = \delta_{iq}/(2 \cdot s)$ is an underestimate on the value of σ_i in order to provide k-anonymity for the uncertain representation of $\overline{X_i}$.

Proof: From equation 10, we would like to choose σ_i such that:

$$k = \sum_{\overline{X_j} \in \mathcal{D}} P(M_j \geq \delta_{ij}/(2 \cdot \sigma_i)) \tag{11}$$

However, if we choose $\sigma_i = L$, then the expression on the right hand side is equal to $\sum_{j=1}^{N} P(M > s \cdot \delta_{ij}/\delta_{ir})$. However, since δ_{ir} is the distance to the nearest neighbor of i, we know that for each value of $j \neq i$ we have:

$$\begin{aligned}
\sum_{j=1}^{N} P(M > s \cdot \delta_{ij}/\delta_{ir}) &= \sum_{j \neq i} P(M > s \cdot \delta_{ij}/\delta_{ir}) + 1 && (12) \\
&\leq \sum_{j=1}^{N-1} P(M > s) + 1 && (13) \\
&= (N-1) \cdot (k-1)/(N-1) + 1 = k && (14)
\end{aligned}$$

Since the value of $\sigma_i = L$ provides an anonymity level less than k, it follows that L is an underestimate on the desired value of σ_i. ■

In the above section, we established both a lower and upper bound on the value of σ_i. Therefore, the true value of σ_i must lie in the range $[L, 10 \cdot \delta_{iq}]$. Now, we can use the monotonicity of the function $\sum_{\overline{X_j} \in \mathcal{D}} P(M_j \geq \delta_{ij}/(2 \cdot \sigma_i))$, in conjunction with binary search on different values of σ_i in order to determine its final value to any degree of desired accuracy. One interesting observation about this model is that the value of σ_i is determined independently for each data point $\overline{X_i}$, and does not affect the anonymity behavior of the other data points. This is quite different from deterministic k-anonymization models in which the transformation of one data point affects the other data points. This is a clear advantage of the uncertain data model in case different portions of the data have different required levels of anonymity [13]. In such cases, we can set the different values of σ_i appropriately, so as to reach the desired level of anonymity for the different points.

B. Uniform Uncertainty Model

In the uniform data model, the density function $f_i(\cdot)$ is a cube with side a_i, which is centered around $\overline{Z_i}$. Then, the density function $f_i(x)$ is defined as follows:

$$\begin{array}{ll} f_i(x - \overline{Z_i}) = 1/a_i^d & \text{if each dimension of } (x - \overline{Z_i}) \text{ is} \\ & \text{at most } a_i/2 \text{ in magnitude} \\ \qquad\qquad 0 & \text{otherwise} \end{array}$$

As in the case of the gaussian model, we wish to determine the probability that for some $j \neq i$, the fit of $\overline{X_j}$ to $\overline{Z_i}$ is at least equal to that of $\overline{X_i}$.

Lemma 2.2: Let $(\overline{Z_i}, f_i(\cdot))$ be the uncertain representation of $\overline{X_i}$ in the d-dimensional database $\mathcal{D}$. Let $f_i(\cdot)$ be a uniform distribution which is the form of a cube centered at $\overline{Z_i}$ and with edge length a_i. Let w_{ij}^k be the kth dimensional component of $\overline{X_i} - \overline{X_j}$. Then, for any $j \neq i$, the probability that the fit of $\overline{Z_i}$ to $\overline{X_j}$ is at least equal to that of $\overline{Z_i}$ to $\overline{X_i}$ is given by:

$$P(\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_j}) \geq \mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_i})) = \\ = \frac{\pi_{k=1}^d \max\{a_i - |w_{ij}^k|, 0\}}{a_i^d}$$

Proof: The value of $\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_i}))$ is always $-d \cdot \log(a_i)$. This is because $\overline{Z_i}$ is generated using a cube of length a_i centered at $\overline{X_i}$. Furthermore, $\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_j})$ is either $-d \cdot \log(a_i)$ or $-\infty$ depending upon whether or not $\overline{Z_i}$ lies in the cube with side a_i centered around $\overline{X_j}$. Therefore, the value of $\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_j})$ can never be strictly greater than $\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_i})$, and in order to satisfy the conditions within the probability expression on the left hand side of the Equation in Lemma 2.2, $\mathcal{F}(\overline{Z_i}, f_i(\cdot), \overline{X_j})$ must be exactly equal to $-d \cdot \log(a_i)$.

As discussed earlier, this can be true only if $\overline{Z_i}$ lies in the cube with side a_i centered around $\overline{X_j}$. Since $\overline{Z_i}$ is generated using the cube centered at $\overline{X_i}$, the probability is equal to the fraction defined by the volume common in the intersection of the two cubes centered at $\overline{X_i}$ and $\overline{X_j}$, out of the cube centered at $\overline{X_i}$. The intersection of the two cubes centered at $\overline{X_i}$ and $\overline{X_j}$ forms a cuboid whose kth side is given by $\max\{a_i - |w_{ij}^k|, 0\}$. The volume of this cuboid is given by $\pi_{k=1}^d \max\{a_i - |w_{ij}^k|, 0\}$, and the volume of the underlying cube centered at $\overline{X_i}$ is given by a_i^d. By taking the corresponding fraction, the result follows. ■

Next, we will define the anonymization level of $\overline{X_i}$, as a function of a_i, which can then be used in order to search for the desired value of a_i as in the previous case.

Theorem 2.3: Let $(\overline{Z_i}, f_i(\cdot))$ be the uncertain representation of $\overline{X_i}$ in the d-dimensional database $\mathcal{D}$ with N data points $\{\overline{X_1} \dots \overline{X_N}\}$. Let $f_i(\cdot)$ be a uniform distribution with mean $\overline{Z_i}$ in the form of a cube with side a_i. Let δ_{ij} be the euclidian distance between $\overline{X_i}$ and any other data point $\overline{X_j}$. Then, the expected anonymity level $A(\overline{X_i}, \mathcal{D})$ of the point $\overline{X_i}$ with respect to the data set $\mathcal{D}$ is given by:

$$A(\overline{X_i}, \mathcal{D}) = \sum_{\overline{X_j} \in \mathcal{D}} \frac{\pi_{k=1}^d \max\{a_i - |w_{ij}^k|, 0\}}{a_i^d} \tag{15}$$

Proof: Similar to that of Theorem 2.1, except that we need to use the results of Lemma 2.2 in the final step. ■

As in the previous case, the value of $A(\overline{X_i}, \mathcal{D})$ is monotonic with the chosen value of a_i. Therefore, we can use a binary search approach in order to obtain the least value of a_i which provides k-anonymity.

C. Local Optimizations

The analysis of the previous sections used a spherical gaussian distribution, and a cubic uniform distribution. In most cases, this is a reasonable assumption when we normalize the data to unit variance in each dimension. However, there can be small local variations in the data distribution. Let $\gamma_{i1} \dots \gamma_{id}$ be the corresponding standard deviations along the d dimensions for the k nearest neighbors of $\overline{X_i}$, where k is the anonymity level. Then, we assume that the standard deviation along the jth dimension is given by $q_i \cdot \gamma_{ij}$. In such cases, the gaussian distribution at $\overline{y} = (y^1 \dots y^d)$ for the data point $\overline{X_i} = (x_i^1 \dots x_i^d)$ is as follows:

$$g_i(\overline{y} - \overline{X_i}) = \pi_{j=1}^d \frac{1}{\sqrt{2 \cdot \pi} \cdot q_i \cdot \gamma_{ij}} e^{-\frac{(y^j - x_i^j)^2}{2 \cdot q_i^2 \cdot \gamma_{ij}^2}} \tag{16}$$

In order to optimize for these variations, a small change can be made to the algorithm. For each data point, we normalize the entire data set such that the variance of the k-nearest neighbors along each dimension is the same. In other words, in Equation 16, we use the scaling $y_j' = y_j/\gamma_{ij}$. This will result in a spherically symmetric gaussian on the scaled data, which has already been analyzed in the earlier sections. A similar transformation can also be made for the case of the uniform distribution. The analysis for each data point is then performed with this locally optimized normalization instead of a single global normalization. As a result, the uncertainty model for each data point is more effective in losing less information for the same amount of privacy. In the final data set, the gaussian distributions associated with each point are elliptically shaped with different kinds of elongations in different directions. The same is true for the case of the uniform distribution, in which we now have cuboids which are elongated differently along different directions.

If desired, the analysis can even be extended to the case of arbitrarily oriented gaussian and uniform distributions. This can be done by appropriate point-specific rotation of the axis in conjunction with scaling. We will discuss more detailed optimizations in an extended version of this paper. For this paper, we will show that even the standard models proposed in the paper are more effective than deterministic methods for condensation [1].

D. Application to Query Estimation

Many confidential databases are queried extensively for determining confidential behavior of the underlying data. Two techniques used for privacy-preserving query processing are *query auditing* and *confidentiality control*. In query auditing, we attempt to restrict a subset of the queries, so as to maintain the privacy of the data. On the other hand, in the case of query confidentiality control, we allow inaccurate responses to queries in order to maintain the privacy of the underlying data. A natural approach in confidentiality control methods is to transform the data to a k-anonymous or other representation, since the response to any query derived from such data will also maintain k-anonymity of the underlying database. In this section, we will discuss the problem of query estimation with the use of the uncertain data representation.

Consider a range query $R = [a_1, b_1], [a_2, b_2], \ldots [a_d, b_d]$ for which we wish to estimate the selectivity. In this query, $[a_i, b_i]$ is the range along the ith dimension. Let $\mathcal{S}(R)$ be the set of data points in R. A naive response is to simply use $|\mathcal{S}(R)|$ as the selectivity estimate for the query. This can sometimes be quite inaccurate, especially when the query contains a small number of data points. A more accurate approach is to use the integral of the uncertain region around the data points in the query. Therefore, the query estimate Q is as follows:

$$Q = \sum_{(\overline{Z_i}, f_i(\cdot)) \in \mathcal{D}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \ldots \int_{a_d}^{b_d} f(\overline{x - Z_i}) dx \qquad (17)$$

We note that in the case of the summation, we are using not just the points in $\mathcal{S}(R)$, but the entire database $\mathcal{D}$. This is because points which are just outside the specified range also have a probability of contributing to the corresponding range query. For the case of the gaussian distribution, this expression can easily be computed since the density distribution can be decomposed into the independent gaussians along each dimension. Therefore, the final result is the product of the integrals along the d different dimensions. Let $F_i(\cdot)$ denote the *cumulative function* for the normal distribution corresponding to data point i. Then, the query estimation Q may be defined as follows:

$$Q = \sum_{(\overline{Z_i}, f_i(\cdot)) \in \mathcal{D}} P(\overline{X_i} \cap_j [a_j, b_j]) \qquad (18)$$

$$= \sum_{(\overline{Z_i}, f_i(\cdot)) \in \mathcal{D}} \pi_{j=1}^{d} (F_i(b_j) - F_i(a_j)) \qquad (19)$$

We note that the above expression is even easier to compute for the case of the uniform distribution, since we only need to determine the fraction of the cube range at which a_i and b_i occur.

The bounds can be tightened further by using the known domain ranges of the underlying data set. For this purpose of this application, the domain range for a given dimension is defined as $[l_i, u_i]$, where l_i is the least value for dimension i, and u_i is the maximum value for dimension i. Then, the use of dimension ranges $[l_i, u_i]$ does not violate the uncertain model of k-anonymity, since it does not affect the potential perturbation function $h^{(f(\cdot), \overline{X})}(\cdot)$ for the purpose of calculating log-likelihood fit. On the other hand, it does affect the conditional function $f_i(\cdot)$ for application purposes, and this tightens the accuracy of the estimation.

Without loss of generality, we can assume that $l_i \leq a_i$, and $b_i \leq u_i$. Then the query estimate is defined in terms of the conditional probabilities.

$$Q = \sum_{(\overline{Z_i}, f_i(\cdot)) \in \mathcal{D}} P(\overline{X_i} \in \cap_j [a_j, b_j] | \overline{X_i} \in \cap_j [l_j, u_j]|) \qquad (20)$$

$$= \sum_{(\overline{Z_i}, f_i(\cdot)) \in \mathcal{D}} \pi_{j=1}^{d} \frac{F_i(b_j) - F_i(a_j)}{F_i(u_j) - F_i(l_j)} \qquad (21)$$

We note that this bound is tighter, since it eliminates the underestimation bias associated with the edge effects of spreading the data over a wider range along each dimension.

E. Application to Classification

In many data mining applications, the use of uncertainty information can be useful in improving the quality of the results [10]. For example, in the case of a nearest neighbor classifier, the uncertainty information about the different records can be used in order to perform the classification. In order to do so, we use the log-likelihood fit $\mathcal{F}(\overline{X_i}, f_i(\cdot), \overline{T})$ of each data point $\overline{X_i}$ to the test instance $\overline{T}$, using the uncertainty function $f_i(\cdot)$. As discussed earlier, the value of $e^{\mathcal{F}(\overline{X_i}, f_i(\cdot), \overline{T})}$ represents the Bayes probability that the test instance T fits the data point $\overline{X_i}$. We determine the q best fits to the test instance T. We partition the q best fits among the different classes, and sum up the corresponding probabilities of fit for the different classes. The class with the highest probability of fit was reported as the result for that test instance.

We note that the use of uncertainty information can greatly affect the final result for the class label. A data point $\overline{X_i}$ with a more widely distributed uncertainty function $f_i(\cdot)$ is likely to have lower fit to the test instance $\overline{T}$, than another test instance at the same distance, if this distance itself is small compared to the uncertainty. On the other hand, if the

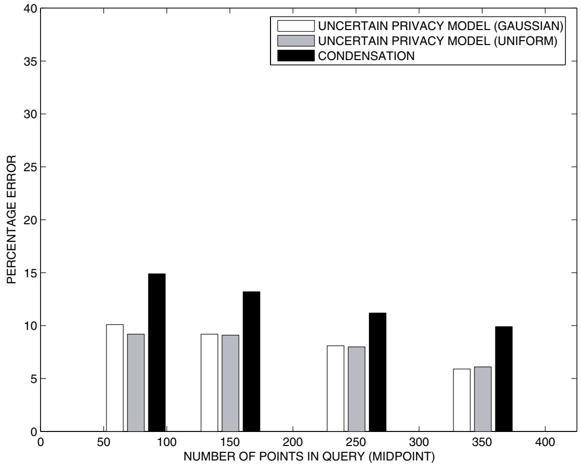

Fig. 1. Query Estimation Error with Increasing Query Size (U10K)

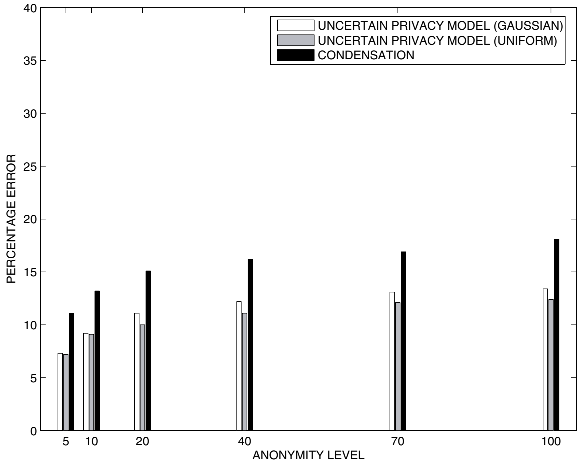

Fig. 2. Query Estimation Error with Increasing Anonymity Level (U10K)

distance is large compared to the uncertainty, then the fit of the more widely distributed function is better. As a result, such a data point may be more or less likely to be among the q best fits to the test instance $\overline{T}$, depending upon the uncertainty function. Such subtle effects of the uncertainty function can greatly influence the final classification behavior of the test instance. In the experimental section, we will test the effect of incorporating uncertainty information in the data, and the relative effectiveness of the two methods.

III. Experimental Results

In this section, we will test the accuracy of the approach for the query estimation and classification problems. In the case of query estimation, we will apply the problem to the method of range selectivity estimation. We will show that the uncertainty based approach is more effective than the condensation method [1] for k-anonymization.

A. Data Sets

We tested the approach on both real and synthetic data sets. The first data set was a uniformly distributed data set

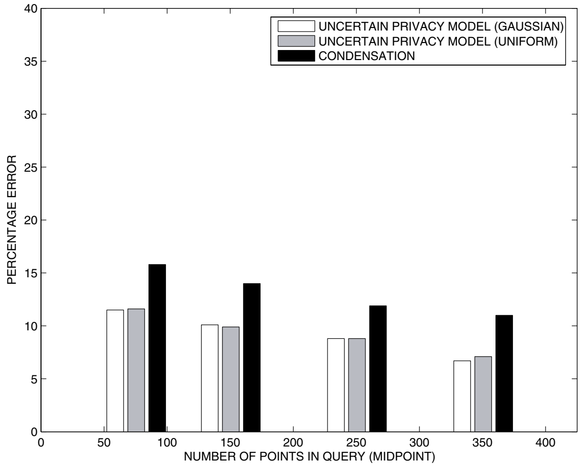

Fig. 3. Query Estimation Error with Increasing Query Size (G20.D10K)

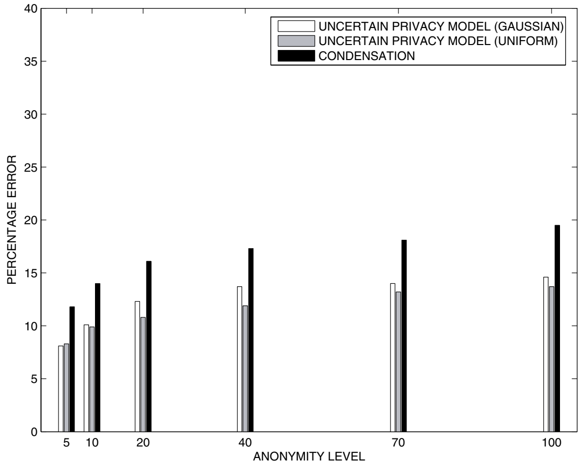

Fig. 4. Query Estimation Error with Increasing Anonymity Level (G20.D10K)

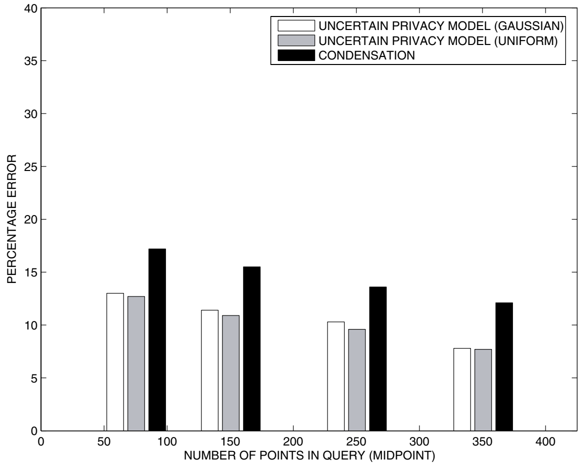

Fig. 5. Query Estimation Error with Increasing Query Size (Adult)

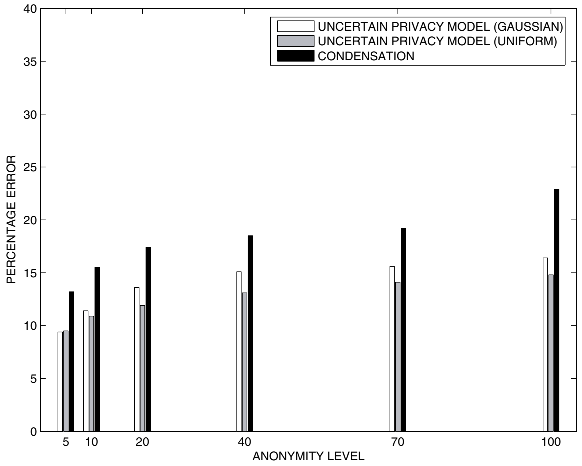

Fig. 6. Query Estimation Error with Increasing Anonymity Level (Adult)

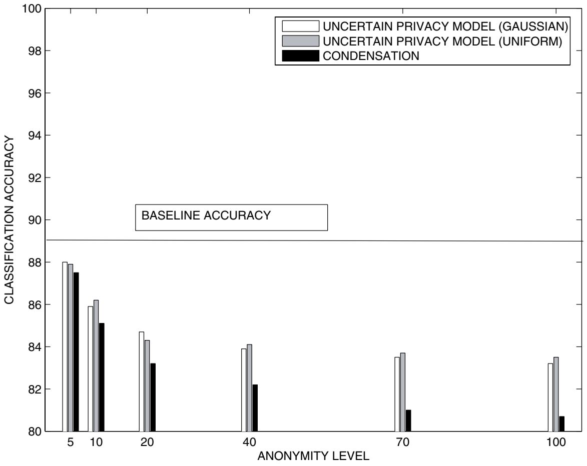

Fig. 7. Classification Accuracy of Data Set G20.D10K

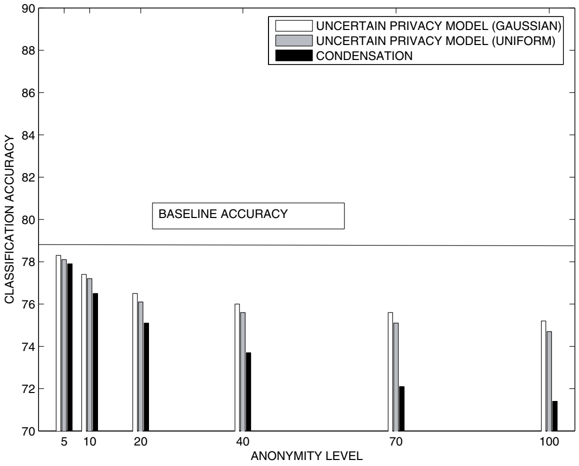

Fig. 8. Classification Accuracy of Adult Data Set

containing 5 dimensions and 10000 data points. Uniform data sets are often quite difficult from a privacy-preservation point of view, because of the inability to find clustered nearest neighbors for anonymization. We refer to this data set as $U10k$ corresponding to the uniformly distributed case.

We also generated a synthetic data set containing clusters. We generated a 5-dimensional synthetic data set containing $r = 20$ clusters. The centers of the clusters were randomly chosen in the unit cube. Each of the clusters was drawn from a gaussian distribution, with a radius randomly chosen in $[0, 0.5]$ along each dimension. The proportional number of points in each cluster were derived by sampling a parameter from a uniform distribution in $[0.5, 1]$. The number of points in each cluster were proportional to the sampled value of the parameter, and 1% of the data points were outliers distributed over the unit cube. A total of $10,000$ data points were generated using the approach. We refer to this data set as $G20.D10k$ corresponding to a data set with 20 clusters and 10000 data points. Since the data set was required to be used for classification, we also created a 2-class data set from $G20.D10K$ by labeling the different data points. For each cluster, we first randomly assigned its class membership, and then labeled the points in it to belong to that class with probability $p = 0.9$. Otherwise the point was assigned to the other class.

The real data set used was the Adult data sets from the UCI machine learning repository. We used all quantitative variables of the Adult data set for the purposes of the experimental evaluation. All data sets were normalized so that the variance along each dimension was 1 unit.

B. *Query Selectivity Estimation*

In this section, we will study the effect of using the method on the query selectivity estimation problem. For the case of selectivity estimation, we used multi-dimensional range queries in the unit cube. The ranges along each dimension were picked randomly, but the queries were classified into different categories depending upon the corresponding selectivity. For each query, we classified it into 4 categories corresponding to its selectivity: (1) 51-100 points (2) 101-200 points (3) 201-300 points and (4) 301-400 points. For each group, we averaged the results over 100 queries. Let S be th etrue selectivity of a query, and S' be the selectivity returned by the estimation method. The error E for each query was defined as follows:

$$E = \frac{|S - S'|}{S} * 100 \qquad (22)$$

In Figure 1, we have illustrated the results for queries of different selectivity and an anonymity level of 10 for the $U10K$ data set. The X-axis of the Figure illustrates the mid point of the query size, whereas the Y-axis illustrates the error of the query, as defined in Equation 22. Thus, there are four midpoints: (1) 75.5 for the query range $[51, 100]$ (2) 150.5 for the query range $[101, 200]$ (3) 250.5 for the query range $[201, 300]$, and (4) 350.5 for the query range $[301, 400]$. These have correspondingly been illustrated on the X-axis. In each case, we have illustrated the error with the use of a uniform distribution for uncertainty modeling, a gaussian distribution for uncertainty modeling, and a condensation approach [1]. It is clear that the relative errors of queries with smaller selectivity are greater. This is consistent with greater statistical robustness of estimating the selectivity of larger queries with the use of stochastic methods.. Another observation is that while the error of the uniform uncertainty function was lower than that of the gaussian uncertainty function, the accuracy of the condensation based approach was the least. This is because the condensation based method often uses principal component analysis using a small number of points. This overfits the true behavior of the underlying data, and increases the errors in the case of the condensation based approach. Furthermore, the uncertain model also uses information about the probability distribution of the data. This is ignored by the condensation model, since the pseudo-data for the condensation model is generated using simplifying assumptions. The additional probabilistic information used by the uncertain model helps in improving its accuracy. In Figure 2, we have illustrated the results with varying anonymity level on the same data set. In this case, we restricted the analysis to queries containing 101-200 data points. It is clear that the accuracy reduces with increasing anonymity level, since greater anonymity also requires greater variance of the uncertain data distributions. However, the reduction in accuracy with increasing anonymity is initially modest, but eventually levels out towards the right end of the graph. This shows a stable and manageable reduction in accuracy with increasing anonymity level. We further note that since the uncertain data transformation of each point is independent of the transformation of the other points, the error figures can be used to approximately extrapolate the behavior of a heterogeneous model in which different data points have different required anonymity level.

We tested the approach for the data set containing gaussian clusters. In Figure 3, we have illustrated the results for the data set $G20.D10K$ for queries of different selectivity. As in the previous case, we fixed the anonymity level at 10. The results are similar to the case of the $U10K$ data set. As in the previous case, the uncertainty modeling methods performed much better than the condensation technique. The relative behavior of queries with different selectivity also showed a similar trend to the previous case, because of greater statistical robustness of larger queries. In Figure 4, we have illustrated the results for data set $G20.D10K$ with increasing anonymity level. These results are for queries with 101-200 data points. As in previous cases, the uncertainty based methods were superior to the condensation based technique. While the error increased with anonymity level, the effectiveness of the approach continues to be retained even when the anonymity level was increased to 100. Finally, the results for the adult data set are illustrated in Figure 5 for the same parameter settings. We fixed the anonymity level at 10, and tested the accuracy of the results for queries of different selectivity. The results are quite similar to the case of the data set $G20.D10K$. We also present the results with varying anonymity level for the Adult data set in Figure 6. As in previous cases, the data set shows a gradual and stable reduction in accuracy with increasing anonymity level. A broad observation is that the trends in both synthetic and real data sets were quite similar and show slow reduction in accuracy with increase in anonymity level.

C. Classification

We also tested the effectiveness of the approach for the classification application. In Figure 7, we have illustrated the effectiveness of the approach for the case of a classification application on the $G20.D10K$ data set. The anonymity level is illustrated on the X-axis, and the classification accuracy is illustrated on the Y-axis. It is clear that in each case, the classification accuracy reduced with increasing anonymity level. In the same graph, we have illustrated the baseline effectiveness of a nearest neighbor classifier on the *original* data set. Since the baseline accuracy is computed without applying any modifications to the data set, it is an optimistic bound on the classification accuracy expected on the uncertain data representation. The results show that while the classification accuracy reduced with increasing anonymity level, the reduction was relatively small over the entire range. Furthermore, the accuracy of the approach was higher than that of condensation with the use of both uncertain data models. This is because the uncertain model used a likelihood fit criterion which adjusted for the different probability distribution around different data points. This was not the case for the condensation approach. Since the condensation approach ignored the probability information, it was unable to provide results which were as accurate as the uncertain data model.

We also tested the results for the case of the adult data set, which was a classification problem containing demographic data of individuals and a binary class corresponding to whether or not the income for the corresponding person was greater than $50K$. The results are illustrated in Figure 8. These results are quite similar to those of the synthetic data set. As in the case of the synthetic data set, the accuracy of the uncertain model was higher than that of the condensation model. The classification accuracy reduced only modestly with increasing anonymity level. The absolute classification accuracy also did not degrade significantly from that obtained by using an exact nearest neighbor classifier on the original data set. The baseline accuracy is illustrated as a horizontal line in Figure 8. The slow degradation in classification accuracy is because of the use of the probabilistic fit on the uncertain data distribution, rather than the use of exact distances. This ensures that each data point is treated differently corresponding to its uncertainty distribution, which improves the accuracy of the classification process.

IV. Conclusions and Discussion

In this paper, we proposed an uncertainty model for k-anonymity. This approach unifies the data models for privacy-preserving and uncertain data mining. This is a key advantage over other privacy-preservation approaches which use diverse representations of the underlying data, each of which requires a different method for processing. On the other hand, since uncertain data management has been widely studied, many of these methods can be used directly with the probabilistic privacy-preserving transformation. We analyzed the gaussian and uniform distributions for uncertainty modeling, and used this analysis to present a technique for performing the transformation to the uncertain privacy-preserving model. This transformation satisfies the k-anonymity condition under the uncertain data model. The uncertain data model has the advantage of retaining the distribution information about the uncertainty, which improves its effectiveness for a variety of applications. Furthermore, the uncertainty introduced in the data for the adversary has the effect of adding privacy to the model. As a result, the approach has the advantage of allowing lower perturbations to the data for the same level of privacy as the deterministic model. We applied the method to the query estimation and classification problems, and illustrated the effectiveness of the uncertain k-anonymity method over the condensation-based approach. Thus, the approach is not only more practical because of its standardized representation, but also allows greater effectiveness over a variety of applications.

REFERENCES

[1] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacy-preserving data mining," in *EDBT Conference Proceedings*, 2004, pp. 183–199.

[2] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM SIGMOD Conference Proceedings*, 2000, pp. 70–81.

[3] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy-preserving data mining algorithms," in *ACM PODS Conference Proceedings*, 2001, pp. 247–255.

[4] A. Machanvajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," in *ICDE Conference Proceedings*, 2006, p. 24.

[5] S. Rizvi and J. Haritsa, "Maintaing data privacy in association rule mining," in *VLDB Conference Proceedings*, 2002, pp. 682–693.

[6] P. Samarati, "Protecting respondents identities in micro-data release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[7] D. Burdick, P. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, "Olap over uncertain and imprecise data," in *VLDB Conference Proceedings*, 2005, pp. 123–144.

[8] L. V. S. Lakshmanan, N. Leon, R. Ross, and V. S. Subrahmanian, "Probview: A flexible database system," *ACM Transactions on Database Systems*, vol. 22, no. 3, pp. 419–469, 1997.

[9] S. I. McClean, B. W. Scotney, and M. Shapcott, "Aggregation of imprecise and uncertain information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 902–912, 2001.

[10] H.-P. Kriegel and M. Pfiefle, "Density-based clustering of uncertain data," in *ACM KDD Conference Proceedings*, 2005, pp. 672–677.

[11] J. Traub, Y. Yemini, and H. Wozniakowski, "The statistical security of a statistical database," *ACM TODS Journal*, vol. 9, no. 4, pp. 672–679, 1984.

[12] J. Kim and W. Winkler, "Masking micro-data files," in *Bureau of the Census*, 1997.

[13] X. Xiao and Y. Tao, "Personalized privacy-preservation," in *ACM SIGMOD Conference Proceedings*, 2006, pp. 229–240.