Anonymizing Healthcare Data: A Case Study on the Blood Transfusion Service

Noman Mohammed* Benjamin C. M. Fung* Patrick C. K. Hung† Cheuk-kwong Lee‡

*CIISE, Concordia University, Montreal, QC, Canada
†University of Ontario Institute of Technology, Oshawa, ON, Canada
‡Hong Kong Red Cross Blood Transfusion Service, Hong Kong
{no_moham, fung}@ciise.concordia.ca patrick.hung@uoit.ca ckleea@ha.org.hk

ABSTRACT

Sharing healthcare data has become a vital requirement in healthcare system management; however, inappropriate sharing and usage of healthcare data could threaten patients' privacy. In this paper, we study the privacy concerns of the blood transfusion information-sharing system between the Hong Kong Red Cross Blood Transfusion Service (BTS) and public hospitals, and identify the major challenges that make traditional data anonymization methods not applicable. Furthermore, we propose a new privacy model called *LKC-privacy*, together with an anonymization algorithm, to meet the privacy and information requirements in this BTS case. Experiments on the real-life data demonstrate that our anonymization algorithm can effectively retain the essential information in anonymous data for data analysis and is scalable for anonymizing large datasets.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Security

Keywords

Privacy, anonymity, classification, healthcare

1. INTRODUCTION

Gaining access to high-quality health data is a vital requirement to informed decision making for medical practitioners and pharmaceutical researchers. Driven by mutual benefits and regulations, there is a demand for healthcare institutes to share patient data with various parties for research purposes. However, health data in its raw form often contains sensitive information about individuals, and publishing such data will violate their privacy. The current practice in data sharing primarily relies on policies and guidelines on the types of data that can be shared and agreements on the use of shared data. This approach alone may lead to excessive data distortion or insufficient protection. In this paper, we study the challenges in a real-life information-sharing scenario in the Hong Kong Red Cross Blood Transfusion Service (BTS) and propose a new privacy model, together with a data anonymization algorithm, to effectively preserve individuals' privacy and meet the information requirements specified by the BTS.

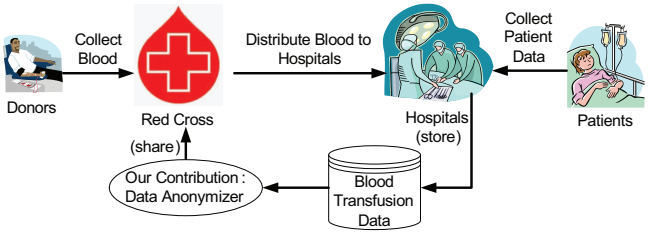


Figure 1: Data flow in Hong Kong Red Cross Blood Transfusion Service (BTS)

Figure 1 illustrates the data flow in the BTS. After collecting and examining the blood collected from donors, the BTS distributes the blood to different public hospitals. The hospitals collect and maintain the health records of their patients and transfuse the blood to the patients if necessary. The blood transfusion information, such as the patient data, type of surgery, names of medical practitioners in charge, and reason for transfusion, is clearly documented and is stored in the database owned by each individual hospital. Periodically, the public hospitals are required to submit the blood usage data, together with the patient-specific surgery data, to the BTS for the purpose of data analysis. This BTS case illustrates a typical dilemma in information sharing and privacy protection faced by many health institutes. For example, licensed hospitals in California are also required to submit specific demographic data on every discharged patient [5]. Our proposed solution, designed for the BTS case, will also benefit other health institutes that face similar challenges in information sharing. We summarize the concerns and challenges of the BTS case as follows.

Privacy concern: Giving the BTS access to blood transfusion data for data analysis is clearly legitimate. However, it raises some concerns on patients' privacy. The patients

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

are willing to submit their data to a hospital because they consider the hospital to be a trustworthy entity. Yet, the trust in the hospital may not necessarily be transitive to a third party. Many agencies and institutes consider that the released data is privacy-preserved if explicit identifying information, such as name, social security number, address, and telephone number, is removed. However, substantial research has shown that simply removing explicit identifying information is insufficient for privacy protection. Sweeney [20] showed that an individual can be re-identified by simply matching other attributes, called *quasi-identifiers* (QID), such as gender, date of birth, and postal code. Below, we illustrate the privacy threats by a simplified BTS example.

EXAMPLE 1. Consider the raw patient data in Table 1, where each record represents a surgery case with the patient-specific information. Job, Sex, and Age are quasi-identifying attributes. The hospital wants to release Table 1 to the BTS for the purpose of classification analysis on the class attribute, $Transfuse$, which has two values, Y and N, indicating whether or not the patient has received blood transfusion. Without a loss of generality, we assume that the only sensitive value in $Surgery$ is $Transgender$. The hospital expresses concern on two types of privacy threats:

Identity linkage: If a record in the table is so specific that not many patients match it, releasing the data may lead to linking the patient's record and, therefore, her received surgery. Suppose that the adversary knows that the target patient is a $Mover$ and his age is 34. Hence, record #3, together with his sensitive value ($Transgender$ in this case), can be uniquely identified since he is the only $Mover$ who is 34 years old in the raw data.

Attribute linkage: If a sensitive value occurs frequently together with some QID attributes, then the sensitive information can be inferred from such attributes even though the exact record of the patient cannot be identified. Suppose the adversary knows that the patient is a male of age 34. In such case, even though there exist two such records (#1 and #3), the adversary can infer that the patient has received a $Transgender$ surgery with 100% confidence since both the records contain $Transgender$. ■

High-dimensionality: Many privacy models, such as K-anonymity [18][20] and its extensions [14][23], have been proposed to thwart privacy threats caused by identity and attribute linkages in the context of relational databases. The usual approach is to generalize the records into equivalence groups so that each group contains at least K records with respect to some QID attributes, and the sensitive values in each QID group are diversified enough to disorient confident inferences. However, [1] has shown that when the number of QID attributes is large, that is, when the dimensionality of data is high, most of the data have to be suppressed in order to achieve K-anonymity. Our experiments confirm this *curse of high-dimensionality on K-anonymity* [1]. Applying K-anonymity on the high-dimensional patient data would significantly degrade the data quality. In order to overcome this bottleneck, we exploit one of the limitations of the adversary: in real-life privacy attacks, it is very difficult for an adversary to acquire *all* the information of a target patient because it requires non-trivial effort to gather each piece of prior knowledge from so many possible values. Thus, it is reasonable to assume that the adversary's prior knowledge is bounded by at most L values of the QID attributes of the patient. Based on this assumption, we define a new privacy model called *LKC-privacy* for anonymizing high-dimensional data.

The general intuition of LKC-privacy is to ensure that every combination of values in $QID_j \subseteq QID$ with maximum length L in the data table T is shared by at least K records, and the confidence of inferring any sensitive values in S is not greater than C, where L, K, C are thresholds and S is a set of sensitive values specified by the data holder (the hospital). LKC-privacy bounds the probability of a successful identity linkage to be $\leq 1/K$ and the probability of a successful attribute linkage to be $\leq C$, provided that the adversary's prior knowledge does not exceed L. Table 2 shows an example of an anonymous table that satisfies $(2, 2, 50\%)$-privacy by generalizing all the values from Table 1 according to the taxonomies in Figure 2. (Ignore the dashed curve for now.) Every possible value of QID_j with maximum length 2 in Table 2 (namely, QID_1, QID_2, and QID_3 in Figure 2) is shared by at least 2 records, and the confidence of inferring the sensitive value $Transgender$ is not greater than 50%. In contrast, enforcing traditional 2-anonymity will require further generalization. For example, in order to make $\langle Professional, M, [30-60)\rangle$ to satisfy traditional 2-anonymity, we may further generalize $[1-30)$ and $[30-60)$ to $[1-60)$, resulting in much higher utility loss.

Information needs: The BTS wants to perform two types of data analysis on the blood transfusion data collected from the hospitals. First, it wants to obtain some general count statistics. Second, it wants to employ the surgery information as training data for building a classification model on blood transfusion. One frequently raised question is: To avoid the privacy concern, why doesn't the hospital simply release the statistical data or a classifier to the BTS? The BTS wants to have access to the blood transfusion data, not statistics, from the hospitals for several reasons. First, the practitioners in hospitals have no expertise and interest in doing the data mining. They simply want to share the patient data with the BTS, who needs the health data for legitimate reasons. Second, having access to the data, the BTS has much better flexibility to perform the required data analysis. It is impractical to continuously request practitioners in a hospital to produce different types of statistical information and fine-tune the data mining results for research purposes.

Contributions: The contributions of this paper are summarized as follows. First, we use the BTS as a real-life example to present the challenges of privacy-aware information sharing for data analysis. Second, to thwart the privacy threats caused by identity and attribute linkage, we propose a new privacy model called *LKC-privacy* that overcomes the challenge of anonymizing high-dimensional relational data without significantly compromising the data quality (Section 3). Third, we present an efficient anonymization algorithm for achieving LKC-privacy with two different adaptations. The first adaptation maximizes the information preserved for classification analysis; the second one minimizes the distortion on the anonymous data for general data analysis. Minimizing distortion is useful when the particular information requirement is unknown during information sharing or the shared data is used for various kinds of data mining tasks (Section 4). Fourth, experiments demonstrate that our developed algorithm is flexible and scalable enough to handle large volumes of blood transfusion data that in-

clude both categorical and numerical attributes. In 2008, the BTS received 150,000 records from the public hospitals (Section 5).

2. RELATED WORK

There is a large body of work on anonymizing relational data. Traditional K-anonymity [18][20], ℓ-diversity [14], and confidence bounding [23] are based on a predefined set of QID attributes. (α, k)-anonymity [24] further requires every QID group to satisfy both K-anonymity and confidence bounding. As discussed earlier, these single QID-based approaches suffer from the curse of high dimensionality [1] and render the high-dimensional data useless for data mining. In this paper, we solve the problem of dimensionality by assuming that the adversary knows at most L values of QID attributes of *any* target patient. [6] proposes a new privacy model called *differential privacy*, which ensures that the removal or addition of a single database record does not significantly affect the overall privacy of the database. Yet, the randomization approach is not applicable to the BTS case because they require data truthfulness at the record level. [8] presents a top-down refinement (TDR) method to flexibly K-anonymize various types of attributes; however, their method does not take attribute linkage and high-dimensionality into consideration.

There are some recent works on anonymizing high dimensional transaction data [10][21][26][27]. [10] divides the transaction data into public and private items; the public items are grouped together based on similarity. Each group is then associated with a set of private items so that the probability of linking private items from public items is bounded. The idea is similar to the privacy model of *Anatomy* [25]. The methods presented in [21][26][27] model the adversary's power by a maximum number of known items as prior knowledge. This assumption is similar to ours, but our problem has major differences. First, a transaction is a *set* of items, whereas our health data is relational. Second, we have different privacy and utility measures. The privacy model of [21] is based on only K-anonymity and does not consider attribute linkages. [26] and [27] aim at minimizing data distortion and preserving frequent item sets, respectively, while we aim at preserving classification quality. Finally, [26] and [27] use suppression, while we use generalization and discretization for anonymizing various types of attributes.

Many techniques have been previously proposed to preserve privacy, but only a few have considered the goal for classification. [12] show that some simple statistical information, like means and correlations, can be preserved by adding noise and swapping values. This technique is studied in data mining for classification [3]. In these works, privacy was measured by how closely the original values of a masked attribute can be estimated, which is very different from the notion of anonymity that quantifies how uniquely an individual can be linked with sensitive information. [28] propose a privacy-preserving approach for building cox regression model. However, unlike this paper, they only target to build an analysis model and fall in the category of privacy preserving data mining (PPDM) research.

Iyengar [11] presented the anonymity problem for classification and proposed a genetic algorithmic solution. The idea is to encode each state of generalization as a "chromosome" and data distortion into the fitness function. Then

Table 1: Raw patient data

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Janitor	M	34	Y	Transgender
2	Doctor	M	58	N	Plastic
3	Mover	M	34	Y	Transgender
4	Lawyer	M	24	N	Vascular
5	Mover	M	58	N	Urology
6	Janitor	M	44	Y	Plastic
7	Doctor	M	24	N	Urology
8	Lawyer	F	58	N	Plastic
9	Doctor	F	44	N	Vascular
10	Carpenter	F	63	Y	Vascular
11	Technician	F	63	Y	Plastic

Table 2: Anonymous data ($L = 2$, $K = 2$, $C = 0.5$)

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Non-Technical	M	$[30-60)$	Y	Transgender
2	Professional	M	$[30-60)$	N	Plastic
3	Non-Technical	M	$[30-60)$	Y	Transgender
4	Professional	M	$[1-30)$	N	Vascular
5	Non-Technical	M	$[30-60)$	N	Urology
6	Non-Technical	M	$[30-60)$	Y	Plastic
7	Professional	M	$[1-30)$	N	Urology
8	Professional	F	$[30-60)$	N	Plastic
9	Professional	F	$[30-60)$	N	Vascular
10	Technical	F	$[60-99)$	Y	Vascular
11	Technical	F	$[60-99)$	Y	Plastic

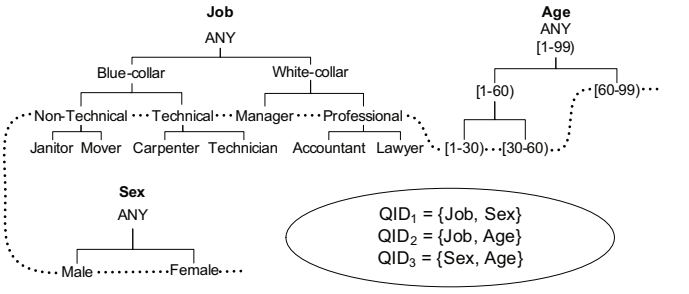


Figure 2: Taxonomy trees and QIDs

they employ the genetic evolution to converge to the fittest chromosome. Similarly, Bayardo and Agrawal [4] also addressed the classification problem using the same classification metric (CM) of [11]. Recently, LeFevre et al. [13] proposed another anonymization technique for classification using multidimensional recoding. Unlike the random genetic evolution and the bottom-up generalization, our approach produces a *progressive* generalization process that users can step through to determine a desired trade-off of privacy and accuracy. We also handle both categorical and numerical attributes. Moreover, all the proposed models for classification analysis do not address the problem of high-dimensionality, which is a primary contribution of this paper.

3. PROBLEM DEFINITION

We first describe the privacy and information requirements, followed by a problem statement.

3.1 Privacy Measure

Suppose a data holder (e.g., a hospital) wants to publish a health data table $T(ID, D_1, \ldots, D_m, Class, Sens)$ (e.g., Table 1) to some recipient (e.g., BTS) for data analysis. ID is an explicit identifier, such as SSN, and it should be removed before publication. Each D_i is either a categorical or a numerical attribute. $Sens$ is a sensitive attribute. A record has the form $\langle v_1, \ldots, v_m, cls, s\rangle$, where v_i is a domain value of D_i, cls is a class value of $Class$, and s is a sensitive value of $Sens$. The data holder wants to protect against linking an individual to a record or some sensitive value in T through some subset of attributes called a *quasi-identifier* or QID, where $QID \subseteq \{D_1, \ldots, D_m\}$.

One recipient, who is an adversary, seeks to identify the record or sensitive values of some target victim patient V in T. As explained in Section 1, we assume that the adversary knows at most L values of QID attributes of the victim patient. We use qid to denote such prior known values, where $|qid| \leq L$. Based on the prior knowledge qid, the adversary could identify a group of records, denoted by $T[qid]$, that contains qid. $|T[qid]|$ denotes the number of records in $T[qid]$. For example, $T[\langle Janitor, M\rangle] = \{ID\#1, 6\}$ and $|T[qid]| = 2$. Then, the adversary could launch two types of privacy attacks:

- *Identity linkage*: Given prior knowledge qid, $T[qid]$ is a set of candidate records that contains the victim patient V's record. If the group size of $T[qid]$, denoted by $|T[qid]|$, is small, then the adversary may identify V's record from $T[qid]$ and, therefore, V's sensitive value. For example, if $qid = \langle Mover, 34\rangle$ in Table 1, $T[qid] = \{ID\#3\}$. Thus, the adversary can easily infer that V has received a $Transgender$ surgery.
- *Attribute linkage*: Given prior knowledge qid, the adversary can identify $T[qid]$ and infer that V has sensitive value s with confidence $P(s|qid) = \frac{|T[qid \wedge s]|}{|T[qid]|}$, where $T[qid \wedge s]$ denotes the set of records containing both qid and s. $P(s|qid)$ is the percentage of the records in $T[qid]$ containing s. The privacy of V is at risk if $P(s|qid)$ is high. For example, given $qid = \langle M, 34\rangle$ in Table 1, $T[qid \wedge Transgender] = \{ID\#1, 3\}$ and $T[qid] = \{ID\#1, 3\}$, hence $P(Transgender|qid) = 2/2 = 100\%$.

To thwart the identity and attribute linkages on *any* patient in the table T, we require every qid with a maximum length L in the anonymous table to be shared by at least a certain number of records, and the ratio of sensitive value(s) in every group cannot be too high. Our privacy model, *LKC-privacy*, reflects this intuition.

DEFINITION 3.1 (LKC-PRIVACY). Let L be the maximum number of values of the prior knowledge. Let $S \subseteq Sens$ be a set of sensitive values. A data table T satisfies *LKC-privacy* if and only if for any qid with $|qid| \leq L$,

1. $|T[qid]| \geq K$, where $K > 0$ is an integer anonymity threshold, and
2. $P(s|qid) \leq C$ for any $s \in S$, where $0 < C \leq 1$ is a real number confidence threshold. Sometimes, we write C in percentage. ■

The data holder specifies the thresholds L, K, and C. The maximum length L reflects the assumption of the adversary's power. LKC-privacy guarantees that the probability of a successful identity linkage to be $\leq 1/K$ and the probability of a successful attribute linkage to be $\leq C$. LKC-privacy has several nice properties that make it suitable for anonymizing high-dimensional data. First, it only requires a subset of QID attributes to be shared by at least K records. This is a major relaxation from traditional K-anonymity, based on a very reasonable assumption that the adversary has limited power. Second, LKC-privacy generalizes several traditional privacy models. K-anonymity [18][20] is a special case of LKC-privacy with $L = |QID|$ and $C = 100\%$, where $|QID|$ is the number of QID attributes in the data table. Confidence bounding [23] is also a special case of LKC-privacy with $L = |QID|$ and $K = 1$. (α, k)-anonymity [24] is also a special case of LKC-privacy with $L = |QID|$, $K = k$, and $C = \alpha$. Thus, the data holder can still achieve the traditional models, if needed.

3.2 Utility Measure

The measure of data utility varies depending on the data analysis task to be performed on the published data. Based on the information requirements specified by the BTS, we define two utility measures. First, we aim at preserving the maximal information for classification analysis. Second, we aim at minimizing the overall data distortion when the data analysis task is unknown.

In this BTS project, we propose a top-down specialization algorithm to achieve LKC-privacy. The general idea is to anonymize a table by a sequence of specializations starting from the topmost general state in which each attribute has the topmost value of its taxonomy tree [8]. We assume that a *taxonomy tree* is specified for each categorical attribute in QID. A leaf node represents a domain value and a parent node represents a less specific value. For a numerical attribute in QID, a taxonomy tree can be grown at runtime, where each node represents an interval, and each non-leaf node has two child nodes representing some optimal binary split of the parent interval. Figure 2 shows a dynamically grown taxonomy tree for *Age*.

A *specialization*, written $v \rightarrow child(v)$, where $child(v)$ denotes the set of child values of v, replaces the parent value v with the child value that generalizes the domain value in a record. A specialization is *valid* if the specialization results in a table satisfying the anonymity requirement after the specialization. A specialization is performed only if it is valid. The specialization process can be viewed as pushing the "cut" of each taxonomy tree downwards. A *cut* of the taxonomy tree for an attribute D_i, denoted by Cut_i, contains exactly one value on each root-to-leaf path. Figure 2 shows a solution cut indicated by the dashed curve representing the anonymous Table 2. Our specialization starts from the topmost cut and pushes down the cut iteratively by specializing some value in the current cut until violating the anonymity requirement. In other words, the specialization process pushes the cut downwards until no valid specialization is possible. Each specialization tends to increase data utility and decrease privacy because records are more distinguishable by specific values. We define two utility measures depending on the information requirement to evaluate the "goodness" of a specialization.

3.2.1 Case 1: Score for Classification Analysis

For the requirement of classification analysis, we use information gain, denoted by $InfoGain(v)$, to measure the *good-*

ness of a specialization. Our selection criterion, $Score(v)$, is to favor the specialization $v \rightarrow child(v)$ that has the maximum $InfoGain(v)$:

$$Score(v) = InfoGain(v). \quad (1)$$

InfoGain(v): Let $T[x]$ denote the set of records in T generalized to the value x. Let $freq(T[x], cls)$ denote the number of records in $T[x]$ having the class cls. Note that $|T[v]| = \sum_c |T[c]|$, where $c \in child(v)$. We have

$$InfoGain(v) = E(T[v]) - \sum_c \frac{|T[c]|}{|T[v]|} E(T[c]), \quad (2)$$

where $E(T[x])$ is the *entropy* of $T[x]$ [17]:

$$E(T[x]) = -\sum_{cls} \frac{freq(T[x], cls)}{|T[x]|} \times log_2 \frac{freq(T[x], cls)}{|T[x]|}, \quad (3)$$

Intuitively, $I(T[x])$ measures the mix of classes for the records in $T[x]$, and $InfoGain(v)$ is the reduction of the mix by specializing v into $c \in child(v)$.

For a numerical attribute, the specialization of an interval refers to the optimal binary split that maximizes information gain on the $Class$ attribute. See [17] for details.

3.2.2 Case 2: Score for General Data Analysis

Sometimes, the data is shared without a specific task. In this case of general data analysis, we use discernibility cost [19] to measure the data distortion in the anonymous data table. The discernibility cost charges a penalty to each record for being indistinguishable from other records. For each record in an equivalence group qid, the penalty is $|T[qid]|$. Thus, the penalty on a group is $|T[qid]|^2$. To minimize the discernibility cost, we choose the specialization $v \rightarrow child(v)$ that maximizes the value of

$$Score(v) = \sum_{qid_v} |T[qid_v]|^2 \quad (4)$$

over all qid_v containing v. Example 3 shows the computation of $Score(v)$.

3.3 Problem Statement

Our goal is to transform a given data set T into an anonymous version T' that satisfies a given *LKC-privacy* requirement and preserves as much information as possible for the intended data analysis task. Based on the information requirements specified by the BTS, we define the problems as follows.

Definition 3.2 (Anonymization for data analysis). Given a data table T, a *LKC-privacy* requirement, and a taxonomy tree for each categorical attribute contained in QID, the *anonymization problem for classification analysis* is to generalize T on the attributes QID to satisfy the LKC-privacy requirement while preserving as much information as possible for the classification analysis. The *anonymization problem for general analysis* is to generalize T on the attributes QID to satisfy the LKC-privacy requirement while minimizing the overall discernibility cost. ■

Computing the optimal LKC-privacy solution is NP-hard. Given a QID, there are $\binom{|QID|}{L}$ combinations of decomposed QID_j with maximum size L. For any value of K and C, each combination of QID_j in LKC-privacy is an instance of the (α, k)-anonymity problem with $\alpha = C$ and $k = K$. [24] has proven that computing the optimal (α, k)-anonymous solution is NP-hard; therefore, computing optimal LKC-privacy is also NP-hard. Below, we provide a greedy approach to efficiently identify a sub-optimal solution.

Algorithm 1 Privacy-Aware Information Sharing (PAIS)

1: Initialize every value in T to the topmost value;
2: Initialize Cut_i to include the topmost value;
3: **while** some $x \in \cup Cut_i$ is valid **do**
4: Find the *Best* specialization from $\cup Cut_i$;
5: Perform *Best* on T and update $\cup Cut_i$;
6: Update $Score(x)$ and validity for $x \in \cup Cut_i$;
7: **end while**;
8: Output T and $\cup Cut_i$.;

4. ANONYMIZATION ALGORITHM

Algorithm 1 provides an overview of our algorithm *privacy-aware information sharing (PAIS)* for achieving LKC-privacy. Initially, all values in QID are generalized to the topmost value in their taxonomy trees, and Cut_i contains the topmost value for each attribute D_i. At each iteration, PAIS performs the *Best* specialization, which has the highest *Score* among the *candidates* that are valid specializations in $\cup Cut_i$ (Line 4). Then, apply *Best* to T and update $\cup Cut_i$ (Line 5). Finally, update the *Score* of the affected candidates due to the specialization (Line 6). The algorithm terminates when there are no more valid candidates in $\cup Cut_i$. In other words, the algorithm terminates if any further specialization would lead to a violation of the LKC-privacy requirement. An important property of PAIS is that the LKC-privacy is *anti-monotone* with respect to a specialization: if a generalized table violates LKC-privacy before a specialization, it remains violated after the specialization because a specialization never increases the $|T[qid]|$ and never decreases the maximum $P(s|qid)$. This anti-monotonic property guarantees that the final solution cut is a sub-optimal solution. PAIS is modified from TDR [8], which is originally designed for achieving only K-anonymity, not LKC-privacy. One major difference is the validity check in Line 6, which will be discussed in detail in Section 4.3.

Example 2. Consider Table 1 with $L = 2$, $K = 2$, $C = 50\%$, and $QID = \{Job, Sex, Age\}$. Initially, all data records are generalized to $\langle ANY\_Job, ANY\_Sex, [1\text{-}99)\rangle$, and $\cup Cut_i = \{ANY\_Job, ANY\_Sex, [1\text{-}99)\}$. To find the *Best* specialization among the candidates in $\cup Cut_i$, we compute $Score(ANY\_Job)$, $Score(ANY\_Sex)$, and $Score([1\text{-}99))$. ■

A simple yet inefficient implementation of Lines 4-6 is to scan *all* data records and recompute $Score(x)$ for all candidates in $\cup Cut_i$. The key to the efficiency of our algorithm is having *direct access* to the data records to be specialized, and updating $Score(x)$ based on some statistics maintained for candidates in $\cup Cut_i$, instead of scanning all data records. In the rest of this section, we explain our scalable implementation and data structures in detail.

4.1 Find the Best Specialization

Initially, we compute *Score* for all candidates x in $\cup Cut_i$. For each subsequent iteration, information needed to calculate *Score* comes from the update of the previous iteration

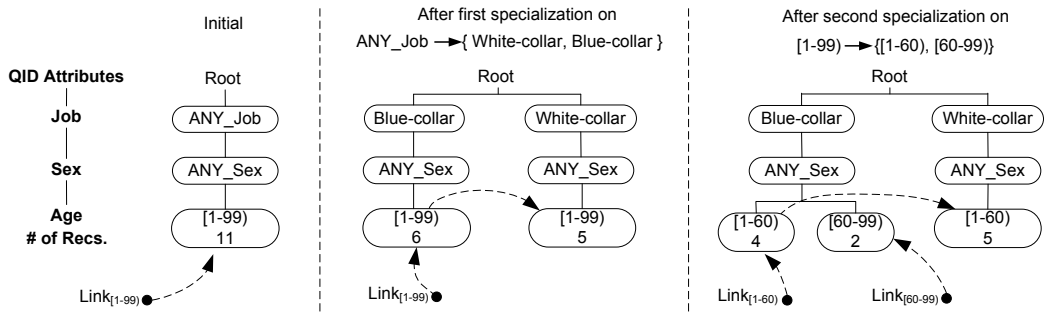


Figure 3: The TIPS data structure

(Line 7). Finding the best specialization $Best$ involves at most $|\cup Cut_i|$ computations of $Score$ without accessing data records. The procedure for updating $Score$ will be discussed in Section 4.3.

EXAMPLE 3. Continue from Example 2. We show the computation of $Score(ANY\_Job)$ for the specialization

$$ANY\_Job \rightarrow \{Blue\text{-}collar, White\text{-}collar\}.$$

For general data analysis, $Score(ANY\_Job) = 6^2 + 5^2 = 61$.
For classification analysis,
$E(T[ANY\_Job]) = -\frac{6}{11} \times log_2 \frac{6}{11} - \frac{5}{11} \times log_2 \frac{5}{11} = 0.994$
$E(T[Blue\text{-}collar]) = -\frac{1}{6} \times log_2 \frac{1}{6} - \frac{5}{6} \times log_2 \frac{5}{6} = 0.6499$
$E(T[White\text{-}collar]) = -\frac{5}{5} \times log_2 \frac{5}{5} - \frac{0}{5} \times log_2 \frac{0}{5} = 0.0$
$InfoGain(ANY\_Job) = E(T[ANY\_Job]) - (\frac{6}{11} \times$
$E(T[Blue\text{-}collar]) + \frac{5}{11} \times E(T[White\text{-}collar])) = 0.6396$
$Score(ANY\_Job) = InfoGain(ANY\_Job) = 0.6396$. ■

4.2 Perform the Best Specialization

Consider a specialization $Best \rightarrow child(Best)$, where $Best \in D_i$ and $D_i \in QID$. First, we replace $Best$ with $child(Best)$ in $\cup Cut_i$. Then, we need to retrieve $T[Best]$, the set of data records generalized to $Best$, to tell the child value in $child(Best)$ for individual data records. We employ a data structure, called *Taxonomy Indexed PartitionS (TIPS)* [8], to facilitate this operation. This data structure is also crucial for updating $Score(x)$ for candidates x. The general idea is to group data records according to their generalized records on QID.

DEFINITION 4.1 (TIPS). TIPS is a tree structure with each root-to-leaf path represents a generalized record over QID. Each leaf node stores the set of data records having the same generalized record for all the QID attributes along the path. Each path is called a *path partition*. For each x in $\cup Cut_i$, P_x denotes a path partition whose generalized record contains x, and $Link_x$ denotes the link of all P_x, with the head of $Link_x$ stored with x. ■

At any time, the generalized data is represented by the path partitions of TIPS, but the original data records remain unchanged. $Link_x$ provides a direct access to $T[x]$, the set of data records generalized to the value x. Initially, TIPS has only one path partition containing all data records, generalized to the topmost value on every attribute in QID. In each iteration, we perform the best specialization $Best$ by refining the path partitions on $Link_{Best}$.

Updating TIPS: We refine each path partition P_{Best} found on $Link_{Best}$ as follows. For each value c in $child(Best)$, a new partition P_c is created from P_{Best}, and data records in P_{Best} are split among the new partitions: P_c contains a data record in P_{Best} if c generalizes the corresponding domain value in the record. An empty P_c is removed. $Link_c$ is created to link up all P_c's for the same c. Also, link P_c to every $Link_x$ to which P_{Best} was previously linked, except for $Link_{Best}$. We emphasize that this is the only operation in the whole algorithm that requires accessing data records. The overhead of maintaining $Link_x$ is small. For each attribute in $\cup QID_j$ and each path partition on $Link_{Best}$, there are at most $|child(Best)|$ "relinkings", or at most $|\cup QID_j| \times |Link_{Best}| \times |child(Best)|$ "relinkings" in total for applying $Best$.

EXAMPLE 4. Initially, TIPS has only one path partition containing all data records and representing the generalized record ⟨*ANY_Job, ANY_Sex, [1-99)*⟩. Let the best specialization be $ANY\_Job \rightarrow \{White\text{-}collar, Blue\text{-}collar\}$ on *Job*. We create two new partitions under the root partition as in Figure 3, and split data records between them. Both the path partitions are on $Link_{ANY\_Sex}$ and $Link_{[1\text{-}99)}$. $\cup Cut_i$ is updated into {*White-collar*, *Blue-collar*, *ANY_Sex*, *[1-99)*}. Suppose that the next best specialization is *[1-99)* → {*[1-60),[60-99)*}, which specializes the two path partitions on $Link_{[1\text{-}99)}$, resulting in the TIPS in Figure 3. ■

A scalable feature of our algorithm is maintaining some statistical information for each candidate x in $\cup Cut_i$ for updating $Score(x)$ without accessing data records. For each new value c in $child(Best)$ added to $\cup Cut_i$ in the current iteration, we collect the following *count statistics* of c while scanning data records in P_{Best} for updating TIPS: $|T[c]|$, $|T[d]|$, $freq(T[c], cls)$, and $freq(T[d], cls)$, where $d \in child(c)$ and cls is a class label. These information will be used in Section 4.3.

TIPS has several useful properties. First, all data records in the same path partition have the same generalized record although they may have different raw values. Second, every data record appears in exactly one path partition. Third, each path partition P_x has exactly one generalized *qid* on QID and contributes the count $|P_x|$ towards $|T[qid]|$. Later, we use the last property to extract $|T[qid]|$ from TIPS.

4.3 Update Score and Validity

This step updates $Score(x)$ and validity for candidates x in $\cup Cut_i$ to reflect the impact of the $Best$ specialization. The key to the scalability of our algorithm is updating

$Score(x)$ using the count statistics maintained in Section 4.2 without accessing raw records again.

4.3.1 Updating Score

The procedure for updating $Score$ is different depending on the information requirement.

Case 1 classification analysis: An observation is that $InfoGain(x)$ is not affected by $Best \rightarrow child(Best)$, except that we need to compute $InfoGain(c)$ for each newly added value c in $child(Best)$. $InfoGain(c)$ can be computed from the count statistics for c collected in Section 4.2.

Case 2 general data analysis: Each path partition P_c keeps the count $|T[qid_c]|$. By following $Link_c$ from TIPS, we can compute $\sum_{qid_c} |T[qid_c]|^2$ for all the qid_c on $Link_c$.

4.3.2 Validity Check

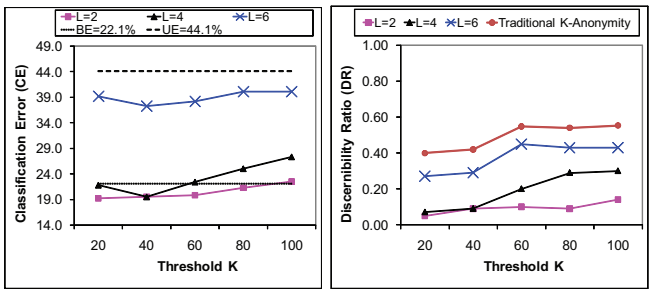
A specialization $Best \rightarrow child(Best)$ may change the validity status of other candidates $x \in \cup Cut_i$ if $Best$ and x are contained in the same qid with size not greater than L. Thus, in order to check the validity, we need to keep track of the count of every qid with $|qid| = L$. Note, we can ignore qid with size less than L because if a table satisfies LKC-privacy, then it must satisfy $L'KC$-privacy where $L' < L$.

We present an efficient method for checking the validity of a candidate. First, given a QID in T, we identify all $QID_j \subseteq QID$ with size L. Then, for each QID_j, we use a data structure, called $QIDTree_j$, to index all qid_j on QID_j. $QIDTree_j$ is a tree, where each level represents one attribute in QID_j. Each root-to-leaf path represents an existing qid_j on QID_j in the generalized data, with $|T[qid_j]|$ and $|T[qid_j \wedge s]|$ for every $s \in S$ stored at the leaf node. A candidate $x \in \cup Cut_i$ is valid if, for every $c \in child(x)$, every qid_j containing c has $|T[qid_j]| \geq K$ and $P(s|qid_j) \leq C$ for any $s \in S$. If x is invalid, remove it from $\cup Cut_i$.

5. EXPERIMENTAL EVALUATION

In this section, our objectives are to study the impact of enforcing various LKC-privacy requirements on the data quality in terms of classification error and discernibility cost, and to evaluate the efficiency and scalability of our proposed anonymization method by varying the thresholds of maximum adversary's knowledge L, minimum anonymity K, and maximum confidence C.

We employ two real-life datasets, *Blood* and *Adult*. *Blood* is a real-life blood transfusion dataset owned by an anonymous health institute. *Blood* has 62 attributes after removing explicit identifiers; 41 of them are QID attributes. *Blood Group* represents the *Class* attribute with 8 possible values. *Diagnosis Codes*, which has 15 possible values representing 15 categories of diagnosis, is considered to be the sensitive attribute. The remaining attributes are neither quasi-identifiers nor sensitive. *Blood* contains 10,000 blood transfusion records in 2008. Each record represents one incident of blood transfusion. The publicly available *Adult* dataset [16] is a *de facto* benchmark for testing anonymization algorithms [4][8][11][14][15][22][23]. *Adult* has 45,222 census records on 6 numerical attributes, 8 categorical attributes, and a binary *Class* column representing two income levels, ≤50K or >50K. See [8] for the description of attributes. We consider *Divorced* and *Separated* in the attribute *Marital-status* as sensitive, and the remaining 13 attributes as QID. All experiments were conducted on an Intel Core2 Quad Q6600 2.4GHz PC with 2GB RAM.



(a) $C = 20\%$ (b) $C = 20\%$

Figure 4: *Blood* Dataset

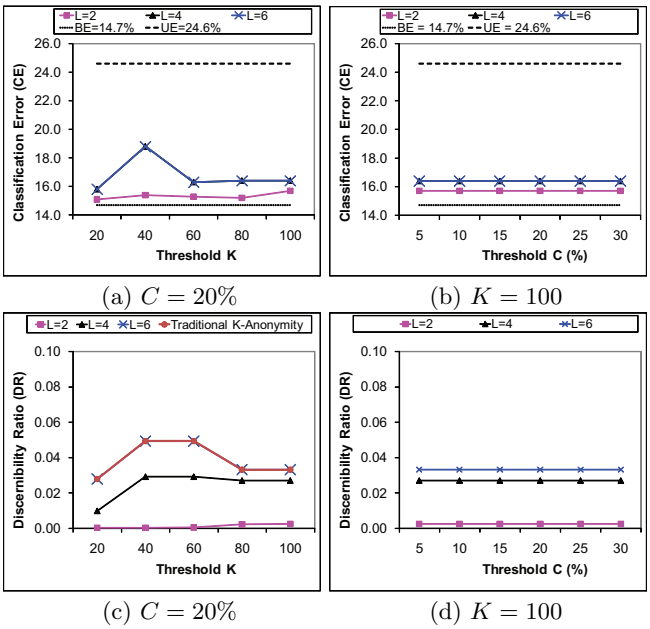
5.1 Data Utility

To evaluate the impact on classification quality (Case 1 in Section 3.2.1), we use all records for generalization, build a classifier on 2/3 of the generalized records as the training set, and measure the *classification error* (CE) on 1/3 of the generalized records as the testing set. For classification models, we use the well-known C4.5 classifier [17]. To better visualize the cost and benefit of our approach, we measure additional errors: *Baseline Error* (BE) is the error measured on the raw data without generalization. $BE - CE$ represents the cost in terms of classification quality for achieving a given LKC-privacy requirement. A naive method to avoid identity and attributes linkages is to simply remove all QID attributes. Thus, we also measure *upper bound error* (UE), which is the error on the raw data with all QID attributes removed. $UE - CE$ represents the benefit of our method over the naive approach.

To evaluate the impact on general analysis quality (Case 2 in Section 3.2.2), we use all records for generalization and measure the discernibility ratio (DR) on the final anonymous data. $DR = \frac{\sum_{qid} |T[qid]|^2}{|T|^2}$. DR is the normalized discernibility cost, with $0 \leq DR \leq 1$. Lower DR means higher data quality.

5.1.1 Blood Dataset

Figure 4a depicts the classification error CE with adversary's knowledge $L = 2, 4, 6$, anonymity threshold $20 \leq K \leq 100$, and confidence threshold $C = 20\%$ on the *Blood* dataset. This setting allows us to measure the performance of the algorithm against identity linkages for a fixed C. CE *generally* increases as K or L increases. However, the increase is not monotonic. For example, the error drops slightly when K increases from 20 to 40 for $L = 4$. This is due to the fact that generalization has removed some noise from the data, resulting in a better classification structure in a more general state. For the same reason, some test cases on $L = 2$ and $L = 4$ have $CE < BE$, implying that generalization not only achieves the given LKC-privacy requirement but sometimes may also improve the classification quality. $BE = 22.1\%$ and $UE = 44.1\%$. For $L = 2$ and $L = 4$, $CE - BE$ spans from -2.9% to 5.2% and $UE - CE$ spans from 16.8% to 24.9%, suggesting that the cost for achieving LKC-privacy is small, but the benefit is large when L is not large. However, as L increases to 6, CE quickly increases to about 40%, the cost increases to about 17%, and the benefit decreases to 5%. For a greater value of L, the difference between LKC-privacy and K-anonymity is very small in terms of classification error since more generalized data does not necessarily worse classification error. This result confirms



(a) $C = 20\%$ (b) $K = 100$

(c) $C = 20\%$ (d) $K = 100$

Figure 5: *Adult* Dataset

that the assumption of an adversary's prior knowledge has a significant impact on the classification quality. It also indirectly confirms the curse of high dimensionality [1].

Figure 4b depicts the discernibility ratio DR with adversary's knowledge $L = 2, 4, 6$, anonymity threshold $20 \leq K \leq 100$, and a fixed confidence threshold $C = 20\%$. DR *generally* increases as K increases, so it exhibits some trade-off between data privacy and data utility. As L increases, DR increases quickly because more generalization is required to ensure each equivalence group has at least K records. To illustrate the benefit of our proposed LKC-privacy model over the traditional K-anonymity model, we measure the discernibility ratio, denoted DR_{TradK}, on traditional K-anonymous solutions produced by the TDR method in [8]. $DR_{TradK} - DR$, representing the benefit of our model, spans from 0.1 to 0.45. This indicates a significant improvement on data quality by making a reasonable assumption on limiting the adversary's knowledge within L known values. Note, the solutions produced by TDR do not prevent attribute linkages although they have higher discernibility ratio.

5.1.2 *Adult* Dataset

Figure 5a depicts the classification error CE with adversary's knowledge $L = 2, 4, 6$, anonymity threshold $20 \leq K \leq 100$, and confidence threshold $C = 20\%$ on the *Adult* dataset. $BE = 14.7\%$ and $UE = 24.5\%$. For $L = 2$, $CE - BE$ is less than 1% and $UE - CE$ spans from 8.9% to 9.5%. For $L = 4$ and $L = 6$, $CE - BE$ spans from 1.1% to 4.1%, and $UE - CE$ spans from 5.8% to 8.8%. These results suggest that the cost for achieving LKC-privacy is small, while the benefit of our method over the naive method is large.

Figure 5b depicts the CE with adversary's knowledge $L = 2, 4, 6$, confidence threshold $5\% \leq C \leq 30\%$, and anonymity threshold $K = 100$. This setting allows us to measure the performance of the algorithm against attribute linkages for a fixed K. The result suggests that CE is insensitive to the change of confidence threshold C. CE slightly increases as the adversary's knowledge L increases.

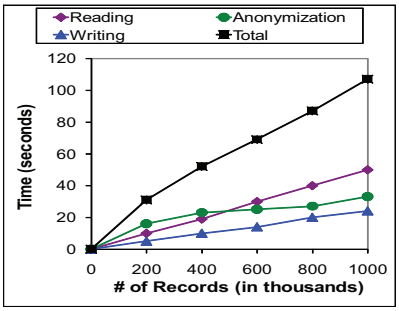


Figure 6: Scalability $(L = 4, K = 20, C = 100\%)$

Figure 5c depicts the discernibility ratio DR with adversary's knowledge $L = 2, 4, 6$, anonymity threshold $20 \leq K \leq 100$, and confidence threshold $C = 20\%$. DR sometimes has a drop when K increases. This is due to the fact that our greedy algorithm identifies only the sub-optimal solution. DR is insensitive to the increase of K and stays close to 0 for $L = 2$. As L increases to 4, DR increases significantly and finally equals traditional K-anonymity when $L = 6$ because the number of attributes in *Adult* is relatively smaller than in *Blood*. Yet, K-anonymity does not prevent attribute linkages, while our LKC-privacy provides this additional privacy guarantee.

Figure 5d depicts the DR with adversary's knowledge $L = 2, 4, 6$, confidence threshold $5\% \leq C \leq 30\%$, and anonymity threshold $K = 100$. In general, DR increases as L increases due to a more restrictive privacy requirement. Similar to Figure 5b, the DR is insensitive to the change of confidence threshold C. It implies that the primary driving forces for generalization are L and K, not C.

5.2 Efficiency and Scalability

One major contribution of our work is the development of an efficient and scalable algorithm for achieving LKC-privacy on high-dimensional healthcare data. Every previous test case can finish the entire anonymization process within 30 seconds. We further evaluate the scalability of PAIS with respect to data volume by blowing up the size of the *Adult* data set. First, we combined the training and testing sets, giving 45,222 records. For each original record r in the combined set, we created $\alpha - 1$ "variations" of r, where $\alpha > 1$ is the blowup scale. Together with all original records, the enlarged data set has $\alpha \times 45,222$ records.

Figure 6 depicts the runtime from 200,000 to 1 million records for $L = 4$, $K = 20$, $C = 100\%$. The total runtime for anonymizing 1 million records is 107s, where 50s are spent on reading raw data, 33s are spent on anonymizing, and 24s are spent on writing the anonymous data. Our algorithm is scalable due to the fact that we use the count statistics to update the *Score*, and thus it only takes one scan of data per iteration to anonymize the data. As the number of records increases, the total runtime increases linearly.

5.3 Summary

The experimental results on the two real-life datasets can be summarized as follows. (1) Our anonymization method PAIS can effectively preserve both privacy and data utility in the anonymous data for a wide range of LKC-privacy requirements. There is a trade-off between data privacy and data utility with respect to K and L, but the trend is less obvious on C. (2) Our proposed LKC-privacy model retains

more information than the traditional K-anonymity model and provides the flexibility to adjust privacy requirements according to the assumption of adversary's background knowledge. (3) PAIS is highly scalable for large data sets. These characteristics make PAIS a promising component for anonymizing healthcare data.

6. CONCLUSION AND LESSON LEARNED

We have proposed a privacy-aware information sharing method for healthcare institutes with the objective of supporting data mining. Motivated by the BTS' privacy and information requirements, we formulated the LKC-privacy model for high-dimensional relational data. Moreover, our developed algorithm can accommodate two different information requirements according to the BTS' information need. Our proposed solution is different from privacy-preserving data mining (PPDM) due to the fact that we allow *data sharing* instead of *data mining result sharing*. This is an essential requirement for the BTS since they require the flexibility to perform various data analysis tasks. We believe that our proposed solution could serve as a model for data sharing in the healthcare sector.

Finally, we would like to share our collaborative experience with the healthcare sector. Health data are complex, often a combination of relational data, transaction data, and textual data. So far, our project focuses only on the relational data, but we notice that some recent works, e.g., [9][10][21][27], are applicable to solve the privacy problem on transaction and textual data in the BTS case. Besides the technical issue, it is equally important to educate health institute management and medical practitioners about the latest privacy-preserving technology. When management encounters the problem of privacy-aware information sharing as presented in this paper, their initial response is often to set up a traditional role-based secure access model. In fact, alternative techniques, such as privacy-preserving data mining and data publishing [2][7], are available to them provided that the data mining quality does not significantly degrade.

7. ACKNOWLEDGMENTS

The research is supported in part by Discovery Grants (356065-2008) and Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada.

8. REFERENCES

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

[2] C. C. Aggarwal and P. S. Yu. *Privacy Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[3] R. Agrawal and R. Srikant. Privacy preserving data mining. In *SIGMOD*, 2000.

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

[5] D. M. Carlisle, M. L. Rodrian, and C. L. Diamond. California inpatient data reporting manual, medical information reporting for california, 5th edition. Technical report, Office of Statewide Health Planning and Development, July 2007.

[6] C. Dwork. Differential privacy: A survey of results. *Theory and Applications of Models of Computation*, 2008.

[7] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*, 2010.

[8] B. C. M. Fung, K. Wang, and P. S. Yu. Anonymizing classification data for privacy preservation. *IEEE TKDE*, 19(5):711–725, May 2007.

[9] J. Gardner and L. Xiong. An integrated framework for de-identifying heterogeneous data. *DKE*, 2009.

[10] G. Ghinita, Y. Tao, and P. Kalnis. On the anonymization of sparse high-dimensional data. In *ICDE*, 2008.

[11] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *SIGKDD*, 2002.

[12] J. Kim and W. Winkler. Masking microdata files. In *ASA Section on Survey Research Methods*, 1995.

[13] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization techniques for large-scale data sets. *ACM TODS*, 2008.

[14] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM TKDD*, 2007.

[15] N. Mohammed, B. C. M. Fung, K. Wang, and P. C. K. Hung. Privacy-preserving data mashup. In *EDBT*, 2009.

[16] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.

[17] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[18] P. Samarati. Protecting respondents' identities in microdata release. *IEEE TKDE*, 2001.

[19] A. Skowron and C. Rauszer. *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory*, chapter The discernibility matrices and functions in information systems. 1992.

[20] L. Sweeney. k-anonymity: A model for protecting privacy. In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[21] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data. In *VLDB*, 2008.

[22] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *SIGKDD*, pages 414–423, August 2006.

[23] K. Wang, B. C. M. Fung, and P. S. Yu. Handicapping attacker's confidence: An alternative to k-anonymization. *KAIS*, 11(3):345–368, April 2007.

[24] 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 *SIGKDD*, 2006.

[25] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, 2006.

[26] 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, December 2008.

[27] Y. Xu, K. Wang, A. W. C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In *SIGKDD*, 2008.

[28] S. Yu, G. Fung, R. Rosales, S. Krishnan, R. B. Rao, C. Dehing-Oberije, and P. Lambin. Privacy-preserving cox regression for survival analysis. In *SIGKDD*, 2008.