Journal of Biomedical Informatics 50 (2014) 107–121

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Quantifying the costs and benefits of privacy-preserving health data publishing

Rashid Hussain Khokhar [a], Rui Chen [b], Benjamin C.M. Fung [c,*], Siu Man Lui [d]

[a] CIISE, Concordia University, Montreal, QC, Canada
[b] Department of Computer Science, Hong Kong Baptist University, Hong Kong
[c] School of Information Studies, McGill University, Montreal, QC, Canada
[d] School of Business, James Cook University, Cairns, QLD, Australia

ARTICLE INFO

Article history:
Received 5 September 2013
Accepted 8 April 2014
Available online 24 April 2014

Keywords:
Privacy
Data utility
Health data
Cost model
Monetary value

ABSTRACT

Cost-benefit analysis is a prerequisite for making good business decisions. In the business environment, companies intend to make profit from maximizing information utility of published data while having an obligation to protect individual privacy. In this paper, we quantify the trade-off between privacy and data utility in health data publishing in terms of *monetary value*. We propose an analytical cost model that can help health information custodians (HICs) make better decisions about sharing person-specific health data with other parties. We examine relevant cost factors associated with the value of anonymized data and the possible damage cost due to potential privacy breaches. Our model guides an HIC to find the optimal value of publishing health data and could be utilized for both perturbative and non-perturbative anonymization techniques. We show that our approach can identify the optimal value for different privacy models, including *K-anonymity*, *LKC-privacy*, and ϵ-*differential privacy*, under various anonymization algorithms and privacy parameters through extensive experiments on real-life data.

1. Introduction

Electronic Health Record (EHR) systems have been widely deployed in recent years [1]. Typically, an EHR system provides stable and secure storage for large volumes of health data, including patient medical histories, laboratory test results, demographics and billing records. Centralized storage facilitates daily operations of different health service providers and provides an ideal environment for supporting effective health data mining. The goal of health data mining is to efficiently and effectively extract hidden knowledge from a large volume of health data with the goal of improving the operations of health service providers or supporting medical research. Data mining on EHRs has been proven to be beneficial to health service providers, researchers, patients, and health insurers [2].

To achieve effective health data mining, the prerequisite is to gain access to high-quality health data. Yet, health data by default is sensitive, and health information custodians (HICs) have the obligation to preserve patients' privacy [3–5] in order to minimize potential risks. The current practice of health data sharing is primarily based on obtaining consent from patients; however, HICs have faced increasing privacy breaches of different natures [6,7] due to either the negligence of administrative staff or the employment of weak de-identification methods.

In the past decade, many new privacy-enhancing techniques have been proposed to thwart different types of privacy attacks [8]. New privacy models and data anonymization methods have been iteratively proposed, broken, and patched with the discovery of new types of privacy attacks [9–11]. Thus, it is very difficult, if not impossible, to claim that the published data is bulletproof for all privacy attacks. Consequently, when an HIC shares patient-specific data with another party, he/she would like to know the answers to the following questions:

- Which privacy model and anonymization algorithm should be employed?
- Given an anonymization algorithm, how do we choose the parameters to provide adequate privacy protection to the patients?
- How useful is the data after anonymization?
- What is the probability of a privacy breach on the released data?
- What are the costs in case of a patient privacy breach?

* Corresponding author.
E-mail addresses: r_khokh@ciise.concordia.ca (R.H. Khokhar), ruichen@comp.hkbu.edu.hk (R. Chen), ben.fung@mcgill.ca (B.C.M. Fung), carrie.lui@jcu.edu.au (S.M. Lui).

http://dx.doi.org/10.1016/j.jbi.2014.04.012
1532-0464/© 2014 Elsevier Inc. All rights reserved.

A practical approach is to identify, minimize, and accept the risks by studying the trade-off between privacy protection and information utility. The recent study [12] shows that the number of health service providers reporting cases of data privacy breaches is increasing every year. The data loss includes patients' sensitive information, medical files, billing information, and insurance records. The average economic impact of data breaches over the last two years is $2.4 million. These data loss incidents have negative impacts on the public's perception of HICs and can result in potential civil lawsuits from patients' compensation claims [13,14]. Measuring the economic consequence of a privacy breach is beneficial, but also challenging. In this paper, we model the associated costs and benefits of sharing person-specific health information under different data anonymization methods at different privacy protection levels in terms of *monetary value*.

The contributions of this paper are summarized as follows. We study the challenges of sharing patient-specific health data (e.g., EHRs) faced by HICs. Different privacy models, such as *K-anonymity* [6], *LKC-privacy* [15] and ϵ*-differential privacy* [16], have been proposed to thwart potential privacy attacks on released data at the cost of degradation of data utility. We develop an analytical cost model to search for the optimal trade-off between privacy and data utility in terms of monetary value. To make our proposed model practical, we take into consideration many possible factors, such as the cost of data distortion, the likelihood of a privacy breach, the expected cost of lawsuits and compensation costs, so that HICs can measure the costs and benefits of releasing health data for secondary and commercial uses. Our model is suitable for both non-perturbative and perturbative anonymization techniques. Finally, we demonstrate the effectiveness of our proposed model by performing an extensive experimental evaluation on real-life data. Nevertheless, we would like to point out that the cost model proposed in this paper is by no means the only feasible model. In fact, there might exist many other reasonable models that may yield different monetary values for anonymized health data. This fact does not undermine our contributions as our goal is to provide a practical basis for HICs to make prudent decisions.

The rest of the paper is organized as follows. In Section 2, we review the related work. In Section 3, we present several ways of quantifying the degree of privacy protection and information utility, followed by an overview of two anonymization algorithms and a problem statement. In Section 4, we provide details of our proposed analytical cost model. In Section 5, we evaluate our proposed model by extensive experiments on real-life person-specific data. In Section 6, we discuss the criteria and the integration of cost factors in our model. Finally, we conclude the paper and discuss possible future work in Section 7.

2. Related work

The research topic of *privacy-preserving data publishing* has received enormous attention from different research communities. In this section, we review the state-of-the-arts with an emphasis on assessing the trade-off between privacy and data utility.

Yassine and Shirmohammadi [17] propose a negotiation process between online consumers and sellers in which consumers can capitalize on their personal information. Danthine and Donaldson [18] employ a risk-based premium method to determine consumers' payoff. The quantified privacy risk is context-dependent for each consumer. Similar to other business risks, the privacy risk could significantly affect the revenue of a company. Jentzsch et al. [19] analyze the monetization of privacy and find that many consumers prefer service providers with lower prices, even if they are more privacy invasive. If the products and prices are similar, then the service provider that collects less personal information gets a significant share of the market by offering privacy-friendly online services. A duopoly model is used to allow consumers to select a service provider depending on privacy concerns and the offers made by providers.

Zielinski and Olivier [20] study the optimum trade-off between privacy and data utility from the perspective of Economic Price Theory. They model the problem as an optimization problem and solve it by using the Lagrange multipliers method. They quantify information utility and privacy based on the preferences of data users and data owners on each identifying variable. Our problem is different from theirs. We propose an analytical cost model that provides a basis to aid in decision making by analyzing different cost factors associated with the value of anonymized data and the potential damage cost. We employ a top-down specialization (TDS) algorithm that uses heuristic search techniques to find the best possible trade-off between information utility and privacy. Furthermore, Zielinski and Olivier's work is limited to non-perturbative microdata anonymization and is only applicable when *global recoding* is used as the anonymization technique. In contrast, our proposed method is applicable to both perturbative and non-perturbative anonymization.

A family of previous works [21–24] discusses the trade-off between privacy and utility, but not in terms of monetary value. Loukides and Shao [21] present a distance-based quality measure that handles both quasi-identifiers (QIDs) and sensitive attributes on equal terms by optimizing the weighted sum of the amount of generalization of QIDs and the amount of protection of sensitive attributes for *K*-anonymous data. Li and Li [22] suggest that it is inappropriate to directly compare privacy with utility. They observe that the trade-off between privacy and utility in data publication is similar to the risk-return trade-off in financial investment, where the aim is to determine the appropriate level of risk. They measure privacy by JS-divergence of the distribution of the sensitive attribute and utility by utility loss against the original data.

There are also some works [25–27] studying the trade-off between disclosure risk and data utility under the confidentiality map, where R denotes disclosure risk and U denotes data utility. The R–U confidentiality map is introduced by Duncan et al. [25]. They quantify disclosure risk and data utility under three different background knowledge states and visualize them on the R–U map in order to determine the parameter values of a disclosure limitation procedure and compare different disclosure limitation procedures. Shlomo and Young [26] further explore the risk-utility trade-off of different statistical disclosure control (SDC) methods and aim to identify the optimal SDC method which reduces the disclosure risk to tolerable risk thresholds while ensuring high quality data that is fit for purpose. Loukides et al. [27] assess the disclosure risk and data utility trade-off offered by several popular transaction data anonymization algorithms using the R–U confidentiality map. Though very relevant to our paper, these works do not aim to derive a cost model that monetizes the costs and benefits of anonymizing health data.

Dwork et al. [16] discuss the *differential privacy* model, which ensures that the addition or removal of a single database record does not significantly affect any computation outcome over a database. It provides privacy protection that is independent of an adversary's background knowledge. Ghosh et al. [23] propose mechanisms that guarantee near-optimal utility to every potential user, independent of side information and preferences. They model the side information as a prior probability distribution over query results, and preferences as a loss function. Alvim et al. [24] model the database query system as an information-theoretic channel and measure the information that an attacker can learn by posting queries on a database and analyzing the response. They prove that differential privacy provides protection by imposing the bound on information leakage and utility. This bound is strong enough to prevent attacks using prior distributions. They use the binary gain function to measure the utility of a query result.

Recently, substantial research has been conducted to study privacy and utility trade-off for other types of data, such as transaction data [21,28], trajectory data [29,30], and network data [31–33].

In this paper, we consider several state-of-the-art anonymization algorithms to achieve some commonly employed privacy models, namely *K-anonymity*, *LKC-privacy*, and *ϵ-differential privacy*. We evaluate the risks, costs and benefits of releasing the anonymized data with respect to different degrees of privacy protection, and then perform a net cost-benefit analysis with the goal of finding the optimal trade-off between privacy protection and information utility.

3. Preliminaries

In this section, we present some measures to quantify the degree of privacy protection and information utility, followed by an overview of two commonly used anonymization algorithms and our problem statement.

3.1. Quantifying privacy

An HIC wants to share a person-specific data table with a health data miner (e.g., a medical practitioner or a health insurance company) for research purposes. A data table for classification analysis typically contains four types of attributes: explicit identifiers, quasi-identifiers (*QIDs*), sensitive attributes and class attributes. Explicit identifiers, such as name and social security number (SSN), are the attributes that contain explicit personally identifiable information. *QID*, such as date of birth, sex and race, is a set of attributes whose values may not be unique, but whose combination may reveal the identity of an individual. Sensitive attributes, such as disease, salary and marital status, contain an individual's sensitive information. A class attribute is the attribute that contains class values for classification analysis.

Without loss of generality, a data table for classification analysis (after removing explicit identifiers) can be defined as $D(A_1, \ldots, A_n, Sens, Class)$, where $\{A_1, \ldots, A_n\}$ are quasi-identifiers that can be either categorical or numerical, *Sens* is a sensitive attribute, and *Class* is a class attribute. A record in D has the form $\langle v_1, v_2, \ldots, v_n, s, cls\rangle$, where v_i is a value of A_i, s is a sensitive value of *Sens*, and *cls* is a class value of *Class*.

3.1.1. Privacy threats

We introduce two most common types of privacy attacks, *record linkage* and *attribute linkage* [8], in the following example.

Example 1. Consider the raw patient data in Table 1, where each record corresponds to the personal and health information of a patient, $QID = \{Age, Gender, Occupation\}$, $Sens = \{Disease\}$, and $Class = \{Blood\ transfusion\}$. An HIC wants to release Table 1 to a researcher for the purpose of classification analysis on the class attribute *Blood transfusion*, which has two values, *Y* and *N*, indicating whether or not the patient needs a blood transfusion.

In a *record linkage* attack [8], an adversary attempts to link a real-life patient to a data record in the released data table. In other words, the adversary wants to identify the record of a target victim in the table. Suppose an adversary has gathered some background knowledge about the target victim who is a female painter, denoted by $qid = \langle F, Painter\rangle$. The adversary searches for the records in the table that are consistent with the background knowledge *qid*. The group of records matching a *qid* is denoted by $D[qid]$. If the group size $|D[qid]|$ is small, the adversary may identify the victim's record and his/her sensitive value. The probability of a successful record linkage is $1/|D[qid]|$. In this example, $D[qid] = \{Rec\#4, 9, 16\}$.

Table 1
Raw patient data.

Rec#	Quasi-identifier (QID)			Sensitive	Class
	Age	Gender	Occupation	Disease	Blood transfusion
1	29	M	Doctor	Migraine	N
2	38	F	Cleaner	HIV	Y
3	64	M	Welder	Asthma	Y
4	38	F	Painter	HIV	Y
5	56	M	Painter	Migraine	N
6	24	F	Lawyer	Migraine	Y
7	36	F	Cleaner	HIV	Y
8	61	M	Lawyer	Asthma	Y
9	39	F	Painter	HIV	Y
10	24	M	Technician	Asthma	N
11	52	M	Painter	HIV	Y
12	41	F	Lawyer	Asthma	N
13	28	M	Lawyer	Migraine	Y
14	37	M	Cleaner	HIV	Y
15	66	M	Welder	Asthma	N
16	36	F	Painter	HIV	Y
17	44	M	Painter	HIV	Y

In an *attribute linkage* attack [8], an adversary may not be able to identify the exact record of a target victim, but could infer his/her sensitive value with high confidence from the released table. With the background knowledge *qid* on a target victim, an adversary can identify $D[qid]$ and infer that the victim has sensitive value s with confidence $P(s|qid) = \frac{|D[qid \wedge s]|}{|D[qid]|}$, where $D[qid \wedge s]$ denotes the set of records matching both *qid* and *s*. $P(s|qid)$ is the percentage of the records in $D[qid]$ containing s. The privacy of the target victim is at risk if $P(s|qid)$ is high. For example, given $qid = \langle M, Painter\rangle$, in Table 1, $D[qid \wedge HIV] = \{Rec\#11, 17\}$, $D[qid] = \{Rec\#5, 11, 17\}$, and $P(HIV|qid) = 2/3 = 66.67\%$.

3.1.2. Privacy models

Various privacy models have been proposed to protect against the aforementioned linkage attacks. In this subsection, we discuss the most widely adopted models in the literature, namely *K-anonymity*, *ℓ-diversity*, *LKC-privacy*, and *ϵ-differential privacy*.

Definition 3.1 (*K-anonymity [6]*). Let $D(A_1, \ldots, A_n)$ be a table and *QID* be its quasi-identifier. *D* satisfies *K*-anonymity if and only if each *QID* group in *D* appears in at least *K* records.

K-anonymity does not provide adequate privacy protection if the sensitive values in an equivalence class (i.e., the group of records matching a QID value) lack diversity, that is, it is subject to attribute linkage attacks.

Machanavajjhala et al. [34,35] proposed a privacy model called *ℓ-diversity* to thwart attribute linkage attacks. The principle of ℓ-diversity is to require every *QID* group contains at least ℓ "well-represented" sensitive values. There are several instantiations of how to ensure the diversity within each *QID* group. A relatively simple instantiation is to ensure that every *QID* group contains at least ℓ distinct sensitive value [34–36]. An alternative diversity privacy model is *entropy ℓ-diversity*.

Definition 3.2 (*Entropy ℓ-diversity [35]*). A table is entropy ℓ-diverse if every *QID* group satisfies $-\sum_{s \in S} P(qid, s) \log(P(qid, s)) \geqslant \log(\ell)$, where S is a sensitive attribute and $P(QID, s)$ is the percentage of records in *QID* group containing the sensitive value s.

LeFevre et al. [37,38] proposed a suite of workload-ware anonymization algorithms to identify a minimally anonymous table satisfying k-anonymity and/or entropy ℓ-diversity with the consideration of releasing the table for classification analysis and answering queries. Due to the curse of high dimensionality [39], enforcing K-anonymity on high-dimensional data would result in significant information loss. To overcome this bottleneck, Mohammed et al. [15] pointed out that, in a real-life privacy attack, it is very difficult for an adversary to acquire the values of all QID attributes of a target victim, leading to the *LKC-privacy* model. In this model, an adversary's background knowledge is bounded by *at most L QID* attributes.

Definition 3.3 (*LKC-privacy [15]*). Let L be the maximum number of QID attributes possessed by an adversary on a target victim and $S \subseteq Sens$ be a set of sensitive values. A data table D satisfies *LKC-privacy* if and only if, for any qid with $0 < |qid| \leqslant L$,

1. $|D[qid]| \geqslant K$, where $K > 0$ is an integer anonymity threshold, *and*
2. for any $s \in S, P(s|qid) \leqslant C$, where $0 < C \leqslant 1$ is a real number confidence threshold.

Intuitively, *LKC-privacy* prevents both record and attribute linkage attacks by ensuring that every qid value with maximum length L in D is shared by at least K records and that 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 HIC. *LKC*-privacy bounds the probability of a successful record linkage to be $\leqslant 1/K$ and the probability of a successful attribute linkage to be $\leqslant C$, provided that the adversary's background knowledge qid does not exceed L attributes. In general, *LKC*-privacy is more flexible than K-anonymity in adjusting the trade-off between privacy and utility.

Dwork et al. [16] propose *differential privacy* that provides strong privacy guarantees independent of an adversary's background knowledge and computational power.

Definition 3.4 (*ϵ-differential privacy [16]*). A sanitization mechanism M_r provides *ϵ-differential privacy*, if for any two databases D_1 and D_2 that differ on at most one record, and for any possible sanitized database $D^* \in Range(M_r)$,

$$\Pr[M_r(D_1) = D^*] \leqslant e^{\epsilon} \times \Pr[M_r(D_2) = D^*],$$

where the probabilities are taken over the randomness of M_r.

Differential privacy originates in the field of statistical disclosure control. It ensures that the addition or removal of a single database record does not significantly affect the outcome of any computation over a database. It follows that almost no risk will be incurred by joining a statistical database. ϵ is a user-specified privacy parameter. A smaller ϵ implies more stringent privacy protection. Recently, Mohammed et al. [40] present a sanitization method to achieve differential privacy on heterogenous data that is a combination of relational data and transaction data, in which the relational data contains raw patient data and the transaction data contains the diagnostic codes.

3.2. Quantifying utility

When an HIC shares health data with a data recipient, he/she may not know the intended use of the published data. Indeed, there are many data analysis/mining tasks that can be performed on the released data. Precise calculation of utility depends on the requirement of the data recipient for which the HIC needs to customize the anonymization process. For this reason, we consider two utility measures in this paper.

The first measure, *discernibility ratio* (DR), aims at quantifying the impact of anonymization on the overall data distortion for general data analysis tasks. Discernibility ratio is suitable for the cases where the analysis task is not known at the time of data publication. Formally, it is defined as:

$$DR = \frac{\sum_{qid} |D[qid]|^2}{|D|^2} \qquad (1)$$

DR is the normalized discernibility cost with the range of $0 < DR \leqslant 1$. A lower value of DR represents higher data quality.

The second measure, *classification accuracy* (CA), aims at quantifying utility for classification analysis, a specific data analysis task. To determine the impact of anonymization on data utility for classification analysis, we can build a classifier on part of the anonymized records as the training set and measure the *classification error* (CE) on the rest anonymized records as the testing set. CA is calculated by $1 - CE$. In this paper, we use the well-known $C4.5$ classifier [41] for classification analysis. *Baseline accuracy* (BA) is the accuracy measured on the raw data without anonymization. $BA - CA$ represents the cost of anonymization in terms of classification accuracy.

Many other utility metrics have been introduced in the literature. Xu et al. [42] present the concept of certainty penalty for utility measure. For a data table D consisting of both numerical and categorical attributes, the total weighted *normalized certainty penalty* (NCP) is the sum of the weighted normalized certainty penalty of all records:

$$NCP(D) = \sum_{r \in D} \sum_{i=1}^{n} (w_i \cdot NCP_{A_i}(r)) \qquad (2)$$

where r denotes a record in the data table D, w_i is the weight associated with an attribute $A_i, NCP_{A_i}(r)$ is defined as $\frac{z_i - y_i}{|A_i|}$ for numerical attributes, and $NCP_{A_i}(r)$ is defined as $\frac{size(u)}{|A_i|}$ for categorical attributes. $z_i - y_i$ is the difference between maximum and minimum values of an equivalence class and $size(u)$ is the number of leaf nodes that are descendants of u in the hierarchy.

Aggregate query answering approach has been extensively used in previous works [22,28,43] to quantify data utility or information loss. The accuracy of a counting query Q is measured by the *relative error* R_{er}:

$$R_{er}(Q) = \frac{|Q_{act} - Q_{est}|}{Q_{act}} \qquad (3)$$

where Q_{act} denotes the accurate answer for query Q when applied to the original data D, Q_{est} denotes the estimated answer when applied to the anonymized data D'. To better estimate the utility of an anonymized dataset, the one can compute *average relative error* (ARE) on a set of queries. Our proposed model can adopt all these models as utility metrics. To demonstrate a concrete instantiation of our model, we focus on discernibility ratio DR and classification accuracy (CA) in the rest of the paper.

3.3. Data anonymization algorithm

Data anonymization algorithms have to be carefully designed to balance the trade-off between privacy and data utility. In this subsection, we discuss two state-of-the-art anonymization algorithms, *Top-Down Specialization* (*TDS*) [15,44] and *Diff*erentially private anonymization based on *Gen*eralization (*DiffGen*) [9], under different privacy models.

Algorithm 1. Top-Down Specialization (TDS) Algorithm [44]

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

3.3.1. Top-down specialization algorithm

Algorithm 1 presents an overview of the *Top-Down Specialization* (*TDS*) algorithm [44] for achieving *K*-anonymity. Mohammed et al. [15] present a variant of TDS that achieves *LKC*-privacy on high-dimensional data.

Initially, all values in *QID* are generalized to the topmost value in their taxonomy tree, as illustrated in Fig. 1. A taxonomy tree is specified for each categorical attribute in *QID*. For each continuous attribute in *QID*, a taxonomy tree can be dynamically grown at runtime, forming a binary tree structure in which each non-leaf node has exactly two child nodes that represent a split of the parent interval. $Mark_i$ initially contains the topmost value for each attribute A_i in a taxonomy tree. At each iteration, the TDS algorithm performs the *Best* specialization, which has the highest *Score* among the *candidates* that are valid specializations in $\cup Mark_i$ (Line 4). Then, *Best* is applied to *D* and $\cup Mark_i$ is updated (Line 5). Finally, it updates the *Score* and validity of the candidates in $\cup Mark_i$ (Line 6).

The algorithm is efficient in updating *Score* and maintaining the statistics for candidates in $\cup Mark_i$ by directly accessing the data records. It terminates if any further specialization would lead to a violation of the privacy requirement. The specialization process can be viewed as pushing the "mark" of each taxonomy tree downwards, which increases utility and decreases anonymity as the values of the records become more distinguishable. Fig. 1 illustrate a solution mark indicated by the dotted lines which leads to the anonymous Table 2.

We discuss the details of the score function for general and classification analysis below.

Score for general analysis. For the cases where data is released for general data analysis or the data analysis task is unknown at the time of data publication, we use a discernibility metric [45] as the score function. The discernibility metric charges a cost to each record for being identical to other records. For each record in an equivalence group $D[qid]$, the penalty cost is $|D[qid]|$. Thus, the penalty cost on the group is $|D[qid]|^2$. To minimize the discernibility penalty cost, we choose the specialization $v \rightarrow child(v)$ that maximizes the value of all *qid* containing v, denoted by qid_v.

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

Rec#	Quasi-identifier (QID)			Sensitive	Class
	Age	Gender	Occupation	Disease	Blood transfusion
1	[1–99]	M	Professional	Migraine	N
2	[1–99]	F	Non-Technical	HIV	Y
3	[1–99]	M	Technical	Asthma	Y
4	[1–99]	F	Non-Technical	HIV	Y
5	[1–99]	M	Non-Technical	Migraine	N
6	[1–99]	F	Professional	Migraine	Y
7	[1–99]	F	Non-Technical	HIV	Y
8	[1–99]	M	Professional	Asthma	Y
9	[1–99]	F	Non-Technical	HIV	Y
10	[1–99]	M	Technical	Asthma	N
11	[1–99]	M	Non-Technical	HIV	Y
12	[1–99]	F	Professional	Asthma	N
13	[1–99]	M	Professional	Migraine	Y
14	[1–99]	M	Non-Technical	HIV	Y
15	[1–99]	M	Technical	Asthma	N
16	[1–99]	F	Non-Technical	HIV	Y
17	[1–99]	M	Non-Technical	HIV	Y

$$Score(v) = DM(v) = \sum_{qid_v} |D[qid_v]|^2 \tag{4}$$

Score for classification analysis. In the case of classification analysis, we use information gain, denoted by $InfoGain(v)$, to measure the *goodness* of a specialization. Our selection criterion, $Score(v)$, is to keep the specialization $v \rightarrow child(v)$ that has the maximum $InfoGain(v)$:

$$Score(v) = InfoGain(v). \tag{5}$$

InfoGain(v): Let D_x denote the set of records in *D* generalized to the value *x*. Let $freq(D_x, cls)$ denote the number of records in D_x having the class value *cls*. Note that $|D_v| = \sum_c |D_c|$, where $c \in child(v)$. So, we have

$$InfoGain(v) = H(D_v) - \sum_c \frac{|D_c|}{|D_v|} H(D_c), \tag{6}$$

$$H(D_x) = -\sum_{cls} \frac{freq(D_x, cls)}{|D_x|} \times \log_2 \frac{freq(D_x, cls)}{|D_x|}, \tag{7}$$

where $H(D_x)$ measures the *entropy* of classes for the records in D_x [41], and $InfoGain(v)$ measures the reduction of the entropy by specializing v into $c \in child(v)$. A smaller entropy $H(D_x)$ implies higher purity of the partition with respect to the class values. Example 2 shows the computation of $InfoGain(v)$.

Example 2. Consider Table 1 with $L = 2, K = 2, C = 50\%$, and $QID = \{Age, Gender, Occupation\}$. Initially, all data records are generalized to $\langle [1\!-\!99], ANY\_Gender, ANY\_Occupation \rangle$, and $\cup Mark_i = \{[1\!-\!99], ANY\_Gen\ der, ANY\_Occupation\}$. To find the *Best* specialization among the candidates in $\cup Mark_i$, we compute

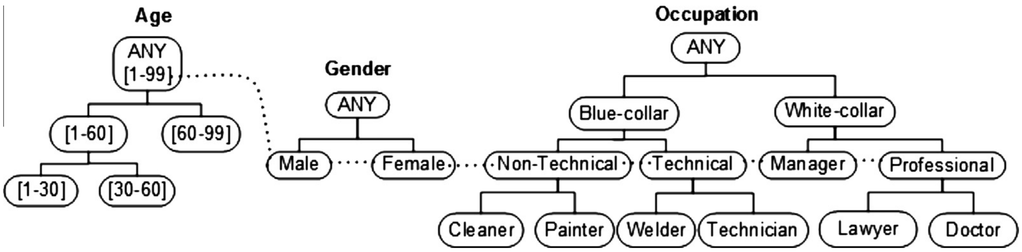

Fig. 1. Taxonomy trees.

$Score([1-99]), Score(ANY\_Gender)$, and $Score(ANY\_ \; Occupation)$. Below we show the computation of $Score(ANY\_ \; Occupation)$ for the specialization

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

For general analysis:

$$Score(ANY\_Occupation) = 12^2 + 5^2 = 169$$

$$DR = \frac{3^2 + 2^2 + 5^2 + 3^2 + 4^2}{17^2} = 0.217993$$

For classification analysis:

$$H(D_{ANY\_Occupation}) = -\frac{12}{17} \times \log_2 \frac{12}{17} - \frac{5}{17} \times \log_2 \frac{5}{17} = 0.8739$$

$$H(D_{Blue-collar}) = -\frac{9}{12} \times \log_2 \frac{9}{12} - \frac{3}{12} \times \log_2 \frac{3}{12} = 0.8112$$

$$H(D_{White-collar}) = -\frac{3}{5} \times \log_2 \frac{3}{5} - \frac{2}{5} \times \log_2 \frac{2}{5} = 0.9709$$

$$InfoGain(ANY\_Occupation) = H(D_{ANY\_Occupation}) - \left(\frac{12}{17} \times H(D_{Blue-collar}) + \frac{5}{17} \times H(D_{White-collar})\right) = 0.0156$$

$$Score(ANY\_Occupation) = InfoGain(ANY\_Occupation) = 0.0156$$

Algorithm 2. DiffGen Algorithm [9]

```
 1: Initialize every value in $D$ to the topmost value;
 2: Initialize $Mark_i$ to include the topmost value;
 3: $\epsilon' \leftarrow \frac{\epsilon}{2(|A_n^{pr}|+2h)}$;
 4: Determine the split value for each $v_n \in \cup Mark_i$ with
    probability $\propto \mathtt{exp}(\frac{\epsilon'}{2\Delta \mathtt{u}} \mathtt{u}(\mathtt{D}, v_{\mathtt{n}}))$;
 5: Compute the score for $\forall v \in \cup Mark_i$;
 6: for $i$ = 1 to $h$ do
 7:     Select $v \in \cup Mark_i$ with probability $\propto \mathtt{exp}(\frac{\epsilon'}{2\Delta \mathtt{u}} \mathtt{u}(\mathtt{D}, v))$;
 8:     Specialize $v$ on $D$ and update $\cup Mark_i$;
 9:     Determine the split value for each new $v_n \in \cup Mark_i$
        with probability $\propto \mathtt{exp}(\frac{\epsilon'}{2\Delta \mathtt{u}} \mathtt{u}(\mathtt{D}, v_{\mathtt{n}}))$;
10:     Update score for $v \in \cup Mark_i$;
11: end for
12: return each group with count $(C + \mathtt{Lap}(2/\epsilon))$
```

3.3.2. Differentially private anonymization algorithm

We present the details of *Diff*erentially private anonymization based on *Gen*eralization (*DiffGen*) [9] in Algorithm 2. *DiffGen* achieves ϵ-differential privacy by making two major extensions on TDS. First, *DiffGen* selects the *Best* specialization based on the exponential mechanism. Second, *DiffGen* adds Laplacian noise to the *qid* counts in the generalized contingency table. The Laplacian noise is calibrated based on the *sensitivity* of a function, which quantifies the maximal impact of a single user on the function.

Initially, all values in the predictor attributes $\mathcal{A}^{pr}$ (i.e., attributes used to predict the class attribute) are generalized to the topmost value in their taxonomy trees (Line 1), and $Mark_i$ contains the topmost value for each attribute A_i^{pr} (Line 2). The algorithm first determines split points for all numerical candidates based on the exponential mechanism (Line 4), and then computes the scores for all candidates $v \in \cup Mark_i$ (Line 5). Different heuristics (e.g., information gain) can be used to calculate the scores. Based on the scores, the algorithm probabilistically selects a candidate $v \in \cup Mark_i$ to specialize (Lines 7–8). Similar specialization steps are iteratively conducted until the given number of iterations has been reached. Finally, the algorithm outputs the noisy count of each group (Line 12) by using the Laplace mechanism. The privacy parameter is carefully distributed to each operation (Line 3) so that the algorithm satisfies ϵ-differential privacy. The general operation of Algorithm 2 is similar to that of Algorithm 1 except that all decisions have to be probabilistically made in order to satisfy differential privacy. A concrete example is available in [9].

3.4. Problem statement

This paper aims at answering the questions raised in Section 1 by proposing an analytical cost model. Let D be a raw patient-specific data table. An HIC would like to anonymize D and share the anonymized version D' with a third party. The HIC wants to quantify the costs and benefits of publishing D' in terms of the level of privacy protection and information utility for future data analysis tasks. Our goal is to propose an analytical cost model that quantifies individual privacy and data utility in terms of *monetary value*. This model provides guidance for HICs in finding the optimal value based on the choices of privacy models, privacy protection levels and anonymization algorithms. Formally, our research problem is defined as follows.

Definition 3.5 (*Problem Definition*). Given an input raw patient-specific data table D, a set of privacy models along with different privacy parameters and a set of anonymization algorithms, the research problem is to develop an analytical cost model that outputs an anonymized table D' that achieves the optimal monetary value.

We note that the optimal value may change with the variations of different qualitative and quantitative factors that influence the outcome of a decision. To make our model practical in different scenarios, we identify a large number of relevant factors that may contribute to the decision-making process.

The problem we consider in this paper is different from traditional optimization problems. In our problem, we are concerned with a small number of variables (e.g., privacy models, privacy parameters and anonymization algorithms) with a few possible values. This implies that it is feasible for an HIC to exhaustively search for the optimal value. For this reason, in this paper, we are *not* concerned with the approximation and computational complexity issues, which are normally important to an optimization problem.

4. Proposed solution

In this section, we present a solution to quantify the trade-off between privacy and utility in data publication in terms of monetary value. Our analytical cost model is applicable to both perturbative and non-perturbative anonymization techniques. In the subsequent analysis, we focus on analyzing person-specific relational data, but our model is also applicable to other types of data, such as set-valued data and sequential data. Our proposed model will be evaluated under several common privacy models, namely *K-anonymity*, *LKC-privacy*, and ϵ*-differential privacy*. Section 4.1 presents the analytical cost model, Section 4.2 discusses the relevant factors of determining the value of anonymized data and the factors that contribute to the potential damage cost, and Section 4.3 introduces the attack model.

4.1. Analytical cost model

Our proposed analytical cost model is the first model that quantifies costs and benefits of releasing anonymized data in terms of *monetary value*. Fig. 2 gives the overview of the proposed cost model, where nodes represent different types of factors, and

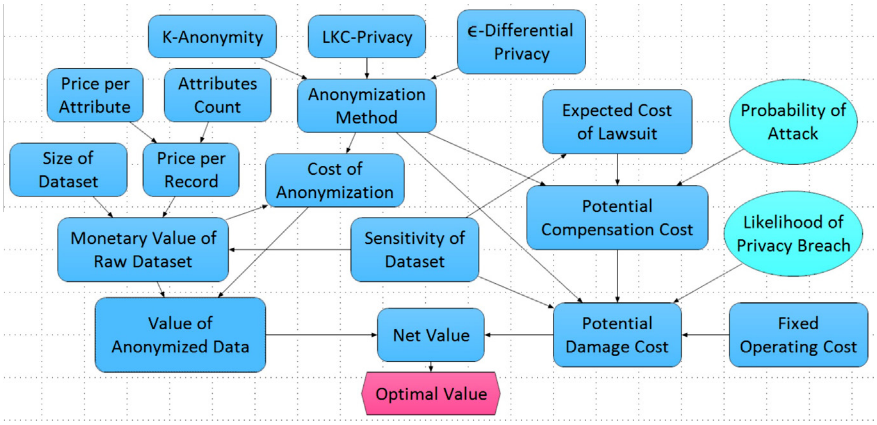

Fig. 2. Our analytical cost model.

arrows indicate the dependencies between different factors. For example, the arrow pointing from *Size of Dataset* to *Monetary Value of Raw Dataset* indicates the dependency of *Monetary Value of Raw Dataset* on *Size of Dataset*. Our model allows a user to choose privacy models along with anonymization algorithms and privacy parameters and then analyze the impact of privacy protection on information utility for health data mining in terms of monetary value. It helps identify the economic consequences of sharing patients' health data.

Broadly, *Value of Anonymized Data* depends upon *Monetary Value of Raw Dataset* and *Cost of Anonymization*. On one hand, data anonymization may impact *Value of Anonymized Data* by hiding potentially relevant information that could be used for data analysis; on the other hand, it may provide benefits from reducing the risk of privacy breaches and therefore costs of potential compensation. The factor *Cost of Anonymization* in the model represents either *Cost of Distortion* for general data analysis or *Cost in terms of Classification Quality* for classification analysis. *Optimal Value* is the model's objective and evaluates the overall value or desirability of possible outcomes. This model can help HICs make better decisions by quantifying the value of their earnings, the impact of a privacy breach, and possible costs of compensation when person-specific health data is shared for secondary and commercial purposes.

We stress that our proposed model is by no means the only reasonable one. There can be other reasonable models for different data sharing scenarios. In fact, we deem that there may not exist a silver bullet for all data sharing scenarios. To make our model applicable to different data sharing scenarios, we take into consideration many possible factors along with their mathematical relationships. We note that, in a particular case, not all these factors are necessary, and an HIC is free to add, delete or replace the factors as needed. We point out that there might be other reasonable factors, nevertheless our cost model is still of significance because it provides a basis for HICs to start with. We expect our model to be of practical use. Moreover, one salient feature of our model is that it guides an HIC to identify the best trade-off between privacy and utility in terms of monetary value.

4.2. Cost factors

To build the analytical cost model in Fig. 2, we need to identify and study the relevant quantitative and qualitative cost factors. We learn the factors from different sources [46] and integrate them into our analytical cost model. In general, the factors fall into two categories: the factors determining the monetary value of anonymized data and the factors resulting in the potential damage cost. It is natural to observe that the net benefit of publishing health data is the difference between these two categories of factors, or, more specifically, the difference between the value of anonymized data and the potential damage cost.

4.2.1. Sensitivity of dataset

The sensitivity of a dataset *SD* is a given qualitative factor, and its level *l* represents the importance of data privacy. Intuitively, a higher sensitivity level of *l* implies a higher monetary value of a raw dataset, and also a higher impact on the potential damage cost. Data privacy risks increase with the increase of data sensitivity.

4.2.2. Size of dataset

The size of a dataset $Size_{ds}$ is a quantitative factor representing the total number of records in the dataset. $Size_{ds}$ increases as the number of records in the dataset increases. Each record has a monetary value. As the number of records increases, the value of a raw dataset also increases.

4.2.3. Price per attribute

The price per attribute Pr_{attr} is a quantitative factor and represents the cost of collecting one successful questionnaire for an attribute.

4.2.4. Attribute count

The attribute count $Count_{attr}$ represents the number of attributes in a single record.

4.2.5. Price per record

The price per record Pr_{rec} is a quantitative factor and represents the unit price of a record. Naturally, it is the product of the price per attribute Pr_{attr} and the attribute count $Count_{attr}$ in a single record. That is,

$$Pr_{rec} = Pr_{attr} \times Count_{attr} \tag{8}$$

The value of a raw dataset increases as the unit price per record increases.

4.2.6. Monetary value of raw dataset

Intuitively, the monetary value of a raw dataset $Cost_{rd}$ is the product of the sensitivity of the dataset *SD*, the size of the dataset $Size_{ds}$, and the price per record Pr_{rec}, which is formulated as follows.

$$Cost_{rd} = SD \times Size_{ds} \times Pr_{rec} \tag{9}$$

In our model, the monetary value of a raw dataset roughly corresponds to the cost of data collection. In some scenarios, the data collection process may not be replicable (e.g., the case of health data collection). We argue that, even for these scenarios, our model is still meaningful in the sense that the derivation of the monetary value of raw data in our model provides a critical foundation for the negotiation between data owners and data recipients. From the viewpoint of the data recipients, the composition of the value of the raw data in our model adds transparency to the negotiation process with the data owner. This is important, especially considering the fact that health data collection is not replicable by external entities, which implies that the data recipients cannot obtain another quote to determine whether the asked price is reasonable. From the viewpoint of the data owners, it is indispensable to have a way to measure the value of the raw data in order to perform cost-benefit analysis no matter whether the data collection process is replicable or not. This is critical for HICs to make a right decision on whether or not to publish a dataset.

4.2.7. Cost of distortion

To determine the cost of distortion CoD for general data analysis, the HIC needs to calculate the discernibility ratio (DR) of the raw dataset before applying any anonymization process, denoted by DR_{before}, and the DR of the anonymized data, denoted by DR_{after}. The information distortion (or utility loss) due to anonymization can be captured by the difference between DR_{before} and DR_{after}. Therefore, the cost of distortion CoD becomes:

$$CoD = Cost_{rd} \times (DR_{after} - DR_{before}) \tag{10}$$

Note that DR_{after} is always greater than or equal to DR_{before}.

4.2.8. Cost in terms of classification quality

To determine the cost in terms of classification quality CCQ, we make use of the difference between baseline accuracy (BA) and classification accuracy (CA). Recall that BA measures the accuracy of classification analysis on raw data while CA measures the accuracy on anonymized data. $BA - CA$ represents the cost of anonymization in terms of classification accuracy. Therefore, the cost in terms of classification quality is defined as

$$CCQ = Cost_{rd} \times (BA - CA) \tag{11}$$

4.2.9. Value of anonymized data

Naturally, the value of anonymized data is the difference of the monetary value of raw data and the cost of anonymization (either CoD or CCQ). It is the earning of an institution by selling anonymized data for research or commercial purposes. For general analysis tasks, the value of anonymized data Val_{AD} is defined as:

$$Val_{AD} = Cost_{rd} - CoD \tag{12}$$

For classification analysis, the value of anonymized data Val_{AD} is defined as:

$$Val_{AD} = Cost_{rd} - CCQ \tag{13}$$

4.2.10. Likelihood of privacy breach

The likelihood of privacy breach L_{pb} measures an adversary's capability of inferring the sensitive attribute value of a victim, in percentage, based on an attack model (see Section 4.3 for details) by using his/her background knowledge. Let us assume that the victim's record is in the released dataset and the adversary knows the victim's QID. Formally, L_{pb} for general and classification analysis cases[1] is defined as:

$$L_{pb} = \frac{Total\ records\ count\ on\ Sen_{val}}{Total\ records\ count\ on\ class\ label\ Sen_{attr}} \tag{14}$$

where Sen_{val} denotes the value of the sensitive attribute and Sen_{attr} denotes the sensitive attribute in the dataset.

4.2.11. Expected cost of lawsuit

The expected cost of lawsuit $Ecost_{lwst}$ is due to monetary fines or penalties applicable by law in real life for data privacy breach incidents. It is a qualitative factor because its monetary value may vary depending on the nature of a privacy breach. $Ecost_{lwst}$ increases as the level of data sensitivity l increases. The approximate value of $Ecost_{lwst}$[1] can be estimated based on the historical trends of privacy breaches. For example, according to the new HITECH penalty scheme [47], the penalty for a violation in which it is known that the violation was due to *reasonable cause* and not to *willful neglect* is an amount not less than $1000 or more than $50,000 for each violation. In fact, the lawsuit cost is not fixed. It varies with the probability of attack and affects the potential compensation cost.

The expected cost of lawsuit should not be taken directly into the account of compensation cost. We attempt to estimate the lawsuit cost which varies depend upon the applied privacy protection measures. An adversary may infer sensitive information from the anonymized dataset using precision and recall measures employed in the equation of probability of attack to exploit inherent weakness of privacy protection method. An HIC may recognize the implications of privacy breach and the associated compensation costs prior to sharing medical dataset. Readers may refer to the study on data privacy breach incidents [48].

4.2.12. Probability of attack

The probability of attack $Prob_{atk}$ is caused by the implicit weakness of privacy protection methods. $Prob_{atk}$ changes with respect to the chosen privacy model and its level of privacy protection. It is taken by calculating the F-measure on the sensitive attribute value Sen_{val}. F-measure is a weighted harmonic mean of precision and recall. Precision and recall are used to measure the quality of results which an adversary can exploit for privacy attacks. Formally, $Prob_{atk}$[1] for general and classification analysis is defined as:

$$Prob_{atk} = \frac{2 \times (Precision\ on\ Sen_{val} \times Recall\ on\ Sen_{val})}{Precision\ on\ Sen_{val} + Recall\ on\ Sen_{val}} \tag{15}$$

4.2.13. Potential compensation cost

The potential compensation cost PCC indicates how the compensation cost would vary in real life in the presence of an attack and its severity level. PCC is affected by the choice of the privacy model and its level of privacy protection. In general, more stringent privacy parameters imply less chance of a privacy attack. We hypothesize that privacy attacks would have an exponential impact on the compensation cost due to costly litigation processes [49]. There is no specific monetary value for compensation cost in [49], but a person who suffers financial loss due to the disclosure of his/her sensitive information may claim for compensation. As the probability of attack $Prob_{atk}$ increases, PCC also increases. Formally, PCC[1] for general and classification analysis case is defined as:

$$PCC = \exp(Prob_{atk}) \times Ecost_{lwst} \tag{16}$$

4.2.14. Fixed operating cost

The fixed operating cost FOC is a quantitative factor, and its value is independent of the employed anonymization process. Fixed operating costs may include, for example, rent, utilities, payments for equipments, and system maintenance. FOC remains the same regardless of the changes in Value of Anonymized Data Val_{AD}.

4.2.15. Potential damage cost

The potential damage cost PDC is the cost associated with data privacy breaches. When an adversary attempts to infer the

[1] We do not use separate notations of $L_{pb}, Ecost_{lwst}, Prob_{atk}, PCC, PDC$, or Opt_{val} for general analysis and classification analysis cases.

sensitive value of a victim from the released anonymized dataset using an attack model (as discussed in Section 4.3), there is a risk of sensitive information disclosure which is measured by the likelihood of a privacy breach in Section 4.2.10. *PDC* signifies the costs of mitigating the effects of a privacy breach. It may include significant costs incurred in sending mandatory breach notifications, dealing with regulatory investigations, hiring external auditors, facing class action litigation, and losing goodwill of the general public due to decreased patient loyalty [50]. Therefore, *PDC* to the HICs is determined by the likelihood of privacy breach L_{pb}, the potential compensation cost *PCC*, and the fixed operating cost *FOC*. As suggested by existing studies [14,51], we deem that L_{pb} would have an exponential impact on the potential damage cost due to the fact that a plaintiff seeks a remedy for alleged harms, such as actual financial loss incurred from the identity theft, emotional distress, or possible future losses [52]. Formally, *PDC*[1] for general and classification analysis is defined as:

$$PDC = \exp(L_{pb}) \times PCC + FOC \tag{17}$$

4.2.16. Net value

The net value *NV* shows due diligence in evaluating the cost factors. *NV* is used in cost-benefit analysis to quantify the difference between the value of anonymized data and the potential damage cost on different privacy protection levels. The net value provides a single monetary value when the values of the previous factors are determined. Formally, NV_{ga} for general analysis and NV_{ca} for classification analysis are calculated respectively as follows.

$$NV_{ga} = Val_{AD_{ga}} - PDC \tag{18}$$

$$NV_{ca} = Val_{AD_{ca}} - PDC \tag{19}$$

4.2.17. Optimal value

The optimal value Opt_{val} is defined to be the maximal net value *NV* for general analysis or classification analysis. It can be obtained by calculating *NV* under different privacy models, different anonymization algorithms and different privacy parameters and choosing the maximal one. In this way, our model guides an HIC to find the optimal value of publishing health data. Formally, Opt_{val}[1] is defined as:

$$Opt_{val} = \max(NV) \tag{20}$$

4.3. Attack model

Let *D* be the raw patient data as shown in Table 1, and *D′* be the anonymized version of patient data as shown in Table 2. Recall that *Disease* is the sensitive attribute and *Blood transfusion* is the class attribute. Assume the anonymized data table *D′* is released together with the classifier. The adversary may have some additional background knowledge about a victim. Without loss of generality, assume that he/she knows that the victim is in the table and knows the victim's *qid*. Our attack model is similar to [53] in the sense that we are thinking from an adversary's perspective and predicting the sensitive attribute value of a target victim who is a participant in the anonymized training data. An adversary cannot link a record to an individual, although he can infer some sensitive values with a high confidence. We set the sensitive attribute *Disease* as the class label and then use classification algorithm *C*4.5 to infer the sensitive attribute of individuals. In our attack model we use precision and recall measures to evaluate the quality of results on the class label *Disease*. Below we provide the details of these measures followed by an example of a confusion matrix.

4.3.1. F-measure

Precision is a measure of exactness or quality that is formally defined as the number of correctly classified positive elements divided by the total number of elements classified as positive. Let *TP* be true positive, *FP* be false positive, and *FN* be false negative.

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

Recall is a measure of completeness or quantity which is formally defined as the number of correctly classified positive elements divided by the total number of actual positive elements.

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

F-measure is the harmonic mean of precision and recall and is formally defined as:

$$F - measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{23}$$

4.3.2. Confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a well-known classification model. The performance of a classification model on a sensitive attribute is evaluated using the data in the matrix.

Example 3. Consider the anonymous table *D′* in Table 2. An adversary sets the sensitive attribute *Disease* as a class on *D′*. This results in a new data table D^*. The adversary then uses the classification model *C*4.5 on D^* to infer sensitive attributes of individuals. The confusion matrix for the three-class classifier is shown in Table 3. The rows correspond to the actual classes of the raw records, and the columns correspond to the predictions made by the model. The values on the diagonal represent the number of correctly classified instances; other values show the errors.

We next show the calculation of the performance measures for the above confusion matrix on sensitive values. For the sensitive value HIV, $TP = 3, FN = 0$, and $FP = 0$. So we obtain $Precision = 1, Recall = 1$, and $F\text{-}measure = 1$. An adversary may use these performance measures to determine the success rate of a privacy attack. *F*-measure represents the probability of attack $Prob_{atk}$. When its value equals 1, it means that there is a 100% chance of a successful attack.

4.3.3. Background knowledge attack

As demonstrated by the attack model discussed in Example 3, an adversary may apply the *C*4.5 classifier on data table D^* to predict the sensitive value of an individual who is a part of the anonymized training data. In addition, we assume that the adversary knows that the victim is in the table and also knows the victim's *qid* (i.e., $\langle F, Painter\rangle$). By applying this background knowledge to the anonymized training data, the adversary finds a total of 4 records on the class attribute *Disease* with the sensitive value *HIV*. So, the likelihood of privacy breach L_{pb} for this case becomes 4/4 that is calculated according to Eq. (14). This implies that the

Table 3
Confusion matrix.

	Predicted class		
	A	**B**	**C**
Actual class			
HIV (**A**)	3	0	0
Asthma (**B**)	0	1	0
Migraine (**C**)	0	1	0

adversary has 100% confidence of inferring the sensitive disease value of the victim.

5. Empirical study

In this section, our objectives are to study the impact of enforcing different data anonymization methods at different privacy protection levels on information utility for data mining in terms of monetary value. More specifically, we perform experiments: (1) to measure the classification accuracy on the class attribute and the sensitive attribute, (2) to measure the cost of distortion, (3) to measure the cost in terms of classification quality, (4) to estimate the probability of attack by using precision and recall performance measures, (5) to quantify the likelihood of a privacy breach impacted by an adversary's background knowledge about victims, and (6) to perform net cost-benefit analysis to measure the value of anonymized data, the potential damage cost, and the optimal value on the released data.

In our experiments, we employ the real-life dataset *Adult*,[2] which has been widely used for different research purposes and has been the *de facto* benchmark for comparing the performance of anonymization algorithms [54,44,9]. It contains 45,222 records with 8 categorical attributes, 6 numerical attributes, and a binary *Income* class attribute (records with unknown instances are removed). In our study, we consider *Income* as the class attribute, denoted by *Class_Income*, and *MaritalStatus* be the sensitive attribute, denoted by *Sens_MaritalStatus*, and the remaining 13 attributes as *QID*. We consider the values *Married-civ-spouse* and *Divorced* of *MaritalStatus* as sensitive. All experiments were performed on an Intel dual core 1.8 GHz PC with 2 GB memory.

5.1. Classification accuracy on class label and sensitive attribute

Fig. 3 depicts the classification accuracy *CA* for general data analysis with privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that the *CA* on the class attribute *Income* generally decreases as *K* or *L* increases, but not monotonically. For example, the *CA* on *Income* increases slightly by 0.1% when *K* increases from 40 to 50 for $L = 2$. Similarly, the *CA* on the sensitive attribute *MaritalStatus* generally decreases as *K* or *L* increases, but with some irregularities. For example, the *CA* on *MaritalStatus* increases by 0.4% when *K* increases from 10 to 20 for $L = 2$, and it increases by 0.2% when *K* increases from 30 to 40 for $L = 6$. In these cases, the *CA* increases because generalization can eliminate some noise. However, as *L* increases to 6, the *CA* of *LKC*-privacy equals the *CA* of the traditional *K*-anonymity model for both *Income* and *MaritalStatus*. This is due to the fact that *Adult* does *not* contain any combination of 6 or more attributes whose every subset of attributes satisfies the privacy model. In other words, all privacy threats involving more than 6 attributes can be eliminated by removing the privacy threats involving less than 6 attributes.

Fig. 4 presents the classification accuracy *CA* for classification analysis with identical parameter settings to those of Fig. 3. We observe the similar trend that the *CA* on the class attribute *Income* generally decreases as *L* increases, but not monotonically with the increase of *K*. For example, *CA* on *Income* increases by 3.1% when *K* increases from 10 to 20 for $L = 4$ and $L = 6$. Similarly, the *CA* on the sensitive attribute *MaritalStatus* generally decreases as *L* increases, but not monotonically with the increase of *K*. For example, the *CA* on *MaritalStatus* increases slightly, by 0.6%, when *K* increases from 30 to 50 for $L = 4$ and $L = 6$. The *CA* of *LKC*-privacy equals the *CA* of the traditional *K*-anonymity for both *Income* and *MaritalStatus* when $L = 4$ and $L = 6$ due to the reason explained before.

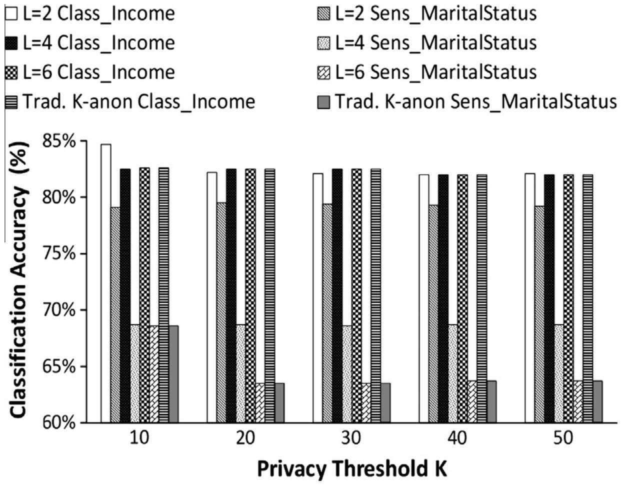

Fig. 3. CA on Income and MaritalStatus for general analysis.

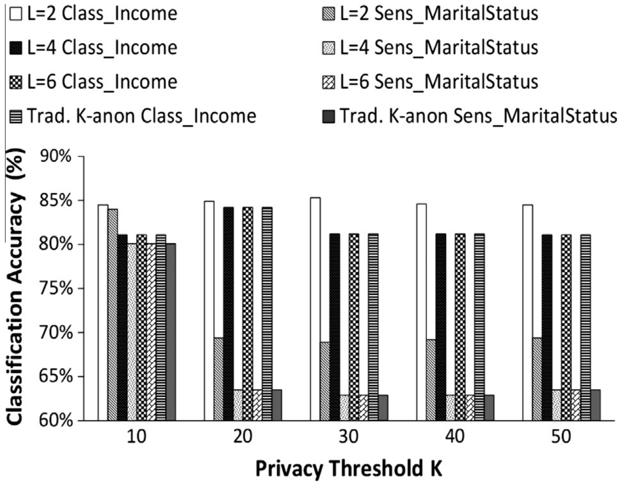

Fig. 4. CA on Income and MaritalStatus for classification analysis.

5.2. Cost of distortion

Suppose the sensitivity of the dataset $SD = 3$, the price per attribute $Pr_{attr} = .1$, the number of attributes per record $Count_{attr} = 13$, the expected cost of lawsuit $Ecost_{lwst}$ = \$10,000, and the size of dataset $Size_{ds}$ = 45,222.

Fig. 5 depicts the cost of distortion *CoD* for general data analysis with a privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. As expected, we observe that *CoD* generally increases as *K* or *L* increases. The *CoD* remains the same for several different parameter settings because there is no change in DR_{after}. For the same reason, the *CoD* of *LKC*-privacy equals the *CoD* of *K*-anonymity when $L = 6$. Though not shown in Fig. 5, *CoD* is insensitive to change of confidence threshold $10\% \leqslant C \leqslant 50\%$.

5.3. Cost in terms of classification quality

Again suppose the sensitivity of the dataset $SD = 3$, the price per attribute $Pr_{attr} = \$.1$, the number of attributes per record $Count_{attr} = 13$, the expected cost of lawsuit $Ecost_{lwst}$ = \$10,000,

[2] Available at: http://archive.ics.uci.edu/ml/datasets/Adult.

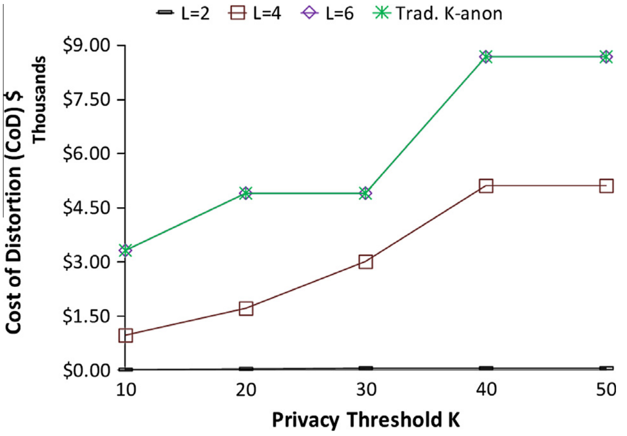

Fig. 5. Cost of distortion for general analysis.

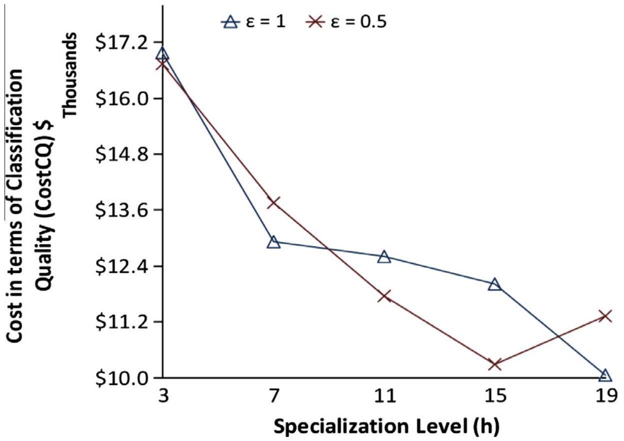

Fig. 7. *CostCQ* of DiffGen for classification analysis.

and the size of dataset $Size_{ds}$ = 45,222. The *baseline accuracy* (*BA*) as calculated on raw data is 85.3%.

Fig. 6 presents the cost in terms of classification quality *CCQ* for classification analysis with privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = 2, 4, 6$; and confidence threshold $C = 50\%$. It can be observed that *CCQ* generally increases as L increases, but does not exhibit obvious monotonicity with the increase of K. For example, *CCQ* decreases by $\$5467.34$ when K increases from 10 to 20 when $L = 4$ and $L = 6$. This is because the cost of anonymization in terms of classification accuracy $(BA - CA)$ is reduced from 4.2% to 1.1%, and this aids in finding the sub-optimal solution. *CCQ* is also insensitive to the change of confidence threshold $10\% \leqslant C \leqslant 50\%$.

Fig. 7 depicts the cost in terms of classification quality *CostCQ* using *DiffGen* for classification analysis with privacy parameters $\epsilon = 0.5$ and 1.0 and specialization levels $3 \leqslant h \leqslant 19$. We use 30,162 records in *Adult* dataset to build the classifier and then measure the accuracy on the remaining 15,060 records. We use 10-fold cross-validation to estimate the average accuracy. We observe that *CostCQ* generally decreases as the specialization level h increases, except when privacy budget $\epsilon = 0.5$ and specialization level h increases from 15 to 19. It is because when ϵ is small, having too many levels makes each specialization less accurate.

5.4. Probability of attack

Suppose the sensitivity of the dataset $SD = 2$, the price per attribute $Pr_{attr} = \$.1$, the number of attributes per record $Count_{attr} = 13$, the expected cost of lawsuit $Ecost_{lwst}$ = $10,000, and the size of dataset $Size_{ds}$ = 45,222.

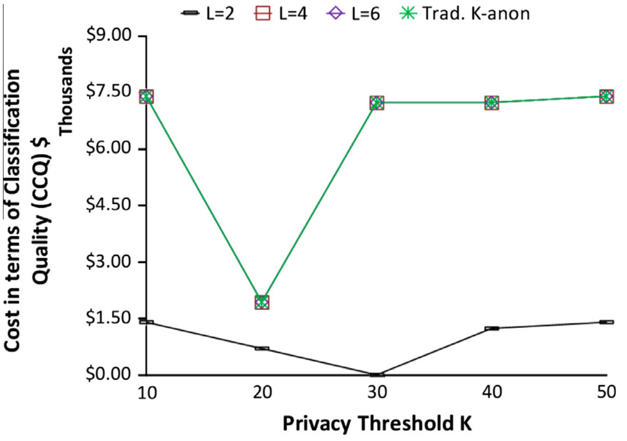

Fig. 6. Cost in terms of classification quality for classification analysis.

Fig. 8 presents the probability of attack $Prob_{atk}$ and its consequences in the terms of potential compensation cost *PCC* for the sensitive value *Married-civ-spouse* in case of general data analysis, where privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that *PCC* generally decreases as $Prob_{atk}$ decreases or L increases, but not monotonically with the increase of K. The potential compensation cost is reduced to $40789.40, corresponding to the lowest probability of attack 71.27% when $L = 6$ or $K \geqslant 20$. This is consistent with the theoretical analysis that more stringent privacy requirements lead to lower probabilities of privacy attacks and thus less potential compensation costs.

Fig. 9 shows the probability of attack $Prob_{atk}$ for the sensitive value *Married-civ-spouse* in the case of classification analysis with privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We can observe the similar trend that $Prob_{atk}$ generally decreases as K or L increases, which also conforms to the theoretical analysis.

5.5. Likelihood of privacy breach

Suppose an adversary has background knowledge about a male victim that his *age* is between 46 and 50, his *education-num* is $\geqslant$ 13, his *native-country* is *Canada*, and his *salary* is $\geqslant$50,000.

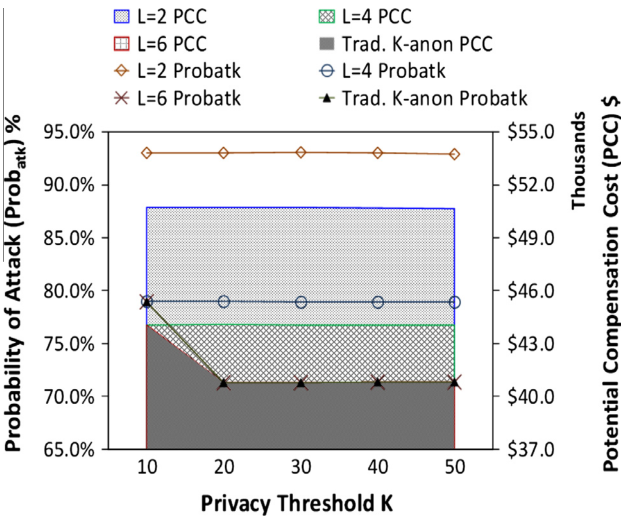

Fig. 8. Probability of attack and *PCC* for general analysis.

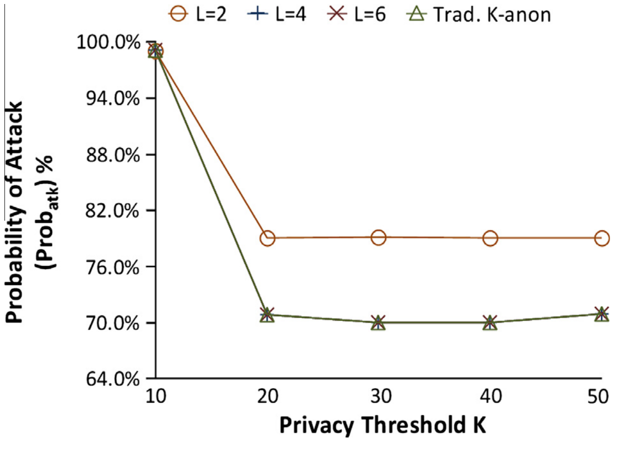

Fig. 9. Probability of attack for classification analysis.

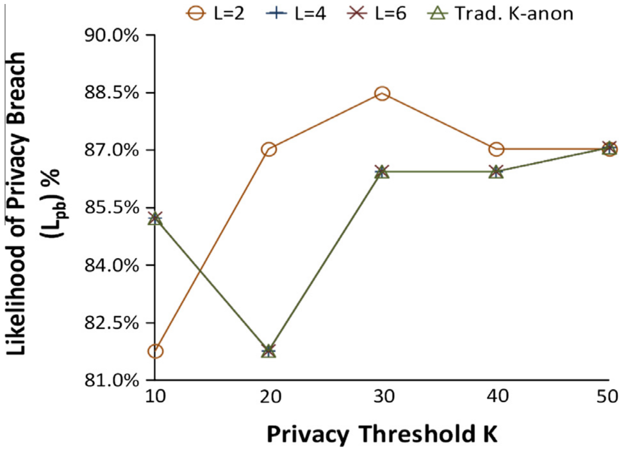

Fig. 11. Likelihood of privacy breach for classification analysis.

Fig. 10 shows the likelihood of privacy breach L_{pb} when the adversary applies her background knowledge on the sensitive value *Married-civ-spouse* in the case of general data analysis, where privacy threshold $10 \leqslant K \leqslant 50$, adversary's knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We observe that L_{pb} changes as the adversary's knowledge L increases, but it is less sensitive to the increase of K when $K \geqslant 20$. It is interesting to observe that L_{pb} increases from 87.04% to 88.49%, as L increases from 2 to 4. This seems counter-intuitive as normally a larger L value implies a smaller likelihood of privacy breach. We believe that it is due to the fact that the TDS algorithm only identifies sub-optimal solutions.

Fig. 11 presents the likelihood of privacy breach L_{pb} for classification analysis with the identical parameter settings to those of Fig. 10. We observe that in general L_{pb} decreases as L increases, but shows some irregularities with the increase of K. This anti-monotonic property of the TDS algorithm helps identify the sub-optimal solution. The L_{pb} of *LKC*-privacy equals the L_{pb} of K-anonymity when $L = 4$ and $L = 6$ because the classification accuracy on the sensitive *MaritalStatus* attribute remains unchanged with the increase of L. Though not shown in the figure, L_{pb} is insensitive to change of confidence threshold $10\% \leqslant C \leqslant 50\%$.

5.6. Net cost-benefit analysis

Suppose the price per attribute $Pr_{attr} = \$.1$, the number of attributes per record $Count_{attr} = 13$, the expected cost of lawsuit $Ecost_{lwst} = \$10,000$, the fixed operating cost $FOC = \$100$, and the size of dataset $Size_{ds} = 45,222$. *Baseline accuracy* (*BA*) calculated on the raw data is 85.3%.

In Fig. 12, we conduct cost-benefit analysis for general data analysis by giving $Val_{AD_{ga}}$, PDC, NV_{ga}, and $Ecost_{lwst}$ with the sensitivity of the dataset $1 \leqslant SD \leqslant 5$, privacy threshold $K = 30$, background knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. It can be observed that $Val_{AD_{ga}}$ and *PDC* generally increase as the sensitivity of the dataset *SD* increases, and decrease as L increases. There are several interesting findings in Fig. 12. First, when $L = 2$, with the increase of *SD*, the *NV* of publishing health data becomes negative, suggesting that insufficient privacy protection can make data owners incur loss of money. Second, publishing health data with a proper trade-off between privacy and data utility can bring in substantial earnings for data owners. For example, when $SD = 5$, choosing $L = 6$ for *LKC*-privacy or $K = 30$ for K-anonymity results in a *NV* greater than $30,000.

In Fig. 13, we identify the optimal value under different privacy models and different privacy parameters for classification analysis. We show $Val_{AD_{ca}}$, PDC and NV_{ca}, and Opt_{val} with the sensitivity of the dataset $SD = 3$, privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = 2, 4, 6$, and confidence threshold $C = 50\%$. We can observe that the optimal value (the maximum NV_{ca} in all parameter settings) $36316.31 is achieved when $L = 4$ or 6 and $K = 20$ for *LKC*-privacy or when $K = 20$ for K-anonymity. This result suggests

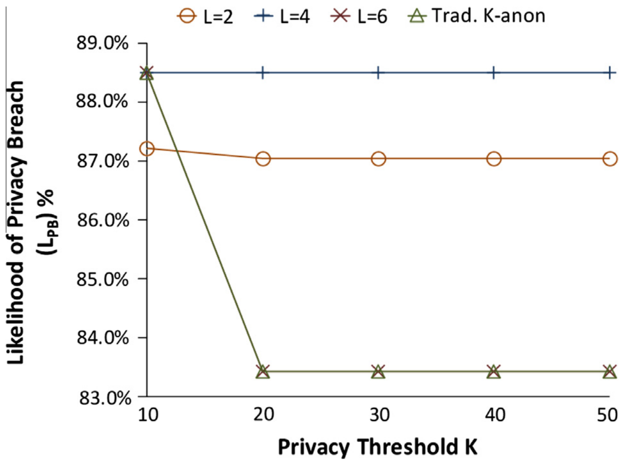

Fig. 10. Likelihood of privacy breach for general analysis.

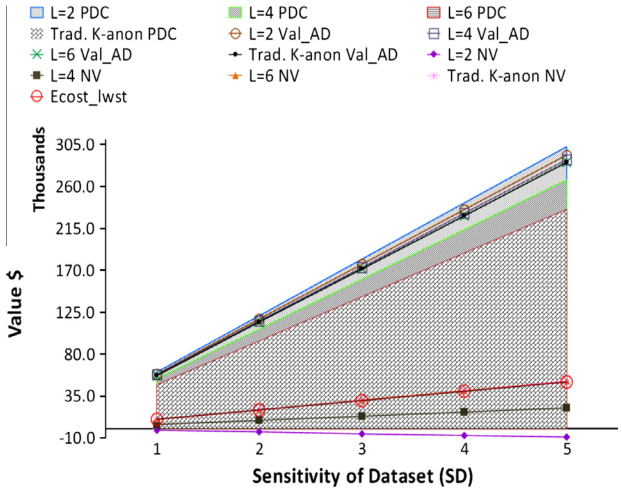

Fig. 12. Cost-benefit analysis for general data analysis.

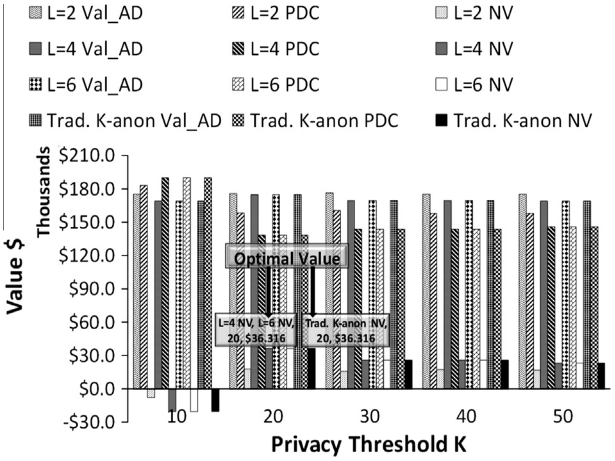

Fig. 13. Optimal value for classification analysis.

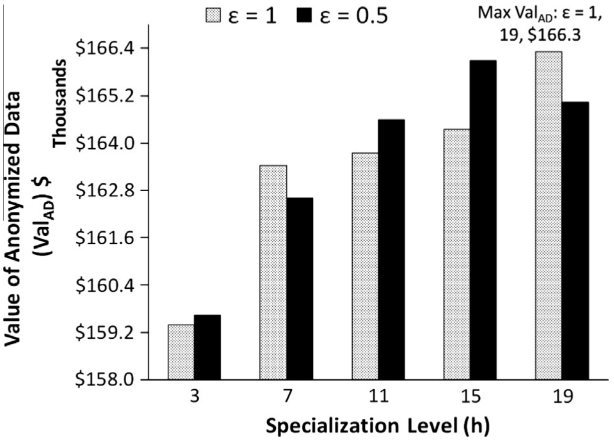

Fig. 14. *Val$_{AD}$* of DiffGen for classification analysis.

that the optimal value is obtained by carefully balancing the trade-off between privacy and data utility. The *NV* values under various settings exhibit huge differences. When $K = 10$, the net values are negative, suggesting that the HIC should not publish the data with weak privacy protection. The result confirm the benefit of employing our proposed model in health data publishing.

Fig. 14 depicts the value of anonymized data using *DiffGen* for classification analysis with privacy parameters $\epsilon = 0.5, 1.0$ and specialization levels $3 \leqslant h \leqslant 19$. We observe that the *Val$_{AD}$* generally increases as the specialization level h increases, because more specializations preserve more information utility. But when $\epsilon = 0.5$, having too many specializations may negatively impact utility, since excessive noise is added to each specialization. This explains the fall when $\epsilon = 0.5$ and $h = 19$. The maximum *Val$_{AD}$* \$166312.95 is achieved when $\epsilon = 1.0$ and $h = 19$. Since *DiffGen* prevents data breaches from adversaries with arbitrary background knowledge, we learn that applying adversary knowledge, as mentioned in Section 5.5, on differentially private data does not result in significant privacy breaches. Therefore, it reflects no potential damage cost.

5.7. Summary of empirical study

We show how to use our proposed model to search for the optimal value through extensive experiments on real-life data. Under our model, an HIC can compare costs and benefits by choosing different privacy models, different privacy parameters and different anonymization algorithms so as to achieve the optimal value. From the experimental results, we learn that the optimal value is obtained by carefully balancing the trade-off between privacy and utility. Choosing either too weak or overly strong privacy protection often leads to less desirable net values. This fact demonstrates the benefit of our analytical cost model for privacy-preserving health data publishing.

6. Discussion

Proposing an analytical cost model to conduct cost-benefit analysis for privacy-preserving health data publishing is a challenging task. In this section, we justify the design of our cost model, identify some of its limitations that are inherent known problems of cost-benefit analysis, and finally discuss the possibility of incorporating alternative cost models.

Under the theory of cost-benefit analysis, the most important steps of building a cost model are the selection of cost factors and their valuations [55]. We follow the procedure used in cost-benefit analysis to determine the factors that should be included in our cost model.

- Identify the problem.
- Gather the stakeholders' requirements.
- Design the scenario as per requirements.
- Study the possible factors that can be incorporated in the scenario.
- Identify each factor's properties in the problem domain.
- Evaluate the importance of each factor.

In our problem, patients, HICs (on behalf of data owners) and data recipients are the most important stakeholders. For these stakeholders, we identify the most important factors, as illustrated in Fig. 2, to reflect patients' requirements on privacy, data recipients' requirements on data utility and HICs' requirements on properly balancing privacy and utility in order to publish health data for profit. Consequently, we identify the following relevant factors:

1. Privacy models for balancing the trade-off between data privacy and utility.
2. Anonymization methods to generalize the data (bottom-up/top-down).
3. Cost of anonymization due to information loss.
4. Monetary value of personal data.
5. Number of records in the dataset.
6. Sensitivity level of dataset.
7. Benefit/value earned from anonymized data.
8. Monetary fines or penalties applicable by law.
9. Compensation cost caused by the weakness of the privacy protection method.
10. Risk of privacy breach.
11. Potential damage cost due to data privacy breach.

However, due to the nature of cost-benefit analysis, there are some inherent limitations in our model. Inaccuracies in cost-benefit analysis may arise in many steps. One main source of inaccuracies comes from the decision of what factors count [55]. In spite of our very best efforts, it is not possible to consider all meaningful factors in our models. This will inevitably incur inaccuracies, which is referred to as *omission errors* in cost-benefit analysis. Similarly, for the identified factors, there are alternative ways to valuate them, leading to *valuation errors*. Indeed, as pointed out by [56], there is no common accepted methodology for estimating the value of personal data. Fortunately, these errors do not diminish the value of cost-benefit analysis, and they are expected to decline over time, for example, due to increased knowledge and subsequent ex post analysis [55]. Under our model, these errors will decrease by adding/removing cost factors and adjusting their monetary values according to the specific application scenario and ex post analysis. We note that any factor that demonstrates its applicability and usefulness in an application scenario could be incorporated in our model. Some possible candidates could be the cost of data preprocessing, standard data format, service cost, hardware and software infrastructure to adapt the change, and hiring experts to develop anonymization methods. It is worth mentioning that our model is open to modifications of factors, and therefore can be tuned for different application scenarios.

In general, there may exist many reasonable alternative models. For example, though not directly relevant to our problem, Yin [57] discusses the costs and benefits of sharing electronic health records. Romanosky and Acquisti [48] provide an economic cost model based on economic theory to analyze the consumer privacy costs. Both of them could be adapted to address our problem. However, in cost-benefit analysis, it is unrealistic (and maybe not necessary) to identify all possible models. In practice, usually only one model will be analyzed with the status quo [55]. Therefore, it is necessary to discuss the connection between our proposed model and other possible cost models. In essence, a cost model is composed of its cost factors and their valuations. Our proposed model can incorporate the factors and their valuations from other cost models based on the application scenario. For example, if the data analysis task is known to be one of those identified in [26] (e.g., statistical hypothesis tests), the cost of anonymization in our model could be accordingly updated by using the corresponding utility metric. We stress that what factors to use should be decided based on the application scenario, and could be continuously adjusted over time.

7. Conclusion

In this paper, we propose an analytical cost model that can benefit health information custodians (HICs) from making better decisions on sharing health data for secondary and commercial uses. Our model quantifies the trade-off between individual privacy and data utility in terms of monetary value for both general data analysis and classification analysis. Our proposed model integrates relevant quantitative and qualitative cost factors associated with the value of anonymized data and the potential damage cost and effectively guides HICs to achieve the optimal value to privacy-preserving health data publishing. Our analytical cost model and the identified factors also apply to other privacy-preserving data publishing scenarios for other types of data, such as transaction data [28], trajectory data [29], and social network data [33]. We expect this work to shed light on future research that studies the trade-off between privacy protection and information utility.

Acknowledgments

The research is supported in part by the Discovery Grants (356065-2013) and Canada Graduate Scholarships from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Neclerio JM, Cheney K, Goldman C, Clark LW. Adopting electronic medical records: what do the new federal incentives mean to your individual physician practice? J Med Pract Manage 2009;25(1):44–8.
[2] Koh HC, Tan G. Data mining applications in healthcare. Healthcare Inf Manag 2005;19(2):64–72.
[3] Beaver K, Herold R. The practical guide to HIPAA privacy and security compliance. Auerbach; 2003.
[4] Al Faresi A, Wijesekera D, Moidu K. A comprehensive privacy-aware authorization framework founded on HIPAA privacy rules. In: Proceedings of the 1st ACM international health informatics symposium. IHI '10. ACM; 2010. p. 637–46. ISBN 978-1-4503-0030-8. http://dx.doi.org/10.1145/1882992.1883093.
[5] Baumer D, Earp JB, Payton FC. Privacy of medical records: IT implications of HIPAA. SIGCAS Comput Soc 2000;30(4).
[6] Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. IEEE Trans Knowledge Data Eng 2001. Available from: http://dataprivacylab.org/dataprivacy/projects/kanonymity/index3.html.
[7] Keckley PH, Coughlin S, Gupta S. Privacy and security in health care: a fresh look. Deloitte Center for Health Solutions; 2011. Available from: https://www.deloitte.com/assets/Dcom-UnitedStates/Local%20Assets/Documents/Health%20Reform%20Issues%20Briefs/US_CHS_PrivacyandSecurityinHealthCare_022111.pdf.
[8] Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. ACM Comput Surv 2010;42(4):14:1–14:53.
[9] Mohammed N, Chen R, Fung BCM, Yu PS. Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '11. ACM; 2011. p. 493–501. ISBN 978-1-4503-0813-7. http://dx.doi.org/10.1145/2020408.2020487.
[10] Kifer D, Machanavajjhala A. No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD international conference on management of data. SIGMOD '11. ACM; 2011. p. 193–204. ISBN 978-1-4503-0661-4. http://dx.doi.org/10.1145/1989323.1989345.
[11] Gardner J, Xiong L, Xiao Y, Gao J, Post AR, Jiang X, et al. SHARE: system design and case studies for statistical health information release. J Am Med Inf Assoc (JAMIA) 2013;20(1):109–16.
[12] Ponemon Institute LLC. The 3rd annual benchmark study on patient privacy and data security. Tech rep; 2012.
[13] Witzleb N. Monetary remedies for breach of confidence in privacy cases. Legal Stud 2007;27(3):430–64.
[14] Backman, P, Levin K. Privacy breaches – impact, notification and strategic plans. Aird and Berlis LLP; 2011. Available form: http://www.lexology.com/library/detail.aspx?g=6b37a60b-e179-419a-a822-c1fe47cf49e3.
[15] Mohammed N, Fung BCM, Hung PC, Lee CK. Anonymizing healthcare data: a case study on the blood transfusion service. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '09. ACM; 2009. p. 1285–94. ISBN 978-1-60558-495-9. http://dx.doi.org/10.1145/1557019.1557157.
[16] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography, lecture notes in computer science, vol. 3876. Heidelberg: Springer Berlin; 2006. p. 265–84. ISBN 978-3-540-32731-8. http://dx.doi.org/10.1007/11681878_14.
[17] Yassine A, Shirmohammadi S. Privacy and the market for private data: a negotiation model to capitalize on private data. In: Computer systems and applications, 2008. AICCSA 2008. IEEE/ACS international conference on; 2008. p. 669–78. http://dx.doi.org/10.1109/AICCSA.2008.4493601.
[18] Danthine J-P, Donaldson JB. Intermediate financial theory. 2nd ed. Elsevier Academic Press; 2005. p. 377.
[19] Jentzsch N, Preibusch S, Harasser A. Study on monetising privacy: an economic model for pricing personal information. ENISA; 2012.
[20] Zielinski MP, Olivier MS. On the use of economic price theory to find the optimum levels of privacy and information utility in non-perturbative microdata anonymisation. Data Knowledge Eng (DKE) 2010;69(5):399–423.
[21] Loukides G, Shao J. Data utility and privacy protection trade-off in k-anonymisation. In: Proceedings of the 2008 international workshop on privacy and anonymity in information society. PAIS '08. ACM; 2008. p. 36–45. ISBN 978-1-59593-965-4. http://dx.doi.org/10.1145/1379287.1379296.
[22] Li T, Li N. On the tradeoff between privacy and utility in data publishing. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '09. ACM; 2009. p. 517–26. ISBN 978-1-60558-495-9. http://dx.doi.org/10.1145/1557019.1557079.
[23] Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms. In: Proceedings of the forty-first annual ACM symposium on theory of computing. STOC '09. ACM; 2009. p. 351–60. ISBN 978-1-60558-506-2. http://dx.doi.org/10.1145/1536414.1536464.
[24] Alvim MS, Andrés ME, Chatzikokolakis K, Degano P, Palamidessi C. Differential privacy: on the trade-off between utility and information leakage. Form Asp Secur Trust, LNCS 2012;7140:39–54.
[25] Duncan GKMS, Stokes S. Disclosure risk vs. data utility: the r-u confidentiality map. Tech rep LA-UR-01-6428, Los Alamos National Laboratory, Los Alamos, NM; 2001.
[26] Shlomo N, Young C. Statistical disclosure control methods through a risk-utility framework. In: Proceedings of the 2006 CENEX-SDC project international conference on privacy in statistical databases. PSD '06. Springer-Verlag; 2006. p. 68–81. ISBN 3-540-49330-1, 978-3-540-49330-3. http://dx.doi.org/10.1007/11930242_7.
[27] Loukides G, Gkoulalas-Divanis A, Shao J. Assessing disclosure risk and data utility trade-off in transaction data anonymization. Int J Software Inf 2012;6(3):399–417.
[28] Chen R, Mohammed N, Fung BCM, Desai BC, Xiong L. Publishing set-valued data via differential privacy. Proc VLDB Endow 2011;4(11):1087–98.
[29] Abul O, Bonchi F, Nanni M. Never walk alone: uncertainty for anonymity in moving objects databases. In: Proc of ICDE; 2008. p. 376–85.

[30] Ghasemzadeh M, Fung BCM, Chen R, Awasthi A. Anonymizing trajectory data for passenger flow analysis. Transp Res Part C: Emerging Technol (TRC) 2014;39:63–79.

[31] Chen R, Fung BCM, Yu PS, Desai BC. Correlated network data publication via differential privacy. VLDB J 2013:1–24. ISSN 1066-8888. http://dx.doi.org/10.1007/s00778-013-0344-8.

[32] Fung BCM, Jin Y, Li J. Preserving privacy and frequent sharing patterns for social network data publishing. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. ASONAM '13. ACM; 2013. p. 479–85. ISBN 978-1-4503-2240-9. http://dx.doi.org/10.1145/2492517.2492603.

[33] Zhou B, Pei J. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. Knowledge Inf Syst 2011;28(1):47–77.

[34] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. ℓ-diversity: Privacy beyond k-anonymity. In: Data engineering, ICDE '06, Proceedings of the 22nd international conference on. Atlanta, GA; 2006. 24 pages. http://dx.doi.org/10.1109/ICDE.2006.1.

[35] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. ℓ-diversity: privacy beyond k-anonymity. ACM TKDD 2007;1(1).

[36] Truta TM, Bindu V. Privacy protection: *p*-sensitive *k*-anonymity property. In: Data engineering workshops, 2006. Proceedings of 22nd international conference on; 2006. p. 94. http://dx.doi.org/10.1109/ICDEW.2006.116.

[37] LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-aware anonymization. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '06. ACM; 2006. p. 277–86. ISBN 1-59593-339-5. http://dx.doi.org/10.1145/1150402.1150435.

[38] LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-aware anonymization techniques for large-scale datasets. ACM Trans Database Syst (TODS) 2008;33(3).

[39] Aggarwal CC. On k-anonymity and the curse of dimensionality. In: Proceedings of the 31st international conference on very large data bases. VLDB '05. VLDB Endowment; 2005. p. 901–9. ISBN 1-59593-154-6.

[40] Mohammed N, Jiang X, Chen R, Fung BCM, Ohno-Machado L. Privacy-preserving heterogeneous health data sharing. J Am Med Inf Assoc (JAMIA) 2013;20(3):462–9.

[41] Quinlan JR. C4.5: programs for machine learning. Morgan Kaufmann; 1993.

[42] Xu J, Wang W, Pei J, Wang X, Shi B, Fu AWC. Utility-based anonymization using local recoding. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2006. p. 785–90.

[43] Loukides G, Gkoulalas-Divanis A, Malin B. Towards utility-driven anonymization of transactions; 2010. Available from: http://arxiv.org/abs/0912.2548.

[44] Fung BCM, Wang K, Yu PS. Anonymizing classification data for privacy preservation. IEEE Trans Knowledge Data Eng (TKDE) 2007;19(5):711–25.

[45] Skowron A, Rauszer C. The discernibility matrices and functions in information systems. In: Słowiński R, editor. Intelligent decision support. Theory and decision library, vol. 11. Springer Netherlands; 1992. p. 331–62. ISBN 978-90-481-4194-4. http://dx.doi.org/10.1007/978-94-015-7975-9_21.

[46] Hirshleifer J, Glazer A, Hirshleifer D. Price theory and applications: decisions, markets, and information. 7th ed. Cambridge University Press; 2005.

[47] Department of Health and Human Services. Modifications to the HIPAA privacy, security, enforcement, and breach notification rules under the HITECH Act and the GINA Act; other modifications to the HIPAA rules 2013;(78 FR 5565):5565–702.

[48] Romanosky S, Acquisti A. Privacy costs and personal data protection: economic and legal perspectives. Berkeley Technol Law J 2014;24(4).

[49] Office of the Privacy Commissioner for Personal Data. Review of the personal data (privacy) ordinance; 2009.

[50] Bevitt A, Retzer K, Lopatowska J. Dealing with data breaches in europe and beyond. Data Protection Handbook 2011/12. Practical Law Company; 2012. Available from: http://www.mofo.com/files/Uploads/Images/110615-Dealing-with-Data-Breaches-in-Europe-and-Beyond.pdf.

[51] Acquisti A, Friedman A, Telang R. Is there a cost to privacy breaches? An event study. ICIS; 2006.

[52] Romanosky S, Hoffman DA, Acquisti A. Empirical analysis of data breach litigation; 2012.

[53] Kifer D. Attacks on privacy and deFinetti's theorem. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. SIGMOD '09. ACM; 2009. p. 127–38. ISBN 978-1-60558-551-2. http://dx.doi.org/10.1145/1559845.1559861.

[54] Hore B, Jammalamadaka RC, Mehrotra S. Flexible anonymization for privacy preserving data publishing: a systematic search based approach. SDM; 2007. p. 497–502.

[55] Boardman AE, Greenberg DH, Vining AR, Weimer DL. Cost-benefit analysis: concepts and practice. Pearson Prentice Hall; 2006.

[56] OECD. Exploring the economics of personal data: a survey of methodologies for measuring monetary value. OECD Digital Economy Papers 2013;(220).

[57] Yin PW, Review of the implementation of Electronic Health Record in Hong Kong. The University of Hong Kong; 2012. Available from: http://hub.hku.hk/handle/10722/184381.