CERIAS Tech Report 2009-16

Modeling and Integrating Background Knowledge in Data Anonymization

by Tiancheng Li; Ninghui Li; Jian Zhang
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

Modeling and Integrating Background Knowledge in Data Anonymization

Tiancheng Li Ninghui Li
Department of Computer Science
Purdue University
{li83,ninghui}@cs.purdue.edu

Jian Zhang
Department of Statistics
Purdue University
jianzhan@stat.purdue.edu

***Abstract*— Recent work has shown the importance of considering the adversary's background knowledge when reasoning about privacy in data publishing. However, it is very difficult for the data publisher to know exactly the adversary's background knowledge. Existing work cannot satisfactorily model background knowledge and reason about privacy in the presence of such knowledge.**

This paper presents a general framework for modeling the adversary's background knowledge using kernel estimation methods. This framework subsumes different types of knowledge (e.g., negative association rules) that can be mined from the data. Under this framework, we reason about privacy using Bayesian inference techniques and propose the skyline (B, t)-privacy model, which allows the data publisher to enforce privacy requirements to protect the data against adversaries with different levels of background knowledge. Through an extensive set of experiments, we show the effects of probabilistic background knowledge in data anonymization and the effectiveness of our approach in both privacy protection and utility preservation.

I. Introduction

A number of privacy models have been proposed for data anonymization, e.g., k-anonymity [1], ℓ-diversity [2], t-closeness [3], and so on. A key limitation of these models is that they cannot guarantee that the sensitive attribute values of individuals are protected when the adversary has additional knowledge (called background knowledge). Background knowledge can come from diverse sources, such as well-known facts, demographic information, public records, and information about specific individuals.

As an example, consider that a hospital has the original patient table T in Table I(a), which contains three attributes *Age*, *Sex*, and *Disease*. The hospital releases a generalized table T^* in Table I(b) which satisfies 3-diversity. Assume that an adversary knows Bob is a 69-year-old male whose record is in the table, the adversary can only find out that Bob is one of the first three records. Without any additional knowledge, the adversary's estimate of the probability that Bob has *Emphysema* is $1/3$. However, the adversary may know the correlations between *Emphysema* and the non-sensitive attributes *Age* and *Sex*, e.g., "the prevalence of emphysema was appreciably higher for the 65 and older age group than the 45-64 age group for each race-sex group" and "the prevalence was higher in males than females and in whites than blacks". [1]

[1]From a data fact sheet published by National Heart, Lung, and Blood Institute ($http://www.nhlbi.nih.gov/health/public/lung/ther/copd_fact.pdf$).

Age	Sex	Disease
69	M	Emphysema
45	F	Cancer
52	F	Flu
43	F	Gastritis
42	F	Flu
47	F	Cancer
50	M	Flu
56	M	Emphysema
52	M	Gastritis

(a) Original table T

Age	Sex	Disease
$[45-69]$	*	Emphysema
$[45-69]$	*	Cancer
$[45-69]$	*	Flu
$[40-49]$	F	Gastritis
$[40-49]$	F	Flu
$[40-49]$	F	Cancer
$[50-59]$	M	Flu
$[50-59]$	M	Emphysema
$[50-59]$	M	Gastritis

(b) Generalized table T^*

TABLE I
ORIGINAL TABLE AND ITS GENERALIZED TABLE

Because Bob is a 69-year-old male, then based on the above external knowledge, the adversary can infer that Bob has a much larger probability of having *Emphysema* than the other two tuples in the first group.

In the above example, the adversary knows the correlations between *Emphysema* and the attribute *Age* (and *Sex*). We call this *correlational knowledge*. In general, correlational knowledge describes the relationships between the sensitive attribute and the non-sensitive attributes, e.g., male does not have *ovarian cancer*. Correlational knowledge is one kind of adversarial background knowledge.

Integrating background knowledge into privacy quantification has been recently studied [4], [5], [6], [7]. They propose different approaches (a formal language [4], [5] or ME constraints [7]) for expressing background knowledge and analyze the privacy risk when the adversary has a certain amount of knowledge. These works, however, are unaware of the exact background knowledge possessed by the adversary. The injector approach [6] considers only background knowledge that can be expressed as negative association rules and does not provide an approach to analyze how an adversary can gain sensitive information from the published data.

In this paper, we try to remedy this drawback by proposing a framework for systematically modeling background knowledge and reasoning about privacy in the presence of background knowledge. This is a challenging task since it is very difficult to know exactly the adversary's background knowledge and background knowledge can vary significantly among different adversaries. In this paper, we reduce our scope to background knowledge that is consistent with the data itself. We discuss our rationale for this reduction and present a general framework for modeling consistent back-

ground knowledge. This framework subsumes different types of background knowledge, including correlational knowledge.

A. Motivation

Background knowledge poses significant challenges in defining privacy for the anonymized data [4], [5], [6], [7]. For example, when background knowledge is present, we cannot simply say that no adversary knows any individual's sensitive attribute value after seeing the released data, because there may exist an adversary who already knows the value of an individual. While the adversary still knows the value after seeing the anonymized data, we cannot say that the anonymized data violates privacy. Intuitively, privacy should mean "no matter what background knowledge an adversary has, the adversary cannot learn too much *new* about the sensitive attribute of any individual". This, however, cannot be achieved when an adversary has background knowledge that is inconsistent with the dataset to be released. Consider an adversary who *incorrectly* believes that 80% of the population has a particular disease and has no other more specific information. In reality, only 30% of the population has the disease and this is reflected in the dataset. In this case, even when one releases only the distribution of the sensitive attribute of the table as a whole (without any potentially identifying information), the adversary would have a significant knowledge gain about every individual. Such knowledge gain cannot be prevented by data anonymization, and one can argue that releasing such information is precisely the most important utility of releasing data, namely, to correct widely-held wrong beliefs.

Thus, we have to limit ourselves to consider only background knowledge that is consistent with the data to be released. We come to the following definition:

> Given a dataset T, we say that an anonymized version of T preserves privacy if and only if, for any adversary that has some *background knowledge that is consistent with* T, and for any individual in T, the adversary's *knowledge gain* about the sensitive attribute of the individual is limited.

B. Contributions & Organization

In this paper, we formalize the above intuitive definition. First, we model all background knowledge that is consistent with the original data. We build on our previous work [6], namely, mining background knowledge from the data to be released. Our rationale is that if certain facts or knowledge exist in the data (e.g., males cannot have *ovarian cancer*), they should manifest themselves in the data and we should be able to discover them using data mining techniques. The Injector approach [6], however, considers only negative association rules that hold with 100% confidence. It does not consider probabilistic knowledge or knowledge such as positive association rules, summary statistics, and knowledge from clustering.

Our novel approach in this paper is to apply kernel estimation techniques [8] to model background knowledge that is consistent with a dataset. We model the adversary's prior belief on each individual as a probability distribution, which subsumes different types of knowledge that exists in the data. The dataset can be viewed as samples from such distributions. Our problem of inferring background knowledge from the dataset to be released is similar to the problem of inferring an distribution from samples, a problem well studied in statistics and machine learning. We apply the widely used technique of kernel regression estimation to this problem. The bandwidth of the kernel function provides a good parameter of how much background knowledge an adversary has, enabling us to model adversaries with different levels of background knowledge.

Second, we propose a general formula for computing the adversary's posterior belief based on the background knowledge and the anonymized data. However, the computation turns out to be a hard problem and even known estimation algorithms have too high a complexity to be practical. To overcome the complexity of exact inference, we generalize the approximation technique used by Lakshmanan et al. [9] and propose an approximate inference method called Ω-estimate. We show that Ω-estimate is practical and accurate through experimental evaluation.

Thirdly, we propose a novel privacy model called (B, t)-privacy. We describe our desiderata for quantifying information disclosure (i.e., distance measures between two probabilistic distributions), and show that several existing measures do not satisfy them. We then propose a novel distance measure that can satisfy all the properties.

Fourthly, we empirically show that the worst-case disclosure risk changes continuously with the background knowledge parameter B. In other words, slight changes of the B parameters do not cause a large change of the worst-case disclosure risk. Therefore, while it is difficult to know the adversary's background knowledge when releasing the data, the data publisher only needs to use a set of well-chosen parameters of B to protect the data against all adversaries.

Finally, through extensive experiments on real datasets, we demonstrate that our approach is efficient, prevents background knowledge attacks, and preserves data utility.

The rest of this paper is organized as follows. We present the general framework for modeling background knowledge using kernel estimation techniques in Section II. We describe how to compute posterior beliefs using Bayesian inference techniques in Section III. In Section IV, we define the skyline (B, t)-privacy model, describe the desiderata for quantifying sensitive information disclosure, and present our distance measure. Experimental results are presented in Section V and related work is discussed in Section VI. In Section VII, we conclude the paper and discuss avenues for future research.

II. Modeling Background Knowledge

In this section, we present a general framework for modeling the adversary's background knowledge using kernel regression techniques [8]. This framework is able to incorporate different types of background knowledge that exists in the data. At the end of this section, we analyze the scope of our approach by illustrating the types of background knowledge that can be described in our framework.

A. Knowledge Representation

Let $T = \{t_1, t_2, ..., t_n\}$ be a microdata table maintained by the data publisher where each tuple $t_i (1 \leq i \leq n)$ corresponds to an individual. T contains d quasi-identifier (QI) attributes $A_1, A_2, ..., A_d$ and a single sensitive attribute S. Let $D[A_i]$ $(1 \leq i \leq d)$ denote the attribute domain of A_i and $D[S]$ denote the attribute domain of S (let $D[S] = \{s_1, s_2, ..., s_m\}$). For each tuple $t \in T$, let $t[A_i]$ denote its value on attribute A_i and $t[QI]$ denote its value on the QI attributes, i.e., $t[QI] = (t[A_1], t[A_2], ..., t[A_d])$.

For simplicity of discussion, we consider only one sensitive attribute in our framework. If the data contains multiple sensitive attributes, one can either consider them separately or consider their joint distribution. Our framework can be extended to consider multiple sensitive attributes using any of the above two approaches.

Representation of the Adversary's Prior Belief. Let $D[QI] = D[A_1] \times D[A_2] \times ... \times D[A_d]$ be the set of all possible QI values and $\Sigma = \{(p_1, p_2, ..., p_m) | \sum_{1 \leq i \leq m} p_i = 1\}$ be the set of all possible probability distributions on the sensitive attribute S. We model the adversary's prior belief as a function $P_{pri} : D[QI] \rightarrow \Sigma$. Therefore, for an individual whose QI value is $q \in D[QI]$, the adversary's prior belief of the sensitive attribute values is modeled as a probability distribution $P_{pri}(q)$ over $D[S]$.

An example of prior belief on a tuple t is $P(HIV|t) = 0.05$ and $P(none|t) = 0.95$. In other words, the probability that t has HIV is 0.05 and the probability that t has some non-sensitive disease such as *flu* is 0.95. In our representation, $P_{pri}(t[QI]) = (0.05, 0.95)$.

Representation of the Original Dataset. Each tuple t in table T can be represented as a pair $(t[QI], P(t))$ where $P(t) \in \Sigma$, all components of the distribution $P(t)$ is 0 except the i-th component where $t[S] = s_i$. Formally, $P(t) = (p_1(t), p_2(t), ..., p_m(t))$ is defined as follows: for all $i = 1, 2, ..., m$,

$$p_i(t) = \begin{cases} 1 & \text{if } t[S] = s_i \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the table T can be represented as a set of n pairs: $\{(t_1[QI], P(t_1)), (t_2[QI], P(t_2)), ..., (t_n[QI], P(t_n))\}$. Each pair in our representation is a tuple in the original dataset. Thus, we can view each pair in our representation as a point describing the sensitive value $P(t)$ that a tuple t takes.

Finally, our goal of modeling background knowledge is to calculate estimations of the adversary's prior belief function P_{pri}, which is defined over all possible QI values in $D[QI]$.

B. Estimating the Prior Belief Function

We build on the work of [6] and generate background knowledge by mining the data to be released. The general rationale is that the adversary's background knowledge about the data should be consistent with the data in T and should manifest themselves in T. For example, if the adversary knows that male cannot have ovarian cancer, this piece of knowledge should exist in table T and we should be able to discover it by mining the data in T. We now present a general framework for modeling background knowledge.

The adversary's prior belief function P_{pri} can be considered as the underlying probability distribution of the sensitive attribute in table T. And the data in the original table $\{(t_1[QI], P(t_1)), (t_2[QI], P(t_2)), ..., (t_n[QI], P(t_n))\}$ can be considered as a data sample that is consistent with the unknown prior belief function P_{pri}. Our goal is to find the underlying prior belief function P_{pri} that fits the original data.

One way of constructing an estimate of the P_{pri} function is to use the maximum likelihood estimator (MLE), where the prior belief for each tuple is estimated as the distribution among tuples with that QI value. There are several problems with this approach: (1) the number of distinct QI values can be very large, in which case the MLE estimator is of high variance and does not provide a reliable estimate; (2) the MLE estimator does not have parameters to allow estimation of different P_{pri} functions; and (3) the MLE estimator models each QI value independently and does not consider the semantic meanings among the QI values.

This leads us to the kernel regression estimation method. The kernel regression method is a non-parametrical technique in statistics to estimate the conditional expectation of a random variable. Specifically, given a dataset, the kernel regression method tries to find the underlying function that is best-fit match to the data at those data points. The kernel regression estimator belongs to the smoothing method family. Kernel methods have been extensively studied in the statistics, machine learning, and data mining communities. Existing work has shown that kernel methods have a number of desirable properties: (1) they can estimate the underlying function very effectively and (2) they are simple and efficient to compute. We choose to use kernel regression method to approximate the probability distribution function P_{pri}.

C. Kernel Regression Estimator

Kernel estimation includes two components: (1) the kernel function K and (2) the bandwidth B. The kernel function K describes the form of the weight distribution, generally distributing most of its weight to points that are close to it. The bandwidth B determines the size of the impact ranges of the data point. The probability distribution at a point is estimated as the sum of the smoothed distributions of kernel functions associated with each point in the dataset.

Formally, for one-dimensional data (i.e., $d = 1$), the kernel regression estimation is defined as follows. Give $q \in D[A_1] = D[QI]$, using Nadaraya-Watson kernel weighted average [10], the probability distribution at q is estimated as:

$$\widehat{P}_{pri}(q) = \frac{\sum_{t_j \in T} P(t_j) K(q - t_j[A_1])}{\sum_{t_j \in T} K(q - t_j[A_1])} \quad (1)$$

Note that the denominator is used to normalize the probability distribution.

Thus, the probability distribution $P(t_j)$ of the sensitive attribute for tuple t_j is smoothed by the function $K(.)$ which

peaks at $t_j[A_1]$. This allows for tailoring the estimation problem to the *local* characteristics of the data.

For d-dimensional data, the kernel function is chosen to be the product of d kernel functions $K_i(.)(i = 1, 2, ..., d)$. More formally, given a QI value $q = (q_1, q_2, ..., q_d) \in D[QI]$, the approximate underlying prior belief function P_{pri} is estimated as:

$$\widehat{P}_{pri}(q) = \frac{\sum_{t_j \in T} P(t_j) \prod_{1 \leq i \leq d} K_i(q_i - t_j[A_i])}{\sum_{t_j \in T} \prod_{1 \leq i \leq d} K_i(q_i - t_j[A_i])} \quad (2)$$

where K_i is the kernel function for the i-th attribute A_i. Again, note that the denominator is used to normalized the distribution.

The choice of the kernel function K is not as important as the choice of the bandwidth B. It has been shown by [11] that using different kernel functions K causes only small effects on the accuracy of the estimator as compared with varying the bandwidth B. So preferences are given to the kernels with low computational complexity. We thus choose to use the *Epanechnikov kernel function*, which is widely used in kernel estimation:

$$K_i(x) = \begin{cases} \frac{3}{4B_i}(1 - (\frac{x}{B_i})^2) & \text{if } |\frac{x}{B_i}| < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $B = (B_1, B_2, ..., B_d)$ is the bandwidth vector.

The bandwidth provides a good measurement of how much background knowledge an adversary can have. Specifically, a large B_i implies that the adversary does not have much knowledge about the relationship between the sensitive attribute S and the i-th quasi-identifier A_i. On the contrary, with a small B_i, the adversary is assumed to have more fine-grained knowledge on the distribution of the sensitive attribute with respect to A_i. Therefore, we are able to tune the bandwidth parameters B to model adversaries with different levels of background knowledge.

Finally, we define the distance between two values of an attribute. Assume the attribute domain of A_i is $D[A_i] = \{v_{i1}, ..., v_{ir}\}$ where $r = |D[A_i]|$. The attribute A_i is associated with a $r \times r$ distance matrix M_i where the (j,k)-th cell d_{jk} $(1 \leq j, k \leq r)$ indicates the semantic distance between v_{ij} and v_{ik}. The distance matrix M_i is specified by the data publisher. One way of defining the distance matrix is as follows. If A_i is a continuous attribute, the distance matrix can be defined as:

$$d_{jk} = \frac{|v_{ij} - v_{ik}|}{R_i}$$

where R_i is the range of the attribute A_i, i.e., $R = \max_j\{v_{ij}\} - \min_j\{v_{ij}\}$. If A_i is a categorical attribute, the distance matrix can be defined based on the domain hierarchy of attribute A_i:

$$d_{jk} = \frac{h(v_{ij}, v_{ik})}{H_i}$$

where $h(v_{ij}, v_{ik})$ is the height of the lowest common ancestor of v_{ij} and v_{ik}, and H_i is the height of the domain hierarchy of attribute A_i.

Given parameters B, let $Adv(B)$ denote the parameterized adversary whose background knowledge can be modeled by bandwidth B. In the following, we denote $P_{pri}(B, q)$ as the prior belief of the parameterized adversary $Adv(B)$ on the sensitive attribute of an individual whose quasi-identifier value is $q \in D[QI]$.

D. Scope of the Framework

We demonstrate the scope of the kernel estimation framework (i.e., the amount of background knowledge that can be modeled in the framework). Our framework has three characteristics: (1) we focus on background knowledge that is consistent with the data; (2) we model background knowledge as probability distributions; and (3) we use kernel regression estimator to compute background knowledge. We demonstrate the scope of our framework along these dimensions.

General Privacy Models. Several existing privacy models, such as ℓ-diversity (which requires the sensitive attribute values in each group to be "well-represented"), do not specifically consider the prior belief that an adversary has (we refer such an adversary as the ignorant adversary). This ignorant adversary can be viewed as an adversary with a prior belief that every sensitive attribute value is equally possible for every individual in the data, i.e., $P_{pri}(q) = (\frac{1}{m}, \frac{1}{m}, ..., \frac{1}{m})$ for every $q \in QI$. This knowledge is inconsistent with the data, when the sensitive attribute is not uniformly distributed in the data. Given this background knowledge, the adversary's knowledge gain is unavoidable. Our framework does not model such an adversary. Equation (1) and Equation (2) show that adversaries modeled in our framework always have the correct belief about the overall distribution of the sensitive attribute in the data. This is consistent with the t-closeness model (which requires the distribution P of each group to be analogous to the distribution Q of the whole table with respect to the sensitive attribute). t-Closeness considers the adversary who knows Q from the released data. Our framework can model the background knowledge of this adversary as follows. For each tuple $t_j \in T$, t_j distributes its probability distribution $P(t_j)$ equally to all tuples in the table and therefore, every tuple in the table receives the same share $\frac{1}{n}P(t_j)$. This type of adversary is a special adversary modeled by Equation (2). In Equation (2), the $B_i(1 \leq i \leq d)$ is defined as the range of the domain $D_i(1 \leq i \leq d)$ and the $K_i(1 \leq i \leq d)$ is defined as the uniform function. In other words, $K_i(x) = 1/B_i$ for all $0 \leq x \leq B_i$. Then, Equation (2) reduces to $\widehat{P}_{pri}(q) = \frac{1}{n}\sum_{t_j \in T} P(t_j)$, which is the distribution of the sensitive attribute in the whole table.

Knowledge about Specific Individuals and Relationships among Individuals. We note that our framework does not model all types of background knowledge that an adversary may have. Three types of knowledge have been considered in the literature [5]: (1) knowledge about the target individual which are negative associations, e.g., Tom does not have *Cancer*; (2) knowledge about others which are positive

associations, e.g., Gary has flu; (3) knowledge about same-value families, e.g., {Alice, Bob, Carol} could belong to the same-value family (i.e., if one of them has a sensitive value, all others tend also to have the same sensitive value).

Our framework models background knowledge as probability distributions and does not consider type-3 background knowledge, i.e., knowledge about the relationship between individuals. That is, we make the *tuple-independent* assumption: the sensitive attribute values of the tuples in the table are independent of each other. The first two types of knowledge can be represented using our prior belief functions. For example, if tuple t_j does not have the sensitive value s_i, then the i-th component of the probability distribution $P_{pri}(t_j[QI])$ is 0.

Knowledge about Algorithms and Optimization Objectives. Knowledge about the algorithms and optimization objectives for anonymizing data can be used to help adversaries infer the original data, as shown recently by Wong et al. [12]. This kind of knowledge cannot be modeled using prior belief function about individuals. It is an interesting research direction to study this and other kinds of knowledge that may enable an adversary to breach individuals' privacy.

III. Computing Posterior Belief

When we have modeled the adversary's prior belief about the sensitive attribute of all individuals in the table, we now explain how an adversary changes her belief when she sees the released table using Bayesian inference techniques.

Before we present our approach for computing the posterior belief, we describe how the data can be anonymized. We then give an example showing how an adversary changes her belief when she sees the released table and describe the general formula for computing posterior belief. As exact inference is hard to compute, we propose an approximation inference method called Ω-estimate.

A. Anonymization Techniques

Two widely-studied data anonymization techniques are generalization [13], [1], [14] and bucketization [15], [4], [16]. In generalization, quasi-identifier values are replaced with values that are less-specific but semantically consistent. Bucketization, on the other hand, first partitions tuples into group and then separates the sensitive attribute from the QI attributes by randomly permuting the sensitive attribute values in each bucket.

The main difference between the two anonymization techniques lies in that bucketization does not generalize the QI attributes. When the adversary knows who are in the table and their QI attribute values, the two anonymization techniques become equivalent. When these techniques are used to anonymize the data, the adversary always knows that a group of individuals take a set of sensitive attribute values, but does not know the exact mapping. For example, in the generalized table in Table I(b), the first three tuples $\{t_1, t_2, t_3\}$ form a group and take values $\{Emphysema, Cancer, Flu\}$. But the exact mapping, e.g., which one of the three tuples has *Emphysema*, is unknown. In this paper, we assume that the adversary knows who are in the table and their QI values. In this case, the adversary's goal is to infer the exact mapping between the set of individuals and the set of sensitive attribute values.

Most existing works consider every mapping between these two sets to be equally probable. For example, in the first group of Table I(b), each of the three tuples t_1, t_2, and t_3 is assumed to have a probability of $1/3$ to take *Emphysema*. However, armed with background knowledge, an adversary can make more precise inference, e.g., t_1 will have a much larger probability than $1/3$ to take *Emphysema*. This section provides a study on how to compute these probabilities based on the adversary's background knowledge.

B. An Example

Consider the example shown in Table II(a) where we have a group of three tuples $\{t_1, t_2, t_3\}$ and their sensitive attribute values are $\{none, none, HIV\}$. Suppose that the adversary wants to find out the probability that t_3 takes the HIV disease.

Assume that the adversary has some prior beliefs on the sensitive attribute of tuples in the table as shown in Table II(b). For example, she knows that both t_1 and t_2 have a probability of 5% to take HIV and a probability of 95% to have some non-sensitive disease such as *flu*.

From Table II(a), the adversary knows that exactly one of the three tuples $\{t_1, t_2, t_3\}$ takes HIV. With this in mind, the adversary lists the three possible cases of which tuple takes HIV as shown in Table II(b). In the following, we use $Prob(E)$ to denote the probability the event E occurs.

In case 1, t_3 takes HIV while t_1 and t_2 take the non-sensitive values. Therefore, the probability that case 1 occurs is:

$$\begin{aligned} P(Case\ 1) \propto p_1 &= P(none|t_1) \times P(none|t_2) \times P(HIV|t_3) \\ &= 0.95 \times 0.95 \times 0.3 = 0.271 \end{aligned}$$

Similarly, we obtain:

$$\begin{aligned} P(Case\ 2) \propto p_2 &= P(none|t_1) \times P(HIV|t_2) \times P(none|t_3) \\ &= 0.95 \times 0.05 \times 0.7 = 0.033 \end{aligned}$$

and

$$\begin{aligned} P(Case\ 3) \propto p_3 &= P(HIV|t_1) \times P(none|t_2) \times P(none|t_3) \\ &= 0.95 \times 0.05 \times 0.7 = 0.033 \end{aligned}$$

We are then able to compute $P(Case\ 1)$ as:

$$P(Case\ 1) = \frac{p_1}{p_1 + p_2 + p_3} = 0.8$$

Thus, the posterior probability that t_3 takes HIV is:

$$\begin{aligned} &P(Case\ 1) \times 1 + P(Case\ 2) \times 0 + P(Case\ 3) \times 0 \\ &= P(Case\ 1) = 0.8 \end{aligned}$$

In summary, the adversary's belief that t_3 has HIV changes from 0.3 to 0.8, which is a significant increase. This shows that inferences using probabilistic background knowledge can breach individuals' privacy.

tuple
t_1
t_2
t_3

disease
none
none
HIV

(a) A group of three tuples

t_1	t_2	t_3
$P(HIV\|t_1) = .05$	$P(HIV\|t_2) = .05$	$P(HIV\|t_3) = .3$
$P(none\|t_1) = .95$	$P(none\|t_2) = .95$	$P(none\|t_3) = .7$

(b) The adversary's prior belief table

	t_1	t_2	t_3
Case 1	none	none	HIV
Case 2	none	HIV	none
Case 3	HIV	none	none

(c) The three possible cases

TABLE II

AN EXAMPLE

C. General Formula

We derive the general formula for computing the posterior belief using Bayesian inference techniques (the idea is illustrated in the example above). We consider a group E of k tuples (namely, $E = \{t_1, t_2, ..., t_k\}$). Let the multi-set S denote all sensitive attribute values in E.

In the following, we use $P(s_i|t_j)$ and $P^*(s_i|t_j)$ to denote the prior belief and the posterior belief that tuple $t_j (1 \leq j \leq k)$ takes the sensitive value $s_i (1 \leq i \leq m)$, respectively.

We denote $P(S|E)$ as the likelihood that the tuples in E take the sensitive attribute value in S, which can be computed as the sum of the likelihood of every possible assignments between E and S. For example, consider the tuples in Table II(a), there are three possible assignments as shown in Table II(c):

$$\begin{aligned} &P(\{none,none,HIV\}|\{t_1,t_2,t_3\}) \\ &\quad = P(none|t_1) \times P(none|t_2) \times P(HIV|t_3) \\ &\qquad + P(none|t_1) \times P(HIV|t_2) \times P(none|t_3) \\ &\qquad + P(HIV|t_1) \times P(none|t_2) \times P(none|t_3) \end{aligned}$$

Based on Bayes' rule, the posterior belief $P^*(s_i|t_j)$ is proportional to the product of the prior belief $P(s_i|t_j)$ and the normalized likelihood that the $k-1$ tuples in $E \backslash \{t_j\}$ take the $k-1$ sensitive attribute values in $S \backslash \{s_i\}$:

$$P^*(s_i|t_j) \propto n_i \times \frac{P(s_i|t_j) \times P(S \backslash \{s_i\}|E \backslash \{t_j\})}{P(S|E)} \quad (3)$$

$$= n_i \times \frac{P(s_i|t_j) \times P(S \backslash \{s_i\}|E \backslash \{t_j\})}{\sum_{j'=1}^{k} P(s_i|t_{j'}) \times P(S \backslash \{s_i\}|E \backslash \{t_{j'}\})} \quad (4)$$

where n_i is the frequency of s_i in the multiset S.

We can compute the likelihood $P(S|E)$ by enumerating all possible assignments between E and S. In general, assume that in the multi-set S, the value $s_i (1 \leq i \leq m)$ appears n_i times, the total number of possible assignments is $\frac{k!}{\prod_{i=1}^{m} n_i!}$ where $\sum_{i=1}^{m} n_i = k$.

This shows that computing the exact formula requires exponential computation time. We note that the likelihood $P(S|E)$ is exactly the *permanent* of the matrix where the (i, j)-th cell is the prior probability $P(s_i|t_j)$ (note that each sensitive value in the multiset S holds a column and it will be a $k \times k$ matrix). The problem of computing the permanent is known to be a $\#P$-complete problem. A number of approximation algorithms have been proposed to compute the permanent of a matrix. The state of the art is the polynomial-time randomized approximation algorithm presented in [17]. However, the time complexity is of order of $O(k^{22})$. It is thus not feasible for the general formula to work for a large k. In the following, we turn to approximation algorithms for computing the posterior belief. The approximation algorithm allows us to compute the posterior belief accurately enough while in time linear to the size of the group.

D. Approximate Inferences: Ω-estimate

In the following, we consider a heuristic to estimate the posterior probability $P^*(s_i|t_j)$. We represent the prior beliefs as a bipartite graph where one set of nodes consists of tuples in the group and the other set of nodes consists of sensitive values in the group. Each edge from tuple t_j to sensitive value s_i is associated with the probability $P(s_i|t_j)$.

Our approach is a generalized version of the O-estimate used by Lakshmanan et al. [9], where they estimate the number of correct mappings between original items and anonymized items. In that context, a item either can be linked to an anonymized item or cannot be linked to the anonymized item. In our context, a tuple can be linked to a sensitive attribute value with a certain probability.

Based on the prior belief, t_j can be linked to s_i with a probability of $P(s_i|t_j)$ and $t_{j'}$ can be linked to s_i with a probability of $P(s_i|t_{j'})$ for all $1 \leq j' \leq k$. Therefore, the probability that t_j takes s_i is given by

$$\frac{P(s_i|t_j)}{\sum_{j'=1}^{k} P(s_i|t_{j'})}$$

We call this heuristic the **Ω-estimate** (denoted as $\Omega(s_i|t_j)$). s_i appears n_i times in S and by summing up this probability across all these n_i values, we get an estimation of the posterior probability:

$$\Omega(s_i|t_j) \propto n_i \times \frac{P(s_i|t_j)}{\sum_{j'=1}^{k} P(s_i|t_{j'})}$$

By normalizing the probability ditribution for each t_j, we obtain

$$\Omega(s_i|t_j) = \frac{n_i \times \frac{P(s_i|t_j)}{\sum_{j'=1}^{k} P(s_i|t_{j'})}}{\sum_{r=1}^{m} n_r \times \frac{P(s_r|t_j)}{\sum_{j'=1}^{k} P(s_r|t_{j'})}} \quad (5)$$

The above estimation technique makes the random world assumption [18], where every reasonable mapping between individuals and sensitive attribute values is equally probable. Specifically, Equation (5) can be directly derived from the formula shown in Equation (4) by assuming $P(S - \{s_i\}|E - \{t_j\}) = P(S - \{s_i\}|E - \{t_{j'}\})$ for all $1 \leq j' \leq k$.

In [2], Machanavajjhala et al. studied the problem of calculating the posterior belief under the framework of *generalization* by employing the random world theory. Not surprisingly,

t_1	t_2	t_3
$P(HIV\|t_1) = 0$	$P(HIV\|t_2) = 0$	$P(HIV\|t_3) = .3$
$P(none\|t_1) = 1$	$P(none\|t_2) = 1$	$P(none\|t_3) = .7$

TABLE III

ANOTHER ADVERSARY'S PRIOR BELIEF TABLE

the results they obtained for *generalization* are consistent with our results for *bucketization*.

We note that the Ω-estimate is not exact. Consider the example shown in Table II(a) again where we have a group of three tuples $\{t_1, t_2, t_3\}$ and their sensitive attribute values are $\{none, none, HIV\}$. Now, assume the adversary has different prior beliefs as shown in Table III and she wants to find out the sensitive value that t_3 takes. Using the general formula for exact inference, the probability can be calculated as follows. First, we have $P(\{none, none\}|\{t_1, t_2\}) = 1 \times 1 = 1$ and $P(\{none, HIV\}|\{t_1, t_2\}) = 1 \times 0 + 0 \times 1 = 0$. Therefore we have:

$$P^*(HIV|t_3) = \frac{P(HIV|t_3) \times 1}{P(HIV|t_3) \times 1 + P(none|t_3) \times 0} = 1$$

It is intuitive that t_3 must take the HIV disease because none of t_1 and t_2 can take the HIV disease. However, based on the Ω-estimate, the probability is calculated as:

$$\Omega(HIV|t_3) = \frac{1 \times \frac{0.3}{0.3}}{1 \times \frac{0.3}{0.3} + 2 \times \frac{0.7}{2.7}} = 0.66$$

Here, the inexactness of the Ω-estimate results from the fact that Ω-estimate assigns a uniform likelihood to the following two events: (1) $\{t_1, t_2\}$ take {none,none} and (2) $\{t_1, t_2\}$ take {none,HIV}. However, these two events have very different likelihoods. In fact, the second event cannot occur under the prior beliefs shown in Table III. In general, the Ω-estimate is accurate enough for use in practice. In Section V, the accuracy of the Ω-estimate is empirically evaluated with real datasets.

IV. PRIVACY MODEL WITH BACKGROUND KNOWLEDGE

The next step is to extend privacy definitions for data publishing to consider background knowledge. We define our (B, t)-privacy model and describe our definition of distance measure between two probability distribution, which quantifies the amount of information disclosed by the released table.

A. Privacy Model

Given the background knowledge parameter B and a target individual r whose quasi-identifier value is $q \in D[QI]$, the adversary $Adv(B)$ has a prior belief $P_{pri}(B, q)$ on r's sensitive attribute. When she sees the released table T^*, she has a posterior belief $P_{pos}(B, q, T^*)$ on r's sensitive attribute. The distance of the two probabilistic beliefs measures the amount of sensitive information about individual r that the adversary $Adv(B)$ learns from the released data. Based on this rationale, we define the (B, t)-privacy principle as follows:

Definition 1 (the (B, t)-privacy principle): Given two parameters B and t, an anonymized table T^* is said to have (B, t)-privacy iff the worst-case disclosure risk for all tuples (with QI value being q) in T is at most t:

$$\max_q \mathsf{D}[P_{pri}(B, q), P_{pos}(B, q, T^*)] \leq t$$

where $\mathsf{D}[P, Q]$ is the distance measure of two distributions P and Q.

The parameter B determines the profile of the adversary (i.e., how much background knowledge she has). $B = \{B_1, B_2, ..., B_d\}$ is a d-dimensional vector, which allows the data publisher to specify values for different components of the vector. For example, an adversary may know more information about attribute A_i than about attribute A_j of the table. In this case, we would set a smaller value for B_i than for B_j to accurately model the knowledge of the adversary. On the other hand, the parameter t defines the amount of sensitive information that is allowed to be learned by this adversary.

The above privacy model only protects the data against adversaries with a particular amount of background knowledge B. While this model gives the data publisher the flexibility to specify the parameter B, the main challenge is how to protect the data against all kinds of adversaries with different levels of background knowledge. Of course, the data publisher can enumerate all possible B parameters and enforce the above privacy model for all these B parameters.

In Section V, we empirically show the continuity of the worst-case disclosure risk with respect to the background knowledge parameters, i.e., slight changes of the B parameter do not cause a large change of the worst-case disclosure risk. Therefore, the data publisher only needs to define the privacy model for a set of well-chosen B parameters.

The data publisher can define a set of background knowledge parameters $B_1, B_2, ..., B_r$ and enforce the following skyline (B, t)-privacy principle to protect the data against adversaries with all levels of background knowledge.

Definition 2 (the skyline (B, t)-privacy principle): Given a skyline $\{(B_1, t_1), (B_2, t_2), ..., (B_r, t_r)\}$, an anonymized table T^* satisfies the skyline (B, t)-privacy requirement iff for $i = 1$ to r, the worst-case disclosure risk for all tuples (with QI value being q) in T is at most t_i:

$$\max_q \mathsf{D}[P_{pri}(B_i, q), P_{pos}(B_i, q, T^*)] \leq t_i$$

In practice, the data publisher specifies a set of background knowledge parameters B_i, together with the t_i parameter for each B_i. This allows the data publisher to specify and enforce privacy requirements for different adversaries simultaneously. As we point out above, the worst-case disclosure risk distributes continuously with respect to the background knowledge parameter. This allows the data publisher to use a set of well-chosen background knowledge parameters to protect the data against adversaries with all levels of background knowledge. Also, the data publisher can set default parameters and has the flexibility to define their own parameters for special cases.

B. Distance Measure: Quantifying Information Disclosure

We study the problem of measuring the distance $\mathsf{D}[P, Q]$ between two probabilistic distributions P and Q. The distance measure quantifies the information revealed to an adversary whose prior belief is P and posterior belief is Q.

In this section, we first identify our desiderata for the design of the distance measure and show that existing distance measures cannot satisfy some of these properties. We then define our distance measure that satisfies all of these properties.

1) Desiderata: From our perspective, a useful distance measure should display the following properties:

1) **Identity of indiscernibles:** An adversary has no information gain if her belief does not change. Mathematically, $\mathsf{D}[P, P] = 0$, for any P.
2) **Non-negativity:** When the released data is available, the adversary has a non-negative information gain. Mathematically, $\mathsf{D}[P, Q] \geq 0$, for any P and Q.
3) **Probability scaling:** The belief change from probability α to $\alpha+\gamma$ is more signification than that from β to $\beta+\gamma$ when $\alpha < \beta$ and α is small. $\mathsf{D}[P, Q]$ should consider reflect the difference.
4) **Zero-probability definability:** $\mathsf{D}[P, Q]$ should be well-defined when there are zero probability values in P and Q.
5) **Semantic awareness:** When the values in P and Q have semantic meanings, $\mathsf{D}[P, Q]$ should reflect the semantic distance among different values. For example, for the "Salary" attribute, the value $30K$ is closer to $50K$ than to $80K$. A semantic-aware distance measure should consider this semantics, e.g., the distance between {30K,40K} and {50K,60K} should be smaller than the distance between {30K,40K} and {80K,90K}.

Note that we do not require $\mathsf{D}[P, Q]$ to be a distance metric (the symmetry property and the triangle-inequality property). First, $\mathsf{D}[P, Q]$ does not always have to be the same as $\mathsf{D}[Q, P]$. Intuitively, the information gain from $(0.5, 0.5)$ to $(0.9, 0.1)$ is larger than that from $(0.9, 0.1)$ to $(0.5, 0.5)$. Second, $\mathsf{D}[P, Q]$ can be larger than $\mathsf{D}[P, R] + \mathsf{D}[R, Q]$ where R is also a probabilistic distribution. In fact, the well-known Kullback-Leibler (KL) divergenceis not a distance metric since it is not symmetric and does not satisfy the triangle inequality property. The Kullback-Leibler (KL) divergenceis defined as:

$$KL[P, Q] = \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i}$$

The KL divergence measure is undefined when $p_i > 0$ but $q_i = 0$ for some $i \in \{1, 2, ..., d\}$ and thus does not satisfy the *zero-probability definability* property. To fix this problem, a variation of KL divergence called the Jensen-Shannon (JS) divergence [19] has been proposed. The JS divergence measure is defined as:

$$JS[P, Q] = \frac{1}{2}[\mathsf{D}[P, avg(P, Q)] + \mathsf{D}[Q, avg(P, Q)]] \qquad (6)$$

where $avg(P, Q)$ is the average distribution $(P + Q)/2$ and $KL[,]$ is the KL divergence measure.

However, none of the above distance measures satisfy the *semantic awareness* property. One distance measure that takes value semantics into consideration is the Earth Mover's Distance (EMD) [20], [3]. The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Unfortunately, EMD does not have the *probability scaling* property. For example, the EMD distance between the two distributions $(0.01, 0.99)$ and $(0.11, 0.89)$ is 0.1, and the EMD distance between the two distributions $(0.4, 0.6)$ and $(0.5, 0.5)$ is also 0.1. However, one may argue that the belief change in the first pair is much more significant than that between the second pair. In the first pair, the probability of taking the first value increases from 0.01 to 0.11, a 1000% increase. While in the second pair, the probability increase is only 25%.

2) Distance Measure: We propose a distance measure that can satisfy all the five properties. The idea is to apply kernel smoothing [11] before using JS divergence. Kernel smoothing is a standard statistical tool for filtering out high-frequency noise from signals with a lower frequency variation. Here, we use the technique across the domain of the sensitive attribute value to smooth out the distribution. For computing distance between two sensitive values, we define a $m \times m$ distance matrix for S using the same method as described in Section II-C. The (i, j)-th cell d_{ij} of the matrix indicates the distance between s_i and s_j.

We use the Nadaraya-Watson kernel weighted average:

$$\widehat{p}_i = \frac{\sum_{j=1}^{m} p_j K(d_{ij})}{\sum_{j=1}^{m} K(d_{ij})}$$

where $K(.)$ is the kernel function, which is chosen to be the Epanechnikov kernel as described in Section II. The bandwidth is determined based on the sensitive attribute. In the experiments, we use "Occupation" as the sensitive attribute with a domain hierarchy of height 2, the bandwidth is chosen to be at least 0.5 so that kernel smoothing can be applied.

We then have a smoothed probability distribution $\widehat{P} = (\widehat{p}_1, \widehat{p}_2, ..., \widehat{p}_m)$ for P. The distribution $\widehat{P}$ reflects the semantic distance among different sensitive values.

To incorport semantics into the distance between P and Q, we compute the distance between $\widehat{P}$ and $\widehat{Q}$ as an estimate instead: $\mathsf{D}[P, Q] \approx \mathsf{D}[\widehat{P}, \widehat{Q}]$. The distance $\mathsf{D}[\widehat{P}, \widehat{Q}]$ can be computed using JS-divergence measure (in Equation (6)) which is well-defined even when there are zero probabilities in the two distributions. We can see that our distance measure has all of the five properties described in Section IV-B.1.

V. Experiments

The main goals of the experiments are: (1) to demonstrate the effects of probabilistic background knowledge on data anonymization, (2) to evaluate the accuracy of the Ω-estimate, (3) to illustrate the continuity of the worst-case disclosure risk with respect to the background knowledge parameter B, (4) to show the efficiency of computing (B, t)-private tables , and (5) to show the effectiveness of the (B, t)-privacy model in utility preservation.

	Attribute	Type	# of values
1	Age	Numeric	74
2	Workclass	Categorical	8
3	Education	Categorical	16
4	Marital_Status	Categorical	7
5	Race	Categorical	5
6	Gender	Categorical	2
7	Occupation	Sensitive	14

TABLE IV

DESCRIPTION OF THE ***Adult*** DATASET USED IN THE EXPERIMENT

	k	ℓ	t	b
para1	3	3	0.25	0.3
para2	4	4	0.2	0.3
para3	5	5	0.15	0.3
para4	6	6	0.1	0.3

TABLE V

PRIVACY PARAMETERS USED IN THE EXPERIMENTS

The dataset used in the experiments is the adult dataset from the UC Irvine machine learning repository, which is comprised of data collected from the US census. We use seven attributes of the dataset, as shown in Table IV, where the sensitive attribute is *Occupation*. Tuples with missing values are eliminated and there are about $30K$ valid tuples in total. All algorithms are implemented in Java and the experiments are performed on a 3.4GHZ Pentium 4 machine with 2.0GB of RAM.

Given the dataset, we use the variations of Mondrian multidimensional algorithm [21] to compute the anonymized tables using different privacy models: (1) distinct ℓ-diversity; (2) probabilistic ℓ-diversity; (3) t-closeness; and (4) (B, t)-privacy.

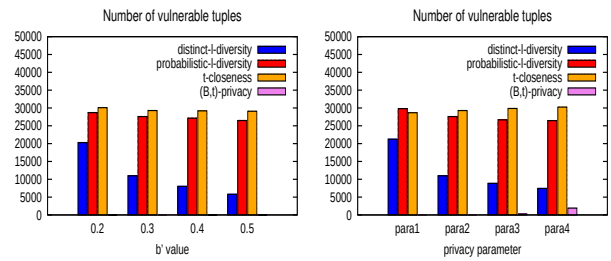
The variations of Mondrian use the original dimension selection and median split heuristics, and check if the specific privacy requirement is satisfied. Note that we can generate the ℓ-diverse table using the anatomizing algorithm [15]. However, Anatomy does not generalize the quasi-identifiers and it would be unfair to compare Mondrian with Anatomy.

The four privacy models protect the data against attribute disclosure. To protect identity disclosure, we also enforce k-anonymity (each group contains at least k records) together with each of the above privacy models.

For each experiment, we evaluate the performance with respect to four sets of privacy parameters in Table V. To make the comparisons easier, we use the same ℓ value for distinct ℓ-diversity and probabilistic ℓ-diversity, the same t for t-closeness and (B, t), the same b value, and $k = \ell$ for all cases as shown in Figure V.

A. Effects of Probabilistic Background Knowledge

We assume that adversary's background knowledge is modeled by the b' parameter, i.e., $B' = (b', b', ..., b')$. To illustrate the effects of probabilistic background knowledge, we apply the prior belief function computed from B' on each of the four anonymized tables, compute the posterior beliefs of each tuple, and report the number of tuples whose privacy is breached under that privacy requirement. These tuples are viewed as vulnerable to the probabilistic background knowledge attacks.



(a) Varied b' values (b) Varied privacy parameters

Fig. 1. Probabilistic Background Knowledge Attack

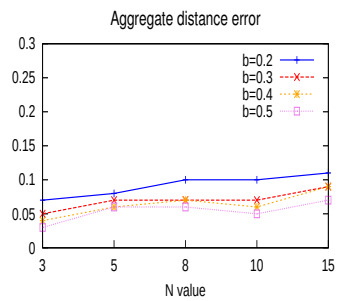


Fig. 2. Accuracy of the Ω-estimate

Our first set of experiments investigates the effect of b' parameter on the number of vulnerable tuples. We fix the privacy parameters $k = \ell = 4$, $t = 0.2$, and $b = 0.3$. Figure 1(a) shows the number of vulnerable tuples in the four anonymized tables with respect to different b' values. As we can see from the figure, the number of vulnerable tuples decreases as b' increases. This is because a larger b' value corresponds to a less-knowledgeable adversary.

The second set of experiment investigates the effect of privacy parameters shown in Table V on the number of vulnerable tuples. We fix the adversary's parameter $b' = 0.3$. Figure 1(b) shows the experimental result.

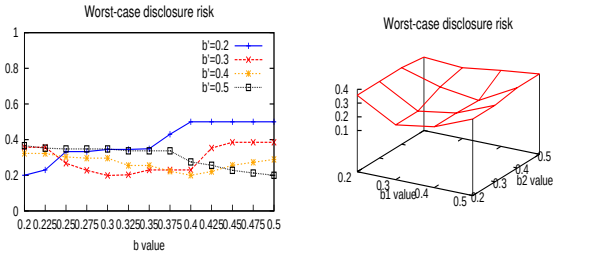
As we can see from these figures, the (B, t)-private table contains much fewer vulnerable tuples in all cases. This shows that the (B, t)-privacy model better protects the data against probabilistic-background-knowledge attacks.

B. Accuracy of the Ω-estimate

To evaluate the accuracy of the Ω-estimate, we randomly pick a group of N tuples from the table and apply both exact inference and the Ω-estimate on the N tuples. Each tuple has a prior distribution P_{pri}, the exact inference distribution P_{exa}, and the Ω-estimate distribution P_{ome}. We then compute the *average distance error*, which is the estimation error averaged over all of the N tuples:

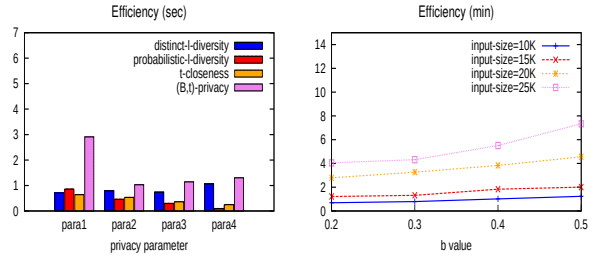
$$\rho = \frac{1}{N} \sum_{j=1}^{N} |\mathsf{D}[P_{exa}, P_{pri}] - \mathsf{D}[P_{ome}, P_{pri}]|$$

We run the experiment 100 times and the average is reported. Figure 2 depicts the *average distance error* with respect to different N values. In all cases, the Ω-estimate is within 0.1-distance with the exact inference. The experiments



(a) Varied b values (b) Varied (b_1, b_2) values

Fig. 3. Continuity of worst-case disclosure risk



(a) Varied privacy models (b) Varied b values

Fig. 4. Efficiency Comparisons

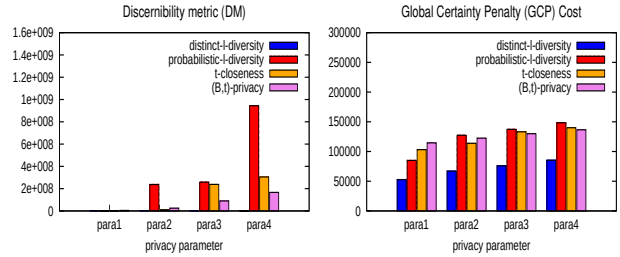
show that the Ω-estimate is accurate enough to be used in practice.

C. Continuity of Disclosure Risk

The goal of this experiment is to show the continuity of the worst-case disclosure risk with regard to the background knowledge parameter B. We first fix the adversary with the background knowledge parameter b' which can be one of the four values $\{0.2, 0.3, 0.4, 0.5\}$. We then generate a set of (B, t)-private tables with different b parameters. For each anonymized table, we compute the worst-case disclosure risk by the adversary. The worst-case disclosure risk is computed as the maximum knowledge gain for all tuples in the table: $\max_q\{\mathsf{D}[P_{pri}(B', q), P_{pos}(B', q, T^*)]\}$. Figure 3(a) shows the results. As we can see from the figure, the worst-case disclosure risk increases/decreases continuously with respect to the b parameter.

We then evaluate the continuity of the disclosure risk with respect to the background knowledge parameters $B = (b_1, b_1, b_1, b_2, b_2, b_2)$, i.e., the adversary's background knowledge on the first three attributes is modeled by b_1 and her background knowledge on the last three attributes is modeled by b_2. Here, we fix the adversary's parameter $b' = 0.3$ and compute the worst-case disclosure risk by the adversary with respect to different (b_1, b_2) values. Figure 3(b) shows the results. As we can see the figures, the worst-case disclosure risks increases/decreases continuously among the domain of (b_1, b_2).

These experiments show that slight changes of the background knowledge parameters will not cause a large change of the worst-case disclosure risk, the conjecture we made in Section IV. This validates our approach of using a set of well-chosen background knowledge parameters to protect the data



(a) Varied privacy models (b) Varied privacy models

Fig. 5. General Utility Measures

against adversaries with all levels of background knowledge.

D. Efficiency

We compare the efficiency of computing the four anonymized tables. We compare the efficiency with regard to different privacy parameters. Figure 4(a) shows the results. As we can see from Figure 4(a), the running time decreases with increasingly stringent privacy requirements because *Mondrian* is a top-down algorithm.

Here, the time to compute the (B, t)-private table does not include the time to run the kernel estimation method to compute the background knowledge. As we can see from Figure 4(a), without considering the time for estimating background knowledge, the running time to compute the (B, t)-private table is roughly the same as the time to compute the other tables, usually within seconds.

We then evaluate the efficiency of computing background knowledge using the kernel estimation method, which is the main efficiency issue of the (B, t)-privacy model. Figure 4(b) shows the results. As we can see from the figures, the time to compute background knowledge is larger than the time to anonymize the data, partially because *Mondrian* runs much faster than many other anonymization algorithms. Moreover, computing background knowledge is still fast enough for large-enough datasets, usually within several minutes.

E. Data Utility

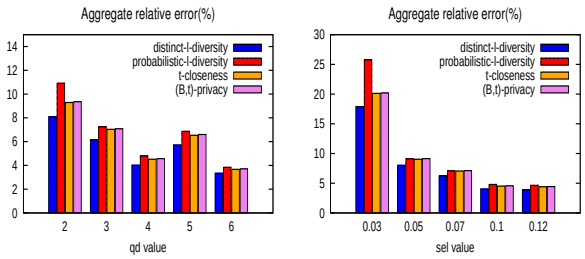
To compare data utility of the four anonymized tables, we evaluate the anonymized data both in terms of general utility measures and accuracy in aggregate query answering.

1) General Utility Measures: We first compare data utility based on two general utility measures: *Discernibility Metric (DM)* [22] and *Global Certainty Penalty (GCP)* [23].

Figure 5(a) shows the DM cost while Figure 5(b) shows the GCP cost for the four anonymized tables. In both experiments, we evaluate the utility measure as a function of the privacy parameters shown in Table V. In both figures, the (B, t)-private table shows comparable utility with the other four anonymized tables.

2) Workload Experiments: We evaluate data utility in terms of performance in aggregate query answering [24], [15], [25].

Figure 6(a) shows the average relative error as a function of the query dimension. As the query dimension increases, average relative error decreases and therefore, the anonymized data performs better for queries with a larger query dimension.



(a) Varied qd values (b) Varied sel values

Fig. 6. Aggregate Query Answering Error

Figure 6(b) shows that as the query selectivity increases, average relative error also decreases. This shows that the anonymized data can answer more accurately on queries with a larger selectivity. In all figures, we can see that the (B, t)-private table can answer queries as accurately as all other anonymized tables.

VI. Related Work

We first review existing work in data anonymization and explain how our technique differs from them. We classify these works into three categories: (1) general privacy models, (2) background knowledge integration, and (3) anonymization techniques. We then examine several research works that have studied background knowledge in other contexts.

General Privacy Models. The k-anonymity model [1], [14] assumes that the adversary has access to some publicly-available databases (e.g., a vote registration list) and the adversary knows who is and who is not in the table. A few subsequent works [2], [26], [27] recognize that the adversary also has knowledge of the distribution of the sensitive attribute in each group. The t-closeness model [3] proposes that the distribution of the sensitive attribute in the overall table should also be public information.

Recently, the σ-presence measure [28] observes that knowing an individual is in the database poses privacy risks. The m-confidentiality model [12] recognizes that knowledge of the mechanism or algorithm of anonymization can leak extra sensitive information. Dynamic dataset re-publication [29], [30], [25], [31] considers the scenario where an adversary has knowledge of previous releases of the dataset. None of these models consider correlational knowledge.

Background Knowledge Integration. In [4], Martin et al. propose a formal language to express background knowledge as implications. They propose the (c, k)-safety model to protect the data in the worst-case when the adversary has knowledge of k implications in the language. Chen et al. [5] extend the framework of [4] by breaking down the adversary's background knowledge into three components which are more intuitive. They then propose a multidimensional approach to protect the data against adversaries with these three types of background knowledge. Du et al. [7] formulate background knowledge as constraints and propose to use maximal entropy modeling to derive the conditional probabilities between the quasi-identifier values and the sensitive attribute values.

These methods provide a framework for defining and analyzing background knowledge, but they are unaware of the exact background knowledge the adversary may have. The injector approach [6] considers negative association rules but it does not model other types of background knowledge and does not provide an approach to analyze how an adversary can gain sensitive information from the published data.

Anonymization Techniques. Most anonymization solutions adopt generalization [13], [14], [32], [22], [33], [34], [35], [36], [21], [37] and bucketization [15], [16], [4], [38], [39]. Other anonymization techniques include clustering [40], [41], [42], [43], space mapping [44], spatial indexing [45], marginals releasing [46], and data perturbation [47], [48]. On the theoretical side, optimal k-anonymity has been proved to be NP-hard for $k \geq 3$ in [49], and approximation algorithms for finding the anonymization that suppresses the fewest cells have been proposed in [49], [50].

Background Knowledge in Other Contexts. The above works focus on data anonymization in the context of privacy-preserving data publishing. A number of research works have examined background knowledge in other contexts. Yang and Li [51] studied the problem of information disclosure in XML publishing when the adversary has knowledge of functional dependencies about the XML data. In [9], Lakshmanan et al. studied the problem of protecting the true identities of data objects in the context of frequent set mining when an adversary has partial information of the items in the domain.

VII. Conclusion and Future Work

In this paper, we present a general framework for modeling and computing background knowledge using kernel methods. We provide efficient techniques to estimate posterior distribution based on the anonymized table and the prior distribution. We present a design of the (B, t)-privacy model, which protects privacy in the presence of adversarial background knowledge. Finally, we show that probabilistic background knowledge is a real concern and demonstrate the effectiveness of our approach through experiments on a real dataset. Here are several future research directions on this topic.

- **Relational Background Knowledge:** Our knowledge representation assumes the *tuple-independent* property and does not model the relationship among individuals, which is an important piece of background knowledge. One example of such kinds of knowledge may be "either Alice or Bob has flu but not both". One way to model relational background knowledge is to use graphs, where nodes represent individuals and edges represent relationships. How to discover such knowledge and how the data publisher can make use of such knowledge in the data anonymization process are interesting problems for future research.
- **Dealing with Other Background Knowledge:** This paper considers background knowledge that can be mined

from the data to be released. In practice, the adversary may have access to additional background knowledge. Wong et al. [12] study how to protect the data against an adversary who has knowledge of the mechanism or algorithm of anonymization. It is interesting to examine other kinds of adversarial knowledge that can lead to privacy breaches.

- **Background Knowledge about Joint Attributes.** Our framework uses the product kernel, which is a simplification of a general multivariate kernel which involves a bandwidth matrix (our model assumes the bandwidth matrix to be diagonal).The general multivariate kernel allows us to express and model knowledge about joint attributes, e.g., knowledge about the joint distribution of the *Age* attribute and the *Sex* attribute $P(S|Age, Sex)$.
- **Uncertain Background Knowledge:** Our framework considers accurate background knowledge. In practice, background knowledge can be vague. For example, an adversary may know that Alice has cancer with some probability in $[0.2, 0.4]$ and has flu with some probability in $[0.7, 0.8]$. One way to represent such uncertain knowledge as probability intervals. We are interested in studying expressions for such knowledge and how to model and integrate uncertain background knowledge in privacy-preserving data publishing.

REFERENCES

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertain. Fuzz.*, vol. 10, no. 5, pp. 557–570, 2002.

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

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in *ICDE*, 2007, pp. 106–115.

[4] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern, "Worst-case background knowledge for privacy-preserving data publishing," in *ICDE*, 2007, pp. 126–135.

[5] B.-C. Chen, R. Ramakrishnan, and K. LeFevre, "Privacy skyline: Privacy with multidimensional adversarial knowledge," in *VLDB*, 2007, pp. 770–781.

[6] T. Li and N. Li, "Injector: Mining background knowledge for data anonymization," in *ICDE*, 2008, pp. 446–455.

[7] W. Du, Z. Teng, and Z. Zhu, "Privacy-maxent: integrating background knowledge in privacy quantification," in *SIGMOD*, 2008, pp. 459–472.

[8] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer-Verlag, 2001.

[9] L. V. S. Lakshmanan, R. T. Ng, and G. Ramesh, "To do or not to do: the dilemma of disclosing anonymized data," in *SIGMOD*, 2005, pp. 61–72.

[10] E. Nadaraya, "On estimating regression," *Theory of Probability and its Applications*, vol. 10, pp. 186–190, 1964.

[11] M. Wand and M. Jones, *Kernel Smoothing (Monographs on Statistics and Applied Probability)*. Chapman & Hall, 1995.

[12] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *VLDB*, 2007, pp. 543–554.

[13] P. Samarati, "Protecting respondent's privacy in microdata release," *TKDE*, vol. 13, no. 6, pp. 1010–1027, 2001.

[14] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertain. Fuzz.*, vol. 10, no. 6, pp. 571–588, 2002.

[15] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in *VLDB*, 2006, pp. 139–150.

[16] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate query answering on anonymized tables," in *ICDE*, 2007, pp. 116–125.

[17] M. Jerrum, A. Sinclair, and E. Vigoda, "A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries," *J. ACM*, vol. 51, no. 4, pp. 671–697, 2004.

[18] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller, "From statistical knowledge bases to degrees of belief," *Artif. Intell.*, vol. 87, no. 1-2, pp. 75–143, 1996.

[19] J. Lin, "Divergence measures based on the shannon theory," *IEEE T. Infom. Theory*, vol. 37, pp. 145–151, 1991.

[20] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[21] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *ICDE*, 2006, p. 25.

[22] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *ICDE*, 2005, pp. 217–228.

[23] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization using local recoding," in *KDD*, 2006, pp. 785–790.

[24] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization," in *KDD*, 2006, pp. 277–286.

[25] X. Xiao and Y. Tao, "M-invariance: towards privacy preserving republication of dynamic datasets," in *SIGMOD*, 2007, pp. 689–700.

[26] ——, "Personalized privacy preservation," in *SIGMOD*, 2006, pp. 229–240.

[27] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *KDD*, 2006, pp. 754–759.

[28] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *SIGMOD*, 2007, pp. 665–676.

[29] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure anonymization for incremental datasets," in *the Third VLDB Workshop on Secure Data Management (SDM)*, 2006, pp. 48–63.

[30] K. Wang and B. C. M. Fung, "Anonymizing sequential releases," in *KDD*, 2006, pp. 414–423.

[31] J.-W. Byun, T. Li, E. Bertino, N. Li, and Y. Sohn, "Privacy-preserving incremental data dissemination," in *JCS*, 2008.

[32] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *KDD*, 2002, pp. 279–288.

[33] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *ICDE*, 2005, pp. 205–216.

[34] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *SIGMOD*, 2005, pp. 49–60.

[35] C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *VLDB*, 2005, pp. 901–909.

[36] T. Li and N. Li, "Optimal k-anonymity with flexible generalization schemes through bottom-up searching," in *PADM*, 2006.

[37] ——, "Towards optimal k-anonymization," *DKE*, vol. 65, no. 1, pp. 22–39, 2008.

[38] G. Ghinita, Y. Tao, and P. Kalnis, "On the anonymization of sparse high-dimensional data," in *ICDE*, 2008, pp. 715–724.

[39] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," in *SIGMOD*, 2008, pp. 473–486.

[40] G. Aggrawal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving anonymity via clustering," in *PODS*, 2006, pp. 153–162.

[41] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," in *DASFAA*, 2007, pp. 188–200.

[42] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," *DKE*, vol. 63, no. 3, pp. 622–645, 2007.

[43] A. Gionis, A. Mazza, and T. Tassa, "k-anonymization revisited," in *ICDE*, 2008, pp. 744–753.

[44] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *VLDB*, 2007, pp. 758–769.

[45] T. Iwuchukwu and J. F. Naughton, "K-anonymization as spatial indexing: Toward scalable and incremental anonymization," in *VLDB*, 2007, pp. 746–757.

[46] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *SIGMOD*, 2006, pp. 217–228.

[47] V. Rastogi, S. Hong, and D. Suciu, "The boundary between privacy and utility in data publishing," in *VLDB*, 2007, pp. 531–542.

[48] Y. Tao, X. Xiao, J. Li, and D. Zhang, "On anti-corruption privacy preserving publication," in *ICDE*, 2008, pp. 725–734.

[49] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *PODS*, 2004, pp. 223–228.

[50] H. Park and K. Shim, "Approximate algorithms for k-anonymity," in *SIGMOD*, 2007, pp. 67–78.

[51] X. Yang and C. Li, "Secure xml publishing without information leakage in the presence of data inference," in *VLDB*, 2004, pp. 96–107.