On the Tradeoff Between Privacy and Utility in Data Publishing

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

ABSTRACT

In data publishing, anonymization techniques such as generalization and bucketization have been designed to provide privacy protection. In the meanwhile, they reduce the utility of the data. It is important to consider the tradeoff between privacy and utility. In a paper that appeared in KDD 2008, Brickell and Shmatikov proposed an evaluation methodology by comparing privacy gain with utility gain resulted from anonymizing the data, and concluded that "even modest privacy gains require almost complete destruction of the data-mining utility". This conclusion seems to undermine existing work on data anonymization. In this paper, we analyze the fundamental characteristics of privacy and utility, and show that it is inappropriate to directly compare privacy with utility. We then observe that the privacy-utility tradeoff in data publishing is similar to the risk-return tradeoff in financial investment, and propose an integrated framework for considering privacy-utility tradeoff, borrowing concepts from the Modern Portfolio Theory for financial investment. Finally, we evaluate our methodology on the Adult dataset from the UCI machine learning repository. Our results clarify several common misconceptions about data utility and provide data publishers useful guidelines on choosing the right tradeoff between privacy and utility.

Categories and Subject Descriptors

H.2.7 [**Database Administration**]: Security, integrity, and protection; H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms, Experimentation, Security, Theory

Keywords

privacy, anonymity, data publishing, data mining

1. INTRODUCTION

Privacy-preserving publishing of microdata has received much attention in recent years. Microdata contains records each of which contains information about a specific entity, such as an individual, a household, or an organization. Each record has a number of attributes: some attributes may be sensitive (such as *disease* and *salary*) and some may be quasi-identifiers (called QI, such as *zipcode*, *age*, and *sex*) whose values, when taken together, can potentially identify an individual.

Publishing microdata enables researchers and policy-makers to analyze the data and learn important information benefiting the society as a whole, such as the factors causing certain diseases, effectiveness of a medicine or treatment, and social-economic patterns that can guide the formulation of effective public policies. In other words, publishing microdata results in *utility gain for the society as a whole*. However, as microdata contains specific information about individuals, publishing microdata could also result in *privacy loss for individuals* whose information is published. Hence before the microdata can be made public, one must ensure that the privacy loss is limited to an acceptable level. This is typically done via *anonymization*, which transforms the microdata to improve the privacy. Because anonymization makes data imprecise and/or distorted, it also causes losses in potential utility gain, when compared with the case of publishing the unanonymized microdata.

A fundamental problem in privacy-preserving data publishing is how to make the right tradeoff between privacy and utility. The vast majority of existing work on privacy-preserving data publishing uses the following approach. First, one chooses a specific privacy requirement, such as k-anonymity [25, 26], ℓ-diversity [21], (α, k)-anonymity [29], t-closeness [19], and δ-presence [23], based on intuitions of what privacy means. Second, one studies the following problem: after fixing a parameter for the privacy requirement (e.g., choosing $k = 10$ in k-anonymity), how to generate an anonymized dataset that maximizes a particular utility measure, which can be the number of equivalence class [21], or the discernibility metric [4]. The above approach is limited in considering the tradeoff between utility and privacy because it is unable to answer two important questions. First, how to choose among the different privacy requirements? Second, how to choose a particular parameter for the particular requirement? For example, one would want to know whether to choose $k = 5$ or $k = 10$ for k-anonymity. In this approach, these issues are considered only from the privacy aspect, and independent of the utility aspect. However, this is inadequate as often times one does not have a clearly defined privacy requirement set in stone, and may be willing to accept a little more privacy loss to get a large gain in utility. In short, we currently lack a framework for thinking about the privacy-utility tradeoff in data publishing.

In a paper that appeared in KDD 2008, Brickell and Shmatikov [5] applied a fresh angle to the tradeoff between privacy and utility. They directly compared the privacy gain with the utility gain caused by data anonymization, and reached an intrigu-

KDD'09, June 28–July 1, 2009, Paris, France.

ing conclusion "even modest privacy gains require almost complete destruction of the data-mining utility." If this conclusion holds, then it would mean that the vast majority of the work on privacy-preserving publishing of microdata is meaningless, because one might as well publish the microdata in some trivially anonymized way. A simplified variant of the arguments made by Brickell and Shmatikov [5] is as follows. (We will present the complete arguments in Section 3.1.) Privacy loss of the published data is defined by certain kinds of information learned by the adversary from the dataset. Utility gain of the published data is defined as the same kinds of information learned by the researchers. Because both the adversary and the researchers see the same dataset and try to learn the same kinds of information, their knowledge gains are the same. Hence any utility gain by the anonymized data must be offset by the same amount of privacy loss. We call the methodology by Brickell and Shmatikov [5] the *direct comparison* methodology.

In fact, the direct-comparison methodology [5] underestimates the seriousness of privacy loss, as it uses *average* privacy loss among all individuals. When measuring privacy loss, one has to bound the *worst-case* privacy loss among *all* individuals. It is not acceptable if one individual's privacy is seriously compromised, even if the average privacy loss among all individuals is low. This is clearly illustrated when New York Times reporters identified a *single* user in the search logs published by AOL, causing AOL to remove the data immediately and fire two employees involved in publishing the data [3].

The above reasoning seems to suggest that data anonymization is even more doomed than being concluded in [5]. In this paper, we show that there are important reasons why this is not the case. Specifically, we show that arguments along the lines in [5] are flawed. It is inappropriate to directly compare privacy with utility, because of several reasons, including both technical and philosophical ones. The most important reason is that privacy is an *individual* concept, and utility is an *aggregate* concept. The anonymized dataset is safe to be published only when privacy for *each* individual is protected; on the other hand, utility gain adds up when multiple pieces of knowledge are learned. Secondly, even if the adversary and the researcher learn exactly the same information, one cannot conclude that privacy loss equals utility gain. We will elaborate this and other reasons why privacy and utility are not directly comparable in Section 3.

If privacy and utility cannot be directly compared, how should one consider them in an integrated framework for privacy-preserving data publishing? For this, we borrow the efficient frontier concept from the Modern Portfolio Theory which guides financial investments [8] (see Figure 1). When making investments, one must balance the expected return with the risk (often defined as the degree of volatility). One can choose an asset class with high risk and high expected return (e.g., stock), or choose an asset class with low risk and low expected return (e.g., cash), or choose a portfolio that combines multiple asset classes to get more attractive tradeoff between risk and return. Here the risk and expected return cannot be directly compared against each other, just as privacy and utility cannot be compared. One can use points on a two-dimensional plane (one dimension is risk, and the other is the expected return) to represent portfolios, and the efficient frontier consists of all portfolios such that there does not exist another portfolio with both lower risk and higher expected return (which would be more efficient). The points representing these efficient portfolios form the northwest frontier on all points. One can then select a portfolio either based on the maximum acceptable risk, or the slope of the curve, which offers the best risk/return tradeoff.

Contributions. This paper studies the tradeoff between privacy and utility in microdata publishing. Our contributions are as follows. First, we identify several important characteristics of privacy and utility. These observations correct several common misconceptions about privacy and utility. In particular, we show that the arguments made in the KDD 2008 paper [5] are flawed.

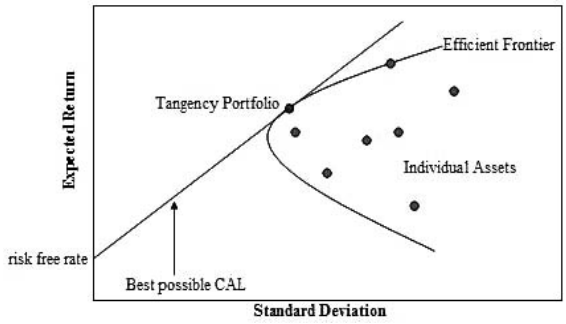


Figure 1: Efficient Frontier. (from Wikipedia)

Second, we present a systematic methodology for measuring privacy loss and utility loss. Privacy loss is quantified by the adversary's knowledge gain about the sensitive values of specific individuals, where the baseline is the trivially-anonymized data where all quasi-identifiers are removed. Utility loss is measured by the information loss about the sensitive values of large populations, where the baseline is the original data (we shall argue that, unlike privacy loss, the utility of the anonymized data should be measured against the original data rather than the trivially-sanitized data, and should be measured as "utility loss" rather than "utility gain" in Section 3.2).

Finally, we evaluate the tradeoff between privacy and utility on the adult dataset from the UCI machine learning repository. Our results show the privacy-utility tradeoff for different privacy requirements and for different anonymization methods. We also give quantitative interpretations to the tradeoff which can guide data publishers to choose the right privacy-utility tradeoff.

The rest of the paper is organized as follows. Section 2 reviews existing work and background information on microdata publishing. Section 3 describes the direct-comparison methodology due to Brickell and Shmatikov [5], clarifies the flaws of the direct-comparison methodology and presents the three characteristics of privacy and utility. Section 4 presents our methodology for measuring privacy and utility tradeoff. Section 5 experimentally evaluates our methodology and Section 6 concludes the paper with directions for future work.

2. BACKGROUND AND RELATED WORK

The general methodology for evaluating privacy-utility tradeoff fixes a privacy requirement with the privacy parameter and tries to find an algorithm that produces an anonymized dataset that maximizes a particular utility measure. The three key components in the above methodology are: (1) anonymization algorithm, (2) privacy requirement, and (3) utility measure. We elaborate on them in the rest of this section.

2.1 Generalization and Bucketization

One popular anonymization method is generalization [25, 26]. Generalization is applied on the quasi-identifiers and replaces a QI value with a "less-specific but semantically consistent value". As a result, more records will have the same set of quasi-identifier values. We define an *equivalence class* of a generalized table to be a set of records that have the same values for the quasi-identifiers.

One problem with generalization is that it cannot handle high-dimensional data due to "the curse of dimensionality" [1]. Bucketization [30, 14, 22] was proposed to remedy this drawback. The bucketization method first partitions tuples in the table into buckets and then separates the quasi-identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. The bucketized data consists of a set of buckets with permuted sensitive attribute values. Finally, another widely-used method is suppression which replaces a QI value by a '*' character.

2.2 Privacy Requirements

Several types of information disclosure in microdata publishing have been identified in the literature [6, 16]. An important type of information disclosure is *attribute disclosure*. Attribute disclosure occurs when a sensitive attribute value is associated with an individual. This is different from both *identity disclosure* (i.e., linking an individual to a record in the database) and *membership disclosure* [7, 23] (i.e., learning whether an individual is included in the database). As in [5], this paper considers *attribute disclosure*.

k-Anonymity [25, 26] (requiring each equivalence class contains at least k records) aims at preventing identity disclosure. Because identity disclosure leads to attribute disclosure (once the record is identified, its sensitive value is immediately revealed), k-anonymity can partly prevent attribute disclosure. But because attribute disclosure can occur without identity disclosure [21, 29] (for example, when all records in the equivalence class have the same sensitive value), k-anonymity does not prevent attribute disclosure.

ℓ-Diversity [21] remedies the above limitations of k-anonymity by requiring that in any equivalence class, each sensitive value can occur with a frequency of at most $1/\ell$. While there are several other definitions of ℓ-diversity such as recursive (c, ℓ)-diversity, the above probabilistic interpretation is the most widely used one in the literature. A similar privacy requirement is the (α, k)-anonymity [29].

ℓ-Diversity ensures that the probability of inferring the sensitive value is bounded by $1/\ell$. However, this confidence bound may be too strong for some sensitive values (e.g., a common form of disease) and too weak for some other sensitive values (e.g., a rare form of cancer). t-Closeness [19] remedies the limitations of ℓ-diversity, by requiring the sensitive attribute distribution in each equivalence class to be close to that in the overall data. A closely-related privacy requirement is the template-based privacy [27] where the probability of each sensitive value is bounded separately.

Similar to t-closeness, semantic privacy [5] also tries to bound the difference between the baseline belief (i.e., the distribution in the overall population) and the posterior belief (i.e., the distribution in each equivalence class). Unlike t-closeness that uses Earth Mover's Distance (EMD) (which is an *additive* measure), semantic privacy uses a *multiplicative* measure which bounds the ratio of the probability of each sensitive value in each equivalence class and that in the overall distribution. One advantage of semantic privacy is that it gives a bound on the adversary's knowledge gain: classification accuracy is bounded when semantic privacy is satisfied. Semantic privacy is quite strong and it does not capture semantic meanings of sensitive values as EMD.

2.3 Utility Measures

It is important that the anonymized data can be used for data analysis or data mining tasks. Otherwise, one can simply remove all quasi-identifiers and output the trivially-anonymized data, which provides maximum privacy.

Also, it is unclear what kinds of data mining tasks will be performed on the anonymized data. Otherwise, instead of publishing the anonymized data, one can simply perform the data mining tasks and output their results. Because of this, most utility measures are workload-independent, i.e., they do not consider any particular data mining workload. For example, the utility of the anonymized data has been measured by the number of generalization steps, the average size of the equivalence classes [21] , the discernibility metric (DM) [4] which sums up the squares of equivalence class sizes, and the KL-divergence between the reconstructed distribution and the true distribution for all possible quasi-identifier values [13].

Several researchers have proposed to evaluate the utility of the anonymized data in terms of data mining workloads, such as classification and aggregate query answering (A comprehensive discussion on the privacy-preserving data publishing is given in [9]). Classification accuracy on the anonymized data has been evaluated in [18, 28, 10, 27, 5]. The main results from these studies are: (1) anonymization algorithms can be tailored to optimize the performance of specific data mining workloads and (2) utility from classification is bounded when attributed disclosure is prevented. Aggregate query answering has also been used for evaluating data utility [30, 14, 24].

2.4 Limitations of the General Methodology

The general methodology (as described in the beginning of Section 2) had several limitations. First, parameters of different privacy requirements usually are not comparable; they may even have different domains. For example, the k parameter in k-anonymity [25, 26] can range from 1 to the total number of records, the ℓ parameter in ℓ-diversity [21] can range from 1 to the total number of sensitive values, the t parameter in t-closeness [19] can be any value in between of 0 and 1, and the δ parameter in semantic privacy [5] can be any positive float number. Therefore, it is not reasonable to compare different privacy requirements based on their parameters because different privacy parameters have different meanings. Second, the privacy parameters put an upper bound on the anonymized data. The actual privacy loss in a particular anonymized dataset may be less than the parameters indicate. Therefore, it is important to measure privacy for a *specific* anonymized dataset. Finally, existing utility measures are limited in several aspects. We will clarify these limitations in Section 3.2. In Section 4, we present our privacy measure and utility measure, and our methodology for evaluating privacy-utility tradeoff.

3. PRIVACY V.S. UTILITY

In this section, we discuss the *direct-comparison* methodology used by Brickell and Shmatikov [5]. We show that the direct-comparison methodology is flawed, and identify three important characteristics of privacy and utility, which lays the foundation for our methodology described in Section 4.

3.1 The Direct Comparison Methodology

Recently, Brickell and Shmatikov [5] applied a fresh angle to the tradeoff between privacy and utility. They directly compared the privacy loss with the utility gain caused by data anonymization. To allow such a comparison, one has to use the *same* measurement for both privacy and utility. In [5], the trivially-anonymized data, where all quasi-identifiers are removed, is used as the benchmark for comparing the anonymized dataset with. Because the trivially-anonymized data contains no identifier information and thus does not reveal sensitive information of any individual (i.e., provides maximum privacy protection in the considered framework). When a non-trivial anonymization is applied, information on quasi-identifiers is revealed, which could cause both privacy loss and utility gain, comparing to the trivially-anonymized data.

In the direct comparison methodology, this privacy loss is measured as the adversary's accuracy improvement in guessing the sensitive attribute value of an individual, and utility gain is measured as the researcher's accuracy improvement in building a classification model for the sensitive attribute. This assumes that both the adversary and the researcher have the same goal, i.e., learning information to predict the sensitive attribute value. Because whatever information that can be discovered by the researcher can also be learned by the adversary, the analysis of privacy-utility tradeoff is trivialized: privacy loss always equals utility gain.

This trivialization is resulted from the following assumptions.

1. Both the adversary and the researcher have the same prior knowledge about the data.
2. Both the adversary and the researcher use the same approach to learn information from the anonymized data.
3. Learning the same kinds of information has the same impact on privacy and utility.

If all of the three assumptions hold, privacy loss would equal utility gain. Because of the first two assumptions, the adversary and the researcher would have exactly the same posterior belief about the data. If the third assumption also holds, the adversary's knowledge gain would equal the researcher's knowledge gain, implying that privacy loss equals utility gain.

To avoid such a trivial result, at least one of the three assumptions must be changed. The direct comparison methodology in [5] changes the first assumption. It assumes that the adversary has less prior knowledge than the researcher. Specifically, it is assumed that the microdata contains some *neutral* attributes that are known to the researcher but not to the adversary; these neutral attributes are not considered as QI's. Then the benchmark trivially-anonymized dataset becomes the dataset with only the neutral attributes and the sensitive attribute, but not the QI's. For anonymized dataset, one compares with this new benchmark for privacy loss and utility gain. Experiments in [5] leads to the intriguing conclusion "even modest privacy gains require almost complete destruction of the data-mining utility". Because this approach gives the apparent impression of limiting the adversary (who does not know the neutral attributes), they further claim that "to protect against an adversary with auxiliary information, the loss of utility must be even greater".

We now show that the above conclusions do not hold. Because the researcher knows the neutral attributes, which often have correlations with the sensitive attribute, the researcher can already learn information about individuals from the new benchmark, and can predict sensitive attributes of individuals quite well. Hence the additional improvement the researcher can get from any anonymized dataset would be small. Because the adversary does not know the neutral attribute values of individuals, the adversary learns little from the new benchmark, and hence is able to gain more from any anonymized dataset. This naturally leads to the conclusion that publishing anonymized dataset is less useful for the researcher than for the adversary. In fact, one can conclude this without running any experiment. It essentially follows from the ways privacy loss and utility gain are defined. Assuming the adversary has less prior knowledge than the researcher allows the adversary to "gain more" from the anonymized data. Under the more natural assumptions that the adversary knows more information than the researcher and the benchmark includes only the sensitive attribute, the comparison between privacy loss and utility gain again becomes a trivial tie.

3.2 Characteristics of Privacy and Utility

From the analysis of the direct-comparison methodology above, one can see that it essentially says that privacy gain equals utility loss. We now argue that directly comparing privacy and utility (as in [5]) is neither reasonable nor feasible, because privacy and utility have very different characteristics, as discussed below.

3.2.1 Specific and Aggregate Knowledge

The direct-comparison methodology implicitly assumes that learning the same piece of information has the *same* impact on both privacy and utility; otherwise one cannot compare them. In fact, this assumption is used quite commonly (though often implicitly) in the literature. For example, Iyengar [12] claims that classification accuracy is maximized when the sensitive values are homogeneous within each equivalence class, which directly contradicts the ℓ-diversity requirement [21]. Similarly, privacy [21, 29, 19] is quantified by $P(SA|QI)$ (i.e., how much an adversary can learn about the sensitive value of an individual from the individual's QI values) while utility [30] is measured by attribute correlations between the QI attributes and the sensitive attribute.

In reality, the same piece of information can have very different impacts on privacy and utility. More specifically, for *different kinds* of knowledge, having the adversary and the researcher learn exactly the same knowledge can be beneficial in some cases and detrimental in other cases. For example, suppose that it is learned from the published data that people living near a small town have a much higher rate of getting cancer (say, 50%) than that among the general population. Learning this piece of information can impact both privacy and utility. On the one hand, this piece of information breaches the privacy of the people in this small town. For example, when they go to purchase health/life insurance, it can adversely affect their ability of getting insurance. On the other hand, by publishing this piece of information, people can investigate the causes of the problem (e.g., find some sources of pollution) and deal with the problem (e.g., by removing the pollution sources or taking precautions). In this case, such *aggregate* information results in more utility gain than privacy loss as it benefits the society on the whole, even for non-participants.

Suppose that, in another case, it is learned from the published data that an individual has a 50% probability of having cancer because the individual's record belongs to an equivalence class containing two records one of which has cancer. Such *specific* information has no utility value to researchers but causes privacy loss. Again, the information gain by the researcher and the adversary are the same, but the utility gain and the privacy loss are very different.

The above arguments leads to the first characteristic of privacy and utility: ***specific* knowledge (that about a small group of individuals) has a larger impact on *privacy*, while *aggregate* information (that about a large group of individuals) has a larger impact on *utility*.**

In other words, privacy loss occurs when the adversary learns more information about specific individuals from the anonymized data. But data utility increases when information about larger-size populations is learned.

Another effect of the aggregate nature of utility is more philosophical than technical. When publishing anonymized dataset, only the individuals whose data are included have potential privacy loss, while everyone in the society has potential utility gain. In fact, this principle is implicit in any kind of survey, medical study, etc. While each participant may loss more than she individually gains, the society as a whole benefit. And everyone is benefiting from the survey and study that one does not participate. Because benefit to society is difficult (if not impossible) to precisely compute, it is unreasonable to require that publishing certain anonymized dataset results in higher "utility gain" than "privacy loss" using some mathematical measure.

3.2.2 Individual and Aggregate Concepts

Another important reason why privacy and utility cannot be directly compared is as follows. For privacy protection, it is safe to publish the data only when *every* record satisfies the privacy parameter (i.e., every individual has a bounded privacy loss). This implies that privacy is an *individual* concept in that each individual's privacy is enforced *separately*. This characteristic is different from utility gain. When multiple pieces of knowledge are learned from the anonymized data, data utility adds up. This implies that utility is an *aggregate* concept in that utility *accumulates* when more useful information is learned from the data. The above arguments lead to the second characteristic of privacy and utility: ***privacy* is an *individual* concept and should be measured *separately* for every individual while *utility* is an *aggregate* concept and should be measured *accumulatively* for all useful knowledge.**

This characteristic immediately implies the following corollary on measuring privacy and utility.

COROLLARY 3.1. For privacy, the *worst-case* privacy loss should be measured. For utility, the *aggregated* utility should be measured.

Hence it is possible to publish anonymized data even if for each individual, both the privacy loss and the utility gain are small, because the utility adds up.

3.2.3 Correctness of Information

Yet another difference between privacy and utility emerges when we consider the correctness of the information learned from the anonymized data. When the adversary learns some *false* information about an individual, the individual's privacy is breached even though the perception is incorrect. However, when the researcher learns some *false* information, data utility deteriorates because it may lead to false conclusions or even misleading public policies.

In fact, some studies have overlooked this difference between privacy and utility. For example, the direct comparison methodology uses the trivially-anonymized data as the baseline for both privacy and utility. While the trivially-anonymized data is appropriate as the baseline for privacy [19, 5], it is inappropriate to be used as the baseline for utility gain. Consider using the anonymized data for aggregate query answering, e.g., determining the distribution of the sensitive values in a large population. Let the estimated distribution be $\hat{P}$. Let the distribution of the sensitive values in the trivially-anonymized data be Q. When the trivially-anonymized data is used as the baseline, the anonymized data adds to utility when $\hat{P}$ is different from Q. However, $\hat{P}$ might be significantly different from the true distribution P. The estimated false information does not contribute to utility; in fact, it worsens the data utility.

This is the third characteristic of privacy and utility: ***any* information that deviates from the prior belief, *false or correct*, may cause *privacy* loss but only *correct* information contributes to utility**. This characteristics implies the following corollary on measuring privacy and utility.

COROLLARY 3.2. Privacy should be measured against *the trivially-anonymized data* whereas utility should be measured using *the original data* as the baseline.

When the original data is used for measuring utility, we need to measure "utility loss", instead of "utility gain". An ideal (but unachievable) privacy-preserving method should result in zero privacy loss and zero utility loss.

To summarize, privacy cannot be compared with utility directly because: (1) privacy concerns information about specific individuals while aggregate information about large populations also contributes to utility, (2) privacy should be enforced for each individual and for the worst-case while utility accumulates all useful knowledge; (3) only participants have potential privacy loss, while the society as a whole benefit, and (4) false information can cause privacy damage but only correct information contributes to utility gain. All reasons suggest that the direct-comparison methodology is flawed. These characteristics also lay the foundation for our proposed methodology in Section 4.

4. METHODOLOGY

In this section, we present our methodology for analyzing the privacy-utility tradeoff in determining how to anonymize and publish datasets. Data publishers often have many choices of privacy requirements and privacy parameters. They can anonymize the data and generate a number of datasets that satisfy different privacy requirements and different privacy parameters. Often times, an important question for them is "which dataset should be chosen to publish?". Our methodology helps data publishers answer this question.

We observe that the privacy-utility tradeoff in microdata publishing is similar to the risk-return tradeoff in financial investment. In financial investment, risk of an asset class or a portfolio is typically defined as volatility of its return rate, which can be measured using, e.g., the standard deviation. Risk cannot be directly compared with return, just as privacy cannot be directly compared with utility. Similarly, different investors may have different tolerance of risks and expectation of returns. Different data publishers may have different tolerance of privacy and expectation of utility.

We borrow the efficient frontier concept from the Modern Portfolio Theory. Given two anonymized datasets $\widehat{D}_1$ and $\widehat{D}_2$, we say that $\widehat{D}_1$ is *more efficient* than $\widehat{D}_2$ if $\widehat{D}_1$ is as good as $\widehat{D}_2$ in terms of both privacy and utility, and is better in at least one of privacy and utility. Two anonymized datasets $\widehat{D}_1$ and $\widehat{D}_2$ may not be comparable because one may offer better privacy but worse utility.

Given a number of anonymized datasets, for each of them we measure its privacy loss P_{loss} relative to the case of publishing a trivial anonymized dataset that has no privacy threat, and its utility loss U_{loss} relative to the case of publishing the dataset without anonymization. We obtain a set of (P_{loss}, U_{loss}) pairs, one for each anonymized dataset. We plot the (P_{loss}, U_{loss}) pairs on a 2-dimensional space, where the x-axis depicts the privacy loss P_{loss} and the y-axis depicts the utility loss U_{loss}. An ideal (but often impossible) dataset would have $P_{loss} = 0$ and $U_{loss} = 0$, which corresponds to the origin point of the coordinate. All datasets that are most efficient will form a curve, and the data publisher can choose a dataset based on the desired levels of privacy and utility and the shape of the curve.

To use our methodology, one must choose a measure for privacy and a measure for utility. Our methodology is independent of the particular choices for such measures. In this paper, we use P_{loss} to measure the degree of attribute disclosure beyond what can be learned from publishing the sensitive attributes without QIs. We introduce a novel utility measure, which is based on the intuition of measuring the accuracy of association rule mining results.

4.1 Measuring Privacy Loss P_{loss}

We propose a worst-case privacy loss measure. Let Q be the distribution of the sensitive attribute in the overall table. As in [19, 5], we use the distribution Q as the adversary's *prior knowledge* about the data, because Q is always available to the adversary even if all quasi-identifiers are suppressed. This is true as long as the sensitive

attribute is kept intact, as in most existing methods. Privacy leaks occur only when the adversary learns sensitive information *beyond* the distribution Q.

When the adversary sees the anonymized data, the adversary's *posterior knowledge* about the sensitive attribute of a tuple t reduces to the equivalence class that contains t. Let the distribution of the sensitive attribute in the equivalence class be $P(t)$. The privacy loss for a tuple t is measured as the distance between Q and $P(t)$. We use the JS-divergence distance measure:

$$P_{loss}(t) = JS(Q, P(t)) = \frac{1}{2}[KL(Q, M) + KL(P(t), M)]$$

where $M = \frac{1}{2}(Q + P(t))$ and $KL(,)$ is the KL-divergence [15]:

$$KL(Q, P) = \sum_i q_i \log \frac{q_i}{p_i}$$

Note that here we JS-divergence rather than KL-divergence because KL-divergence is not well-defined when there are zero probabilities in the second distribution P. Therefore, using KL-divergence would require that for every equivalence class, all sensitive attribute values must occur at least once. However, most existing privacy requirements such as ℓ-diversity [21], t-closeness [19], and sematic privacy [5] do not have such a property. Finally, the worst-case privacy loss is measured as the maximum privacy loss for all tuples in the data:

$$P_{loss} = \max_t P_{loss}(t)$$

It should be noted that one has the option to use other distance measures. Whichever distance measure one chooses, it should be the case that less privacy loss occurs when the distance is smaller. Studying which distance measure one should choose is beyond the scope of this paper.

4.2 Measuring Utility Loss U_{loss}

In general, there are two approaches to measure utility. In the first approach, one measures the amount of utility that is remained in the anonymized data. This includes measures such as the average size of the equivalence classes [21] and the discernibility metric [4]. This also includes the approach of evaluating data utility in terms of data mining workloads. In the second approach, one measures the loss of utility due to data anonymization. This is measured by comparing the anonymized data with the original data. This includes measures such as the number of generalization steps and the KL-divergence between the reconstructed distribution and the true distribution for all possible quasi-identifier values [13].

It should be noted that, when the utility of the original data is low, it should be expected that the utility of the anonymized data is also low. In this case, the first approach may conclude that data anonymization has destroyed data utility while in fact, the low data utility is due to low utility of the original data. Similarly, the fact the anonymized data can be used for a variety of data mining tasks does not imply that the anonymization method is effective; another anonymization method may provide even higher data utility with less privacy loss. Due to these reasons, the first approach provides little indication with regard to whether an anonymization method is effective or not. Our methodology therefore adopts the second approach. This is also consistent with our arguments about data utility in Section 3.2: utility should be measured as *utility loss* against *the original data*.

When we measure utility loss, we need to decide which data mining task should be chosen. Previous studies have evaluated data utility in terms of classification [12, 18, 27, 5]. Because classification can also be used by the adversary to learn the sensitive values of specific individuals, when the adversary's knowledge gain is bounded, data utility of classification is also bounded (see Section 4.2 of [5]). Therefore, data utility of classification will not constitute a major part of data utility because it is bounded for a safely-anonymized dataset. Because of this, we do not measure data utility in terms of classification. Note that, we do not intend to underestimate the potential use of the anonymized data for classification purposes. In fact, we agree with the previous studies on the utility of classification.

Instead of classification, we use the anonymized data for association rule mining [31] and aggregate query answering [30, 14, 24]. For both workloads, an important task is to reconstruct the sensitive attribute distribution for large populations. This is also consistent with our arguments about data utility in Section 3.2: information on large populations contributes to utility. A large population can be specified by a support value and a predicate involving only quasi-identifiers, e.g., "$Age \geq 40 \&\& Sex = Male$". The support value is the number of records in the data that satisfy the predicate. We therefore adopt the following methodology for measuring utility loss U_{loss}.

First, we find all large populations whose support values are at least $minSup$ (where $minSup$ is a user-defined threshold value). To allow the large populations to be defined in terms of generalized predicate such as "$Age \geq 40$", we use generalized predicates that involve not only values from the attribute domain of the quasi-identifiers but also values from the generalization hierarchy of the quasi-identifiers (see for example [20] and other data mining literature on generalized association rule mining). We use the FP-tree [11] algorithm for discovering large populations.

Next, for each large population y, we compute the estimated distribution $\bar{P}_y$ of the sensitive attribute from the anonymized data and the true distribution P_y of the sensitive attribute from the original data. We adopt the uniform distribution assumption: every value in a generalized interval is equally possible and every sensitive value in an equivalence class is also equally possible. We measure the difference between P_y and $\bar{P}_y$ as the researcher's information loss when analyzing the the large population y. Again, we use the JS-divergence as the distance measure, i.e., $U_{loss}(y) = JS(P_y, \bar{P}_y)$.

Finally, because utility is an *aggregate* concept, we measure the utility loss U_{loss} by averaging utility loss $U_{loss}(y)$ for all large population y.

$$U_{loss} = \frac{1}{|Y|} \sum_{y \in Y} U_{loss}(y)$$

where Y is the set of all large populations. The anonymized data provides maximum utility when $U_{loss} = 0$. In our experiments (see Section 5), we also empirically evaluate data utility in terms of aggregate query answering.

4.3 Special Cases

There are two special cases for the privacy-utility tradeoff. The first case is to publish the trivially-anonymized data where all quasi-identifiers are completely suppressed. In this case, the estimated distribution of the sensitive attribute for every individual equals to the overall distribution Q. Because $JS[Q, Q] = 0$, we have $P_{loss}(t) = 0$ for every tuple t. Therefore, $P_{loss} = 0$, achieving maximal privacy protection.

Similarly, the estimated distribution of the sensitive attribute for every large population also equals to the overall population Q. Because the overall distribution Q may be quite different from the true distribution, utility loss could be significant. This trivially-anonymized dataset is the first baseline that ensures $P_{loss} = 0$ but U_{loss} can be large.

The second special case is to publish the original dataset where

	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 1: Description of the *Adult* dataset.

all quasi-identifiers are kept intact. In this case, any estimated information is correct and the estimated distribution equals to the true distribution, i.e., $\bar{P}_y = P_y$ for every population y. Because $JS(P_y, P_y) = 0$, we have $U_{loss}(y) = 0$ for every population y. Therefore, $U_{loss} = 0$, achieving maximum utility preservation. However, because the sensitive value for every individual is revealed, which can be quite different from the overall distribution, privacy loss P_{loss} is significant. The original dataset is the second baseline that ensures $U_{loss} = 0$ but P_{loss} can be significant.

4.4 Advantages

Our evaluation methodology has a number of advantages when compared with existing work. First, one can use this methodology to compare datasets anonymized using different requirements. E.g., both ℓ-diversity and t-closeness are motivated by protecting against attribute disclosure, by choosing one privacy loss measure, one can compare datasets anonymized with ℓ-diversity for different ℓ values and those anonymized with t-closeness for different t values.

Second, we measure utility loss against the original data rather than utility gain. Utility gain is not well-defined in data publishing. In order to measure utility gain, a baseline dataset must be defined. Because only correct information contributes to utility, the baseline dataset must contain correct information about large populations. In [5], Brickell and Shmatikov used the trivially-anonymized data as the baseline, in which every distribution is estimated to be the overall distribution and therefore causes incorrect information.

Third, we measure utility for aggregate statistics, rather than for classification. This is because, as several studies have shown, the utility of the anonymized data in classification is limited when privacy requirements are enforced.

Finally, we measure privacy loss in the worst-case and measure the accumulated utility loss. Our methodology thus evaluates the privacy loss for *every* individual and the utility loss for *all* pieces of useful knowledge.

5. EXPERIMENTS

We implemented Mondrian [17] to enforce four requirements: k-anonytmity [26], ℓ-diversity [21], t-closeness [19], and semantic privacy [5]. We used both generalization and bucketization.

We used the Adult dataset (which has been widely used in previous studies) from the UCI Machine Learning Repository [2]. The data contains 45222 records and we use seven attributes of the data, as described in Table 1.

5.1 Utility Loss U_{loss} V.S. Privacy Loss P_{loss}

For each privacy requirement, we use the Mondrian algorithm to compute the anonymized data that satisfies the privacy requirement. Then, privacy loss P_{loss} and utility loss U_{loss} are measured for the anonymized data.

We plot the privacy loss on the x-axis and utility loss on the y-axis. Experimental results are shown in Figure 2. We choose the privacy parameters (i.e., k, ℓ, t, and δ) such that all privacy requirements span a similar range of privacy loss on the x-axis. Specifically, we choose $k \in \{10, 50, 100, 200, 500, 1000, 2000, 5000\}$. For example, when $k = 5000$, the evaluated privacy loss $P_{loss} = 0.086$ and the evaluated utility loss $U_{loss} = 0.0288$, which corresponds to the leftmost point on the k-anonymity curve in Figure 2(a). We choose $\ell \in \{3.0, 3.5, 4.0, 4.25, 4.5, 4.75, 5.0, 5.5\}$, $t \in \{0.075, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$, and $\delta \in \{1.0, 1.2, 1.4, 1.5, 1.7, 1.9, 2.0, 2.1\}$. Therefore, we have 8 points on each privacy-requirement curve and they span a similar range on the x-axis, from 0 to 0.6 (see Figure 2). Note that we choose $\delta \geq 1$ because the Mondrian algorithm returns one single equivalence class when $\delta < 1$. For t-closeness, we use JS-divergence as the distance measure. For utility measure U_{loss}, we fix the minimum support value as $minSup = 0.05$.

Results. Figure 2(a) shows the utility loss v.s. privacy loss with respect to different privacy requirements. We stress that these results are affected by our choice of measures for privacy and utility. If one chooses a different measure for privacy (or utility), then the figure may look differently. As we can see from the figure, t-closeness performs better than other privacy requirements. Based on the figure, one would probably choose one of the three left-most points for t-closeness ($t = 0.075, t = 01, t = 0.15$) to publish, since they offer the best trade-off between privacy and utility. ℓ-Diversity does not perform well because it aims at bounding the posterior belief rather than the distance between the prior belief and the posterior belief. Therefore, even when ℓ-diversity is satisfied, the posterior belief can still be far away from the prior belief, thus leaking sensitive information, based on the privacy loss measure P_{loss}.

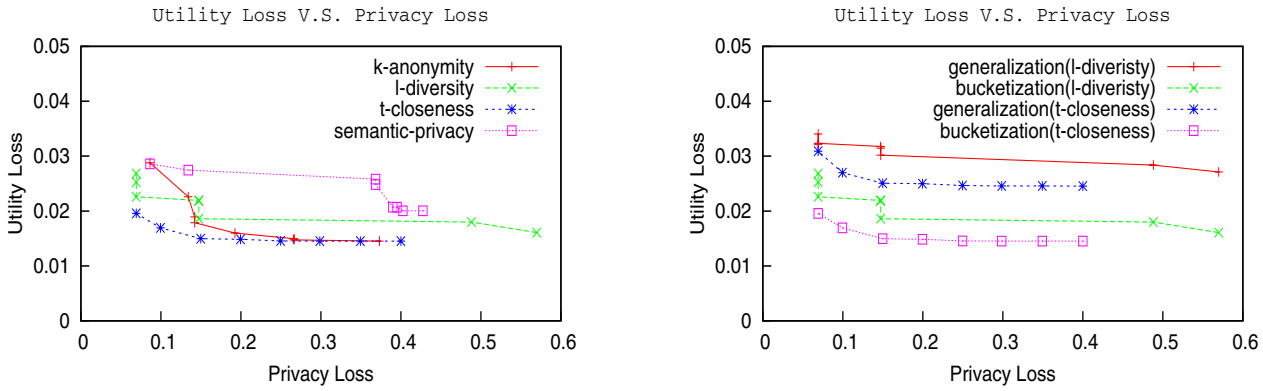
Interestingly, semantic privacy does not perform well either. Semantic privacy bounds the ratio of the posterior belief and the prior belief for every sensitive value. Semantic privacy thus provides a good privacy measure (note that δ has to be non-negative in order for semantic privacy to be achievable). However, semantic privacy is difficult to achieve in that the number of equivalence classes (or buckets) is small, especially when the sensitive attribute domain size is large. In our experiments, there are 14 sensitive values in the attribute domain of "Occupation", and requiring the ratio for each of the 14 sensitive values for each equivalence class (bucket) to be bounded is very difficult to achieve in practice.

Our results demonstrate the similarity between the privacy-utility tradeoff in data publishing and the risk-return tradeoff (Figure 1) in financial investment. One difference is that in data publishing, we measure utility loss rather than utility gain. We believe that, as in financial investment, there exists an efficient frontier in data publishing, which consists of all anonymized datasets such that there does not exist another anonymized dataset with both lower privacy loss and lower utility loss. The data publishers should only consider those "efficient" anonymized dataset when publishing data. For Figure 2(a), the efficient frontier should be somewhere below the t-closeness line.

Figure 2(b) shows the tradeoff for two anonymization methods: generalization and bucketization. We use both ℓ-diversity and t-closeness for the experiment. The results show that bucketization provides substantially better data utility than generalization, when only attribute disclosure is considered.

Interpretation of the privacy loss. We quantitatively illustrate the amount of privacy loss. Specifically, we want to answer the following question: suppose an individual's sensitive value is revealed, what is the privacy loss for that individual?

The overall distribution of the sensitive attribute "Occupation" is $Q = (0.0314, 0.1331, 0.1063, 0.1196, 0.1323, 0.1329, 0.0452, 0.0657, 0.1225, 0.0327, 0.0512, 0.0052, 0.0216, 0.0003)$. If



(a) Varied privacy requirements

(b) Varied anonymization methods

Figure 2: Privacy-Utility Tradeoff: U_{loss} V.S. P_{loss}

an individual's sensitive value is revealed, the privacy loss (computed through JS-divergence) is 0.692 when the sensitive value is "Armed-Forces" (which is the least frequent sensitive value with a frequency of 0.0003) and the privacy loss (computed through JS-divergence) is 0.488 when the sensitive value is "*Craft-repair*" (which is the most frequent sensitive value with a frequency of 0.1331). The above calculation shows that when an individual's privacy is revealed, the privacy loss is in between of 0.488 and 0.692 for the sensitive attribute "Occupation" of the Adult dataset.

This means that privacy loss cannot be greater than 0.692. However, when the privacy loss is larger than 0.488, it does not mean that at least one individual's sensitive value is revealed, because it may be the case that there is a large amount of privacy loss on the least-frequent sensitive value "Armed-Forces" even though the equivalence class (bucket) satisfies ℓ-diversity where $\ell \in \{3, 3.5\}$, as shown by the rightmost two points on the ℓ-diversity curve shown in Figure 2(a). Note that ℓ-diversity requires that even the least-frequent (i.e., the most sensitive) sensitive value must occur with a probability of at least $1/\ell$.

Interpretation of the utility loss. We also quantitatively illustrate the amount of utility loss. Specifically, we want to answer the following question: what is the utility loss when all quasi-identifiers are removed? The utility loss is calculated by averaging the utility loss for all large populations, where the estimated distribution is always the overall distribution Q. Our calculation shows that when all quasi-identifiers are removed, the utility loss is 0.05. In Figure 2, utility loss is lower than 0.04 in all cases, and is lower than 0.02 in many cases, showing that publishing the anonymized data does improve the quality of data utility than publishing trivially anonymized dataset.

5.2 Aggregate Query Answering

Our second experiment evaluates the utility of the anonymized data in terms of aggregate query answering, which has been widely used for measuring data utility [30, 14, 24].

We consider the "COUNT" operator where the query predicate involves the sensitive attribute, as in [30, 20]. A query predicate is characterized by two parameters: (1) the predicate dimension qd and (2) the query selectivity sel. The predicate dimension qd indicates the number of quasi-identifiers involved in the predicate. The query selectivity sel indicates the fraction of selected values for each quasi-identifier. For each selected parameter, we generate a set of 1000 queries for the experiments.

For each query, we run the query on the both the original table and the anonymized table. We denote the actual count from the original table as act_count and the reconstructed count from the anonymized table as rec_count. Then the average relative error is computed over all queries as:

$$\rho = \frac{1}{|Q|} \sum_{q \in Q} \frac{|rec\_count_q - act\_count_q|}{act\_conf_q} * 100$$

where Q is the set of queries generated based on the two parameters, qd and sel. In our experiments, we randomly generate 1000 aggregate queries of the above form, i.e., $|Q| = 1000$.

Results. We plot the privacy loss on the x-axis and the average relative error on the y-axis. Figure 3(a) shows the tradeoff with respect to different privacy requirements. Interestingly, the figure shows a similar pattern as that in Figure 2(a) where utility is measured as U_{loss}, instead of average relative error. The experiments confirm that our utility measure U_{loss} captures the utility of the anonymized data in aggregate query answering. One advantage of U_{loss} is to allow evaluating data utility based on the original data and the anonymized data, avoiding the experimental overheads of evaluating a large random set of aggregate queries.

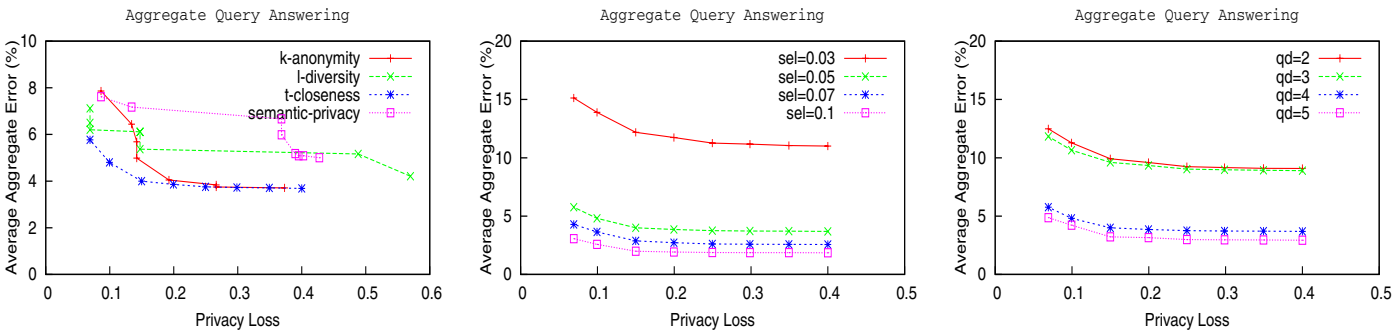
Figure 3(b) measures the tradeoff with respect to different sel values. We use t-closeness and bucketization and fix $qd = 4$. Our experiments show that the average relative error is smaller when sel is larger. Because a larger sel value corresponds to queries about larger populations, this shows that the anonymized data can be used to answer queries about larger populations more accurately.

Figure 3(c) measures the tradeoff with respect to different qd values. We again use t-closeness and bucketization and fix $sel = 0.05$. Interestingly, the results show that the anonymized data can be used to answer queries more accurately as qd increases. This is because when query selectivity is fixed, the number of points in the retrieved region is larger when qd is larger, implying a larger query region. This also shows that the anonymized data can answer queries about larger populations more accurately.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we identify three important characteristics about privacy and utility. These characteristics show that the direct-comparison methodology in [5] is flawed. Based on these characteristics, we present our methodology for evaluating privacy-utility tradeoff. Our results give data publishers useful guidelines on choosing the right tradeoff between privacy and utility.

One important question is how to use the anonymized data for data analysis (such as aggregate query answering) and data mining (such as association rule mining). We use the random distri-



(a) Varied privacy requirements (qd=4, $sel = 0.05$, bucketization)

(b) Varied query selectivity (t-closeness, $qd = 4$, bucketization)

(c) Varied query dimension (t-closeness, $sel = 0.05$, bucketization)

Figure 3: Average relative error V.S. P_{loss}

bution assumption as in most previous studies. Specifically, we interpret a generalized interval as that every value in the interval is equally possible. And when we use bucketization, given a bucket of records, we assume that every sensitive value in the bucket is equally possible. This uniform-distribution assumption has been studied for the privacy aspect [20] but there is no work on the utility aspect as far as we know. We believe that a more sophisticated approach of using the anonymized data can improve the data utility, and thus improve the quality of the privacy-utility tradeoff.

7. REFERENCES

[1] C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[2] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[3] M. Barbaro and T. Z. Jr. A face is exposed for aol searcher no. 4417749. *New York Times*, 2006.

[4] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, pages 217–228, 2005.

[5] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *KDD*, pages 70–78, 2008.

[6] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *J. Am. Stat. Assoc.*, pages 10–28, 1986.

[7] C. Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.

[8] E. Elton and M. Gruber. *Modern Portfolio Theory and Investment Analysis*. John Wiley & Sons Inc, 1995.

[9] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Survey*, 2009.

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

[11] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *SIGMOD*, pages 1–12, 2000.

[12] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.

[13] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD*, pages 217–228, 2006.

[14] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang. Aggregate query answering on anonymized tables. In *ICDE*, pages 116–125, 2007.

[15] S. L. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[16] D. Lambert. Measures of disclosure risk and harm. *J. Official Stat.*, 9:313–331, 1993.

[17] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, page 25, 2006.

[18] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *KDD*, pages 277–286, 2006.

[19] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *ICDE*, pages 106–115, 2007.

[20] T. Li and N. Li. Injector: Mining background knowledge for data anonymization. In *ICDE*, pages 446–455, 2008.

[21] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. In *ICDE*, page 24, 2006.

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

[23] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD*, pages 665–676, 2007.

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

[25] P. Samarati. Protecting respondent's privacy in microdata release. *TKDE*, 13(6):1010–1027, 2001.

[26] L. Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzz.*, 10(5):557–570, 2002.

[27] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In *ICDM*, pages 466–473, 2005.

[28] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, pages 249–256, 2004.

[29] 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*, pages 754–759, 2006.

[30] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In *VLDB*, pages 139–150, 2006.

[31] Y. Xu, B. C. M. Fung, K. Wang, A. W.-C. Fu, and J. Pei. Publishing sensitive transactions for itemset utility. In *ICDM*, pages 1109–1114, 2008.