Electronic Commerce Research and Applications 17 (2016) 19–37

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

Privacy-preserving data mashup model for trading person-specific information

Rashid Hussain Khokhar [a], Benjamin C.M. Fung [b,*], Farkhund Iqbal [c], Dima Alhadidi [c], Jamal Bentahar [a]

[a] CIISE, Concordia University, Montreal, QC, Canada
[b] School of Information Studies, McGill University, Montreal, QC, Canada
[c] College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article history:
Received 17 August 2015
Received in revised form 22 January 2016
Accepted 29 February 2016
Available online 10 March 2016

Keywords:
Privacy
Data utility
Data mashup
Business model
Monetary value

ABSTRACT

Business enterprises adopt cloud integration services to improve collaboration with their trading partners and to deliver quality data mining services. *Data-as-a-Service* (*DaaS*) mashup allows multiple enterprises to integrate their data upon the demand of consumers. Business enterprises face challenges not only to protect private data over the cloud but also to legally adhere to privacy compliance rules when trading person-specific data. They need an effective privacy-preserving business model to deal with the challenges in emerging markets. We propose a model that allows the collaboration of multiple enterprises for integrating their data and derives the contribution of each data provider by valuating the incorporated cost factors. This model serves as a guide for business decision-making, such as estimating the potential risk and finding the optimal value for publishing mashup data. Experiments on real-life data demonstrate that our approach can identify the optimal value in data mashup for different privacy models, including *K-anonymity*, *LKC-privacy*, and ϵ-*differential privacy*, with various anonymization algorithms and privacy parameters.

1. Introduction

Business enterprises have widely adopted web-based mashup technologies for collaboration with their trading partners. A web-based mashup involves the integration of information and services from multiple sources into a single web application. For example, real estate companies mashup their data and other third-party data with Google Maps for comprehensive market analysis. *Enterprise Mashup Markup Language* (*EMML*) is a standard proposed by the Open Mashup Alliance to improve collaboration among business enterprises and to reduce the risk and cost of mashup implementation (Roebuck 2012). Several companies including IBM, StrikeIron, Kapow Technologies, and others have been actively involved in leveraging various web-based mashup technologies such as Quick and Easily Done Wiki (QEDWiki), IBM Mashup Center, and *Data-as-a-Service* (*DaaS*). Business enterprises need to focus on a data-oriented perspective along with the initiatives of *Service-Oriented Architecture* (*SOA*).

DaaS is a cloud computing paradigm that provides data on demand to consumers over the Internet (Arafati et al. 2014). It is becoming popular in commercial setups because it provides flexible and cost-effective collaboration among business enterprises. In the e-market industry, enterprises conduct online market research to collect feedback about their products and services and to identify the demographic characteristics of customers by various means such as surveys, social networks, online purchases, posts, blogs, Internet browsing preferences, phone calls, or apps. The primary purpose in collecting personal information is to provide better services, which in turn generate higher revenue.

Fig. 1 presents an overview of a privacy-preserving data mashup e-market for trading person-specific information. The process consists of five steps. First, data providers register their available data on the registry hosted by the mashup coordinator, who can be a cloud service provider or one of the data providers. Second, data consumers (or data recipients) submit their data requests to the mashup coordinator. A "data request" can be a simple count query or a complicated data mining request. To provide a concrete scenario in the rest of the paper, we assume the data request is a data mining request for classification analysis. Third, a mashup coordinator dynamically determines the group of data providers, since a single data provider may not be able to fulfill the data requests from a data consumer, whose data can collectively fulfill

* Corresponding author.
E-mail addresses: r_khokh@ciise.concordia.ca (R.H. Khokhar), ben.fung@mcgill.ca (B.C.M. Fung), Farkhund.Iqbal@zu.ac.ae (F. Iqbal), Dima.Alhadidi@zu.ac.ae (D. Alhadidi), bentahar@ciise.concordia.ca (J. Bentahar).

http://dx.doi.org/10.1016/j.elerap.2016.02.004
1567-4223/

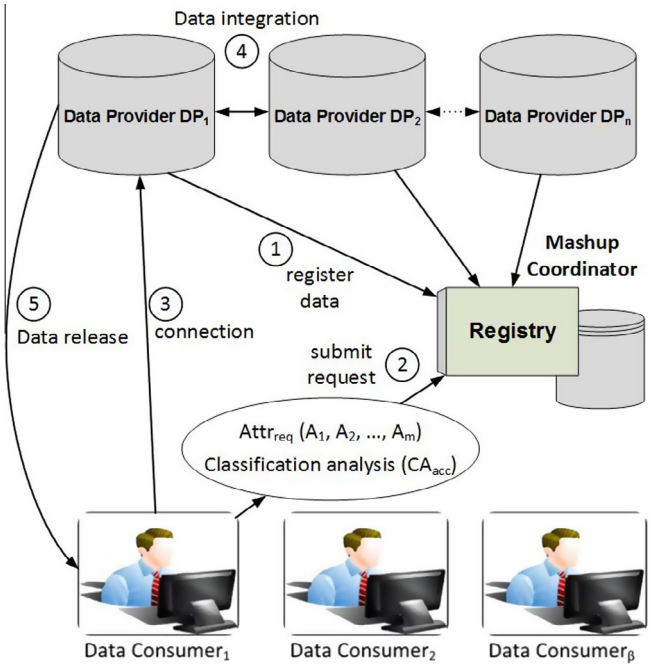


Fig. 1. Privacy-preserving data mashup architecture for trading person-specific information.

the demand of a data consumer by connecting with them. Fourth, the data providers quantify their costs and benefits using joint privacy requirements and integrate their data over the cloud. Finally, the anonymous mashup data is released to the data consumers. The data consumers have the option to perform the data mining operations on the cloud or take the data and perform the data mining operations locally on their own machines.

In the proposed architecture, business enterprises face four major challenges for trading person-specific information: First, extensive research has shown that simply removing explicit identifying information such as name, social security number, birth date, telephone number, and account number is insufficient for privacy protection. Many organizations believe that enforcing regulatory compliance, such as the Gramm–Leach–Bliley Act (GLBA), which protects the privacy and security of individually identifiable financial information, or simply employing common de-identification methods, such as Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor method, which involves removing 18 types of identifiers from health data, is sufficient for privacy protection. Indeed, an individual can be re-identified by matching the *quasi-identifiers QID* with an external data source (Samarati and Sweeney 2001). Second, the data providers collaborate in order to fulfill the demands of a data consumer and to generate more profit by offering better classification utility. In addition, they would avoid sharing information other than the final integrated data because the collaborating data providers could be competitors. Third, a cloud service provider may not be a trusted party. The cloud service provider can be a third-party who offers data integration services over the cloud or one of the data providers. Fourth, the data providers want to ensure that the mashup data can facilitate the queries of data consumers. So, there is a trade-off between data utility and privacy protection in terms of monetary reward. In this paper we propose a model that examines the intangible benefits and potential risks of sharing person-specific data for classification analysis. Our model allows the data providers to quantify the costs and benefits and to generate the monetary value from trading person-specific information.

Our contributions are summarized as follows: the first three challenges, discussed in the previous paragraph, have already been widely studied in the current literature (Arafati et al. 2014, Samarati and Sweeney 2001, Fung et al. 2010, Fung et al. 2012, Aljafer et al. 2014, Mohammed et al. 2014). Here we focus on the fourth challenge that addresses both scientific and business needs for trading person-specific information in the e-market. We develop a business model that identifies the consumers' (e.g., data recipients) requirements and performs the valuation on important parameters associated with revenue and costs for a business. Our business model is suitable for multiple data providers in making decisions where they have the following goals: (a) to find the optimal value on the trade-off between data privacy and data utility and (b) to derive the contribution of each data provider in terms of monetary value. Finally, we show that our proposed approach can effectively achieve both goals by performing extensive experimental evaluations on real-life, person-specific data. The proposed model captures only the relevant factors that are crucial for cost-benefit analysis in our research problem. However, the model provides flexibility for users to include additional factors based on the specific requirements of other scenarios.

The rest of the paper is organized as follows: in Section 2, we review the related work. In Section 3, we explain the challenges faced by business enterprises, followed by the problem definition. In Section 4, we present preliminaries to quantify the data privacy and information utility. In Section 5, we present our model as a privacy-preserving data mashup solution for e-markets. In Section 6, we discuss the limitations of our proposed model. In Section 7, we evaluate our proposed model based on the incorporated factors for multiple data providers by conducting extensive experiments on real-life data. Finally, we provide the conclusion in Section 8.

2. Related work

We summarize the literature of the following related areas: monetizing data privacy for business value generation, trade-off between privacy and utility in data integration, statistical disclosure control methods, and policies and regulations with the perspective of data protection.

2.1. Monetizing data privacy for business value generation

Many organizations are embracing innovations in digital economy to maximize their business value through data. Wixom et al. (2015) conducted seven case studies on companies that monetize data by selling information-based products and/or services. They hypothesize that a company whose business model draws upon six sources, such as data, data architecture, data science, domain leadership, commitment to client action, and process mastery, can bring a competitive advantage for information business value. Wixom and Markus (2015) further identified an approach that they termed "Data Value Assessment" to analyze the costs, benefits, and risks of selling information-based products and services by business enterprises. Li et al. (2014) propose a theoretical framework for private data pricing in an interactive setting. There are three main actors in their proposed architecture: *Data owners* contribute their personal data; a *buyer* submits an aggregate query and pays its price to a *market maker*; and a *market maker*, a trusted party to both, answers *buyer* queries on behalf of *data owners* by adding an appropriate noise (Dwork et al. 2006) in response to the query. The *market maker* compensates the *data owners* whenever they suffer from a privacy loss in response to a *buyer's* query. Riederer et al. (2011) propose a mechanism called "transactional privacy" to control the disclosure of personal information in a

privacy-preserving system. This mechanism allows end users to release personally identifiable information (PII) by giving them the choice to valuate their personal information. Their system leverages prior work on auctions and particularly the exponential mechanism (McSherry and Talwar 2007) to guarantee truthfulness in the bidding process. In this paper, we follow a distributed approach in a non-interactive setting for data mashup of multiple data providers, which is different from our previous work (Khokhar et al. 2014) in which the challenges were to quantify the costs and benefits between privacy and utility from the perspective of a single data custodian. In addition, the business model presented in this paper can derive the contribution of each data provider in terms of monetary value by computing the information gain on the data mashup.

2.2. Trade-off between privacy and utility in data integration

Arafati et al. (2014) propose a cloud-based framework for a privacy-preserving Data-as-a-Service (DaaS) mashup that enables data providers to integrate their person-specific data on demand depending on a consumer's request for data analysis. In their framework, a data consumer can submit a request with a set of attributes, bid price, and classification accuracy. They introduce a greedy algorithm that can dynamically determine the group of DaaS providers offering the lowest price per attribute. They employ a Privacy-Preserving High-Dimensional Data Mashup algorithm (Fung et al. 2012) for secure data integration and to preserve the privacy of mashup data using the LKC-privacy model (Mohammed et al. 2009). Mohammed et al. (2014) propose a differentially private data release algorithm to securely integrate person-specific data from two parties so that integrated data maintains the necessary information to support data utility. They present a scenario for a distributed setup to integrate the vertically partitioned data, where different attributes for the same set of individuals are held by two parties. No additional information is leaked to any party as a result of integrating data. In this paper, the data mashup model employs the approach that was presented in Fung et al. (2012) and Mohammed et al. (2014) for vertically partitioned data to satisfy *LKC*-privacy and ϵ-differential privacy requirements, respectively. There are some other papers (Mohammed et al. 2010, Jurczyk and Xiong 2009) that address the problem of integrating horizontally partitioning data in a distributed manner. This would yield different costs and benefits when quantifying the privacy and utility from the integrated data using horizontal partitioning.

2.3. Statistical disclosure control methods

Many non-perturbative and perturbative anonymization methods, such as global and local recoding (Waal and Willenborg 1998, Takemura 1999), suppression and local suppression (Waal and Willenborg 1998, Little 1993), sampling (Skinner et al. 1994), micro-aggregation (Domingo-Ferrer and Mateo-Sanz 2002), noise addition (Kim 1986), data swapping (Dalenius and Reiss 1982), and post randomization (Kooiman et al. 1997) have been adopted in the past with the goal of providing confidentiality and privacy in publishing person-specific data. According to Gehrke (2010), the statistical methods that are being used for limiting information disclosure do not formally address how much sensitive information an adversary would glean from the published data. Waal and Willenborg (1998) discuss global recoding and local suppression methods to protect person-specific data. In the case of a global recoding method, specific attribute values are mapped to the same generalized value in all records; in the case of local suppression, the specific value of an attribute in a record changes to a 'missing' value, but the attribute values in other records remain unchanged (Waal and Willenborg 1999). Global recoding is the preferable method when there are many unsafe combinations to eliminate in the person-specific data and when one wants to obtain a uniform categorization of attributes (Waal and Willenborg 1998). Truta et al. (2003) use a microaggregation statistical disclosure control technique to measure the trade-off in disclosure risk and information loss on synthetic data based on the criteria specified by the data owner.

2.4. Policies and regulations for data protection

Currie and Seddon (2014) discuss the cross-country approaches to data privacy, regulation, and rules. They did a survey in six countries to collect the views of people on the benefits and risks for adopting cloud computing in a healthcare setup. Generally, healthcare professionals are in favor of adopting cloud computing, but stakeholders involved in the setup have to provide a guarantee for the protection of personal data subject to the regulations enforced in their jurisdictions. They address an important issue of how international governments harmonize an effective legal and regulatory framework for trans-border data flows over the cloud environment. Recent studies (Currie and Seddon 2014, Kuner 2011) show that more than 60 countries in the world have adopted privacy and data protection laws that regulate trans-border data flows. Hu et al. (2012) provide Law-as-a-Service (LaaS) as an emergent technology for cloud service providers to ensure that legal policies are compliant with the laws for users. They provide a conceptual layout of the law-aware semantic policy infrastructure in which a semantic cloud of Trusted Legal Domains (TLDs) are established over the Trusted Virtual Domains (TVDs). Each TLD has a super-peer that provides data integration services for its peers. The super-peer specifies how compliant legal policies are unified and enforced in a domain. Legal policies are composed of OWL-DL ontologies and stratified Datalog rules with negation for a policy's exceptions handling through defeasible reasoning. Description Logic (DL)-based ontologies provide data integration, while Logic Program (LP)-based rules provide data query and protection services.

3. Challenges and problem definition

In this section, we explain the privacy challenges that are realized when integrating data from heterogeneous sources, followed by the problem definition.

3.1. The challenges

The research problem is identified in Data Management Platforms Buyer's Guide (2013), where the challenges are to integrate marketing data from heterogeneous sources and to ensure the privacy of the customers. We generalize the problem as follows: suppose two data providers, DP_1 and DP_2, own raw data tables D_1 and D_2, respectively. Each data provider owns a different set of attributes about the same set of records identified by the common Record IDs, such that DP_1 owns $D_1(Rec.ID, Age, Job)$ and DP_2 owns $D_2(Rec.ID, Sex, Education)$. The data providers want to integrate their data to improve the data utility for classification analysis in order to maximize their profit. The attributes in data tables D_1 and D_2 are classified into four categories for classification analysis: explicit identifier, quasi-identifier (QID), sensitive attribute, and class attribute. An explicit identifier attribute explicitly identifies a person, such as name, social security number (SSN), and account number. A quasi-identifier attribute, such as date of birth, sex, and education, is a set of predictor attributes whose values are used to predict class attribute. A sensitive attribute, such as

disease, salary, and marital status, contains an individual's sensitive information. A class attribute contains the class values for classification analysis. In the following example we discuss the privacy threats that can arise as a result of simply joining the raw data tables of data providers DP_1 and DP_2.

Example 1. Consider the raw data tables of two data providers in Table 1. $Rec.ID, Sensitive$, and $Class$ are shared between data providers DP_1 and DP_2. DP_1 and DP_2 own data tables $D_1(Age, Job)$ and $D_2(Sex, Education)$, respectively. Each record corresponds to the personal information for an individual person. The two data providers want to develop a data mashup service to integrate their data in order to perform classification analysis on the shared *Class* attribute *Loan approval*, which has two values, *Y* and *N*, indicating whether or not the loan is approved.

In a *record linkage* attack (Fung et al. 2010), an adversary attempts to identify the record of a target victim in the released data table. Assume an adversary knows that the target victim is a female cleaner, denoted by $qid = \langle F, Cleaner\rangle$. The group of records matching *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, $D[qid] = \{Rec\#3, 11, 17\}$.

In an *attribute linkage* attack (Fung et al. 2010), an adversary may not be able to accurately identify the record of a target victim but can infer a sensitive value with high confidence if it occurs frequently in the released table. With the prior knowledge *qid* about a target victim, an adversary can identify a group of records $D[qid]$ and can 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*. For example, given $qid = \langle M, Cleaner\rangle$, in Table 1, $D[qid \wedge Divorced] = \{Rec\#10, 18\}$, $D[qid] = \{Rec\#4, 10, 18\}$, and $P(Divorced|qid) = 2/3 = 66.67\%$. □

K-Anonymity (Samarati and Sweeney 2001) and ℓ-diversity (Machanavajjhala et al. 2007) have been proposed to protect against the aforementioned record and attribute linkage attacks in the relational raw data tables. *K-anonymity* prevents record linkage attacks by generalizing the records into equivalence groups of *K* size with respect to some *QID* attributes; however, it could suffer from an attribute linkage attack if the sensitive values are not diversified in an equivalence group. The principle of ℓ-diversity overcomes this problem by requiring every *QID* group to contain at least ℓ well-diversified values for the sensitive attribute. This model presents a stronger notion of privacy to protect from *homogeneity attacks* and *background knowledge attacks*. Mohammed et al. (2009) propose a *LKC-privacy* model in which they assume that the adversary's background knowledge is bounded by *at most 'L' QID* attributes. This model provides better data utility in comparison to *K*-anonymity on high-dimensional data. Dwork et al. (2006) propose a *differential privacy* model that ensures the addition or removal of a single database record does not significantly affect the outcome of any computation over a database. It provides strong privacy guarantees to an individual independent of an adversary's background knowledge and computational power.

The aforementioned privacy models are discussed from the perspective of a single data custodian. Another challenge is related to the data mashup of multiple data custodians when consumer data requests cannot be fulfilled by a single data provider. The data mashup is a process over the cloud infrastructure that enables multiple data providers to integrate their data in order to fulfill the demands of data consumers. The cloud service provider may be one of the data providers or a third party, but the mashup scenario for the integration of data from multiple data custodians should not reveal person-specific information of the customers to unauthorized parties. The trust of a customer in an exchange of services with one data provider by sharing person-specific information does not necessarily extend trust to the other data providers. So, there is a need to avoid disclosure of sensitive information during the data mashup process and in the final release of mashup data. There are some known approaches that do not ensure privacy of an individual, such as (1) *mashup-then-generalize* and (2) *generalize-then-mashup*. The first approach integrates the raw data tables from two data providers and then generalizes using single table anonymization methods (Fung et al. 2007, LeFevre et al. 2006). This approach fails to preserve privacy because once the mashup coordinator or any other third party holds the integrated raw data it will instantly discover all the private information of both data providers. The second approach generalizes the data providers' tables individually using single-table anonymization methods, then integrates the generalized tables. This approach seems to preserve privacy locally at an individual data provider's end, but it does not guarantee the privacy for a quasi-identifier that spans multiple data providers' tables.

To address the above-mentioned privacy issues that arise in the data mashup when data is owned by multiple providers, Fung et al. (2012) propose an extended version of the *LKC*-privacy model to

Table 1
Raw data table of data providers.

Rec.ID	Data provider DP_1		Data provider DP_2		Sensitive	Class
	Age	Job	Sex	Education	Marital-status	Loan approval
1	39	Painter	F	12th	Divorced	N
2	43	Doctor	M	Doctorate	Never-married	Y
3	37	Cleaner	F	12th	Divorced	Y
4	56	Cleaner	M	10th	Never-married	N
5	64	Welder	M	8th	Married-civ-spouse	Y
6	49	Doctor	F	Doctorate	Married-civ-spouse	Y
7	33	Lawyer	F	Masters	Never-married	Y
8	41	Lawyer	F	Doctorate	Married-civ-spouse	N
9	32	Painter	F	12th	Divorced	N
10	52	Cleaner	M	Bachelors	Divorced	Y
11	39	Cleaner	F	11th	Divorced	Y
12	61	Lawyer	M	Doctorate	Married-civ-spouse	Y
13	24	Technician	M	11th	Married-civ-spouse	N
14	44	Technician	F	Bachelors	Divorced	N
15	34	Lawyer	M	Masters	Never-married	Y
16	27	Painter	M	11th	Divorced	N
17	35	Cleaner	F	10th	Divorced	Y
18	41	Cleaner	M	11th	Divorced	Y
19	63	Welder	M	8th	Married-civ-spouse	N

apply to a multiple data providers scenario. The *LKC*-privacy model is suitable to apply on high-dimensional data, as would normally be the case when integrating data from multiple data providers. This overcomes the problem of high-dimensionality when using *K*-anonymity. *K*-Anonymity (Samarati and Sweeney 2001) is known to be a special case of *LKC*-privacy with adversary knowledge $L = |QID|$ and confidence $C = 100\%$, where $|QID|$ is the number of quasi-identifying attributes in the data table (Mohammed et al. 2009). Mohammed et al. (2014) have propose a differentially private data release algorithm for multiple data providers in a distributed setup. Our model employs the approaches presented in Fung et al. (2012) and Mohammed et al. (2014) for data mashup of multiple data providers and sets the joint privacy requirements of contributing data providers in order to ensure that no extra information is leaked to any provider as a result of data integration.

3.2. Problem definition

Suppose data providers $DP_1, \ldots, DP_n$ own data tables $D_1, \ldots, D_n$, respectively. They want to generate an integrated anonymous dataset D' that fulfills the demands of data consumers and generates more profit in terms of monetary value for the data providers. Our proposed model enables the collaboration between data providers to set their joint privacy requirement for data mashup. It can also benefit data providers to quantify their costs and benefits in trading person-specific information and in determining the contribution of each data provider. Formally, the research problem is defined as follows.

Definition 3.1 (*Problem Definition*). Given multiple person-specific raw data tables $D_1, \ldots, D_n$ from data providers $DP_1, \ldots, DP_n$ and a set of requested attributes $Attr_{req}$ for classification analysis from a data consumer, the research problem is to develop a business model that performs the valuation on cost factors to find the optimal value from the anonymized integrated data table $D\prime$ under the joint privacy requirements of the data providers and to derive the contribution of each data provider $DP_1, \ldots, DP_n$ in terms of monetary value. □

4. Preliminaries

In this section, we first present some measures to quantify the data privacy and information utility, followed by an overview of our employed privacy-preserving data mashup algorithms.

4.1. Quantifying privacy

Consider a raw data table $D(Rec.ID, A_1, \ldots, A_m, Sens, Class)$ of two data providers DP_1 and DP_2 as shown in Table 1. Both data providers want to release an integrated anonymized dataset D' to the data consumer for joint classification analysis. *Rec.ID* is shared between the data providers' tables and is used to uniquely identify a record; it is used to join the data tables. Each A_i is either a categorical or a numerical attribute. *Sens*, *Class* are also shared between data providers DP_1 and DP_2 representing a sensitive attribute and a class attribute, respectively. Each data provider owns a different set of attributes on the same set of Record IDs, such that DP_1 owns D_1 and DP_2 owns D_2. A record in D has the form $\langle v_1, v_2, \ldots, v_m, s, cls\rangle$, where v_i is a value in A_i, s is a sensitive value in *Sens*, and *cls* is a class value in *Class*. In Section 3.1 we discussed privacy threats that arise by simply joining the raw data tables of DP_1 and DP_2.

Privacy models. In this subsection, we present the formal definitions of four widely adopted models from the perspective of a single data custodian, namely *K-anonymity*, *ℓ-diversity*, *LKC-privacy*, and *ϵ-differential privacy*.

Definition 4.1 (*(K-anonymity) Samarati and Sweeney 2001*). Let $D(A_1, \ldots, A_m)$ be a data table and *QID* be its quasi-identifier. D satisfies *K*-anonymity if, and only if, each group of *QID* appears in at least K records in D. □

Definition 4.2 (*(Entropy ℓ-diversity) Machanavajjhala et al. 2007*). A table is entropy ℓ-diverse if every *QID* group satisfies $-\sum_{s\in Sens} P(qid, s)log(P(qid, s)) \geqslant log(\ell)$, where *Sens* is a sensitive attribute $P(QID, s)$ is the percentage of records in a *QID* group containing the sensitive value s. □

Definition 4.3 (*(LKC-privacy) Mohammed et al. 2009*). Let L be the maximum number of *QID* attributes acquired by an adversary as prior knowledge about 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 representing the anonymity threshold, *and*
2. for any $s \in S, P(s|qid) \leqslant C$, where $0 < C \leqslant 1$ is a real number representing the confidence threshold. □

Definition 4.4 (*(ϵ-differential privacy) Dwork et al. 2006*). A sanitization mechanism M_{rnd} provides *ϵ-differential privacy*, if for any two datasets D_1 and D_2 that differ on at most one record (i.e., symmetric difference $|D_1 \triangle D_2| \leqslant 1$), and for any possible sanitized datasets $\widehat{D}$,

$$\Pr[M_{rnd}(D_1) = \widehat{D}] \leqslant e^{\epsilon} \times \Pr[M_{rnd}(D_2) = \widehat{D}],$$

where the probabilities are taken over the randomness of M_{rnd}. □

4.2. Quantifying utility

The information utility is measured depending on the requirements for data analysis. In this paper we present classification analysis as a utility measure on the consumer's specified service request and analysis task.

Score for 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 $InfoGain(v)$:

$$Score(v) = InfoGain(v) \tag{1}$$

Let D_x denote the set of records in the data table D generalized to the value x. Let $freq(D_x, cls)$ denote the number of records in D_x having the class *cls*. Note that $|D_v| = \sum_c |D_c|$, where $c \in child(v)$. The information gain $InfoGain(v)$ and entropy $H(D_x)$ are defined as follows:

$$InfoGain(v) = H(D_v) - \sum_c \frac{|D_c|}{|D_v|} H(D_c) \tag{2}$$

$$H(D_x) = -\sum_{cls} \frac{freq(D_x, cls)}{|D_x|} \times log_2 \frac{freq(D_x, cls)}{|D_x|} \tag{3}$$

where $H(D_x)$ measures the *entropy* of classes for the records in D_x (Quinlan 1993), and $InfoGain(v)$ measures the reduction of the entropy by specializing v into $c \in child(v)$. A smaller entropy $H(D_x)$ implies a higher purity of the partition with respect to the class values.

We build a classifier on 2/3 of the records of the anonymized dataset as the training set and measure the *Classification Error* (*CE*) on 1/3 of the records of the anonymized records as the testing set to determine the impact of anonymization on data utility for classification analysis. *Classification Accuracy* (*CA*) is calculated by $1-(CE)$. In this paper, we use the well-known C4.5 classifier (Quinlan 1993) for classification analysis.

4.3. Data mashup algorithms

In this section, we discuss state-of-the-art anonymization algorithms for data mashup in a multiple data-providers scenario: *Top-Down Specialization*(*TDS*) (Fung et al. 2012 and *Diff*erentially private anonymization based on *Gen*eralization (*DiffGen*) (Mohammed et al. 2014).

4.3.1. Top-down specialization algorithm for multiple data providers

Algorithm 1 presents an overview of the *Top-Down Specialization* (*TDS*) algorithm to integrate data in a scenario of multiple data providers (Fung et al. 2012).

Consider multiple data providers $DP_1,\ldots,DP_n$, who own private data tables $D_1,\ldots,D_n$ having a common record identifier *Rec.ID*. Initially, every data provider generalizes all of its own attribute values to the topmost value according to the taxonomy trees, as illustrated in Fig. 2, and maintains a mark $Mark_i$ that contains the topmost value for each attribute A_i in *QID*. A *taxonomy tree* is specified for each categorical attribute in *QID*. A leaf node represents a precise value and a parent node represents a more general value. For continuous attributes in *QID*, taxonomy trees 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 (Quinlan 1993). The $\cup Mark_i$ on all attributes represents a generalized table D, denoted by D_g. $\cup Mark_i$ also contains the set of candidates for specialization. A specialization $v \to child(v)$ is *valid*, written as $IsValid(v)$, if the generalized table D_g still satisfies the privacy requirements stated in Definitions 4.1 and 4.3 *after* the specialization on v. At each iteration, the TDS multiple data providers mashup (TDSmdpm) algorithm identifies the winner candidate by communicating the *Score* with all the participating data providers (Lines 4–5). The valid candidate that has the highest *Score*, among all the candidates, performs the winner specialization (Lines 7–11) and updates the *Score* and the *IsValid* status of the new and existing candidates in the mark (Line 14). TDSmdpm terminates when there are no valid candidates in the mark.

Suppose that winner candidate w is local to data provider DP_1 that performs $w \to child(w)$ on its copy of $\cup Mark_i$ and D_g. This means specializing each record $r \in D_g$ containing w into $r'_1,\ldots,r'_z$; the child values are in $child(w)$. Similarly, all the other data providers $DP_2,\ldots,DP_n$ update their $\cup Mark_i$ and D_g and partition $D_2[r]$ into $D_2[r'_1],\ldots,D_2[r'_z]$ $\ldots D_n[r]$ into $D_n[r'_1],\ldots,D_n[r'_z]$. Since all the other participating data providers do not have w, DP_1 needs to instruct $DP_2,\ldots,DP_n$ on how to partition their records in terms of *Rec.ID*s.

Algorithm 1. TDS multiple providers data mashup (Fung et al. 2012).

1: Initialize every record values in D to the topmost generalized values D_g.
2: Initialize $\cup Mark_i$ to include only topmost values and update $IsValid(v)$ for every $v \in \cup Mark_i$;
3: **while** $\exists v \in \cup Mark_i$ s.t. $IsValid(v)$ **do**
4: Find the local winner candidate x of DP_i that has the highest $Score(x)$;
5: Communicate $Score(x)$ with all the other participating data providers to determine the global winner w;

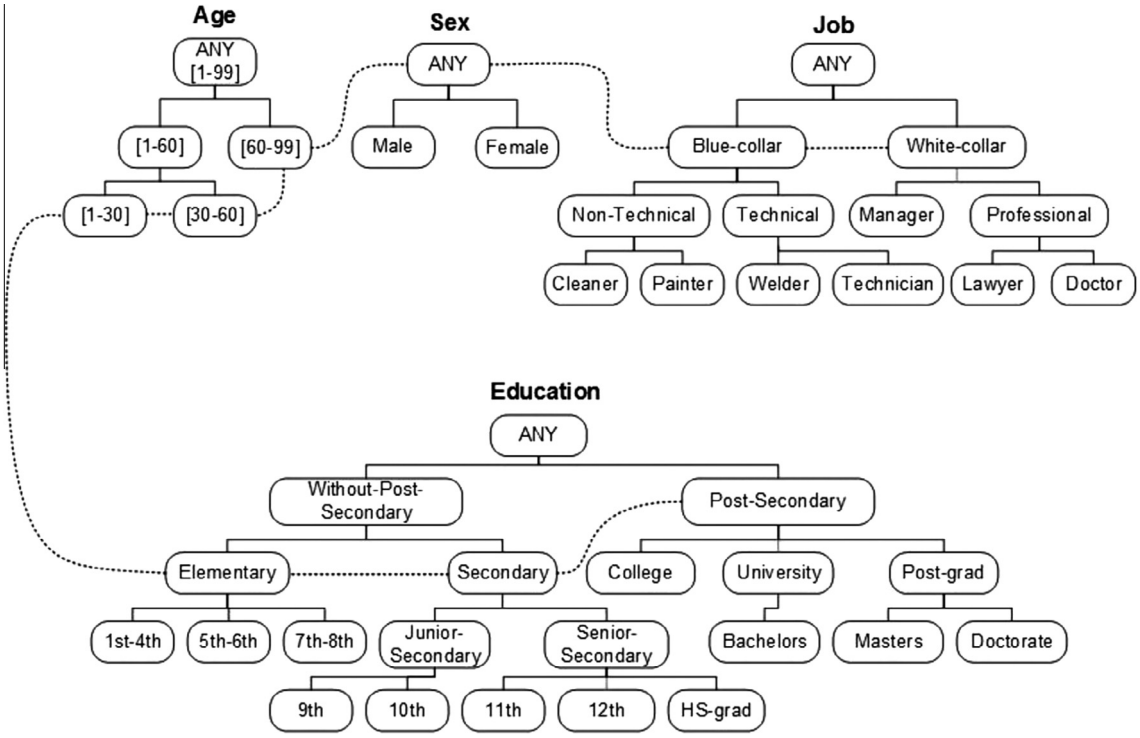


Fig. 2. Taxonomy trees.

6: **if** the winner w is local **then**
7: Specialize w on D_g;
8: Instruct all the other data providers to specialize w;
9: **else**
10: Wait for the instruction from the winner data provider;
11: Specialize w on D_g using the instruction;
12: **end if**
13: Replace w with $child(w)$ in the local copy of $\cup Mark_i$;
14: Update $Score(x)$ and $IsValid(x)$ for every candidate $x \in \cup Mark_i$;
15: **end while**
16: **return** D_g and $\cup Mark_i$;

4.3.2. DiffGen anonymization algorithm for multiple data providers

Algorithm 2 provides an extension of the two-party *Diff*erentially private anonymization based on *Gen*eralization (Mohammed et al. 2014) to differentially integrate multiple private data tables $D_1, \ldots, D_n$ sharing a common identifier *Rec.ID*, which is owned by data providers $DP_1, \ldots, DP_n$ for classification analysis. However, the distributed exponential mechanism is limited to two parties. *DiffGen* (Mohammed et al. 2011) is an extension of the TDS algorithm to achieve ϵ-differential privacy. The two major extensions over the TDS algorithm include: (1) *DiffGen* selects the *Best* specialization based on the exponential mechanism, and (2) *DiffGen* perturbs the generalized contingency table by adding the Laplacian noise to the *qid* counts. The Laplacian noise is calibrated based on the *sensitivity* of a utility function, which quantifies the maximal impact of adding or deleting a single record on a function. This algorithm provides secure data integration of two parties under the definition of the semi-honest adversary model.

Initially, all values in the predictor attributes $\mathcal{A}^{pr}$ (i.e., attributes used to predict the class attribute) of each data provider are generalized to the topmost value in their taxonomy trees (Line 1), as illustrated in Fig. 2, and $Mark_i$ contains the topmost value for each attribute A_i^{pr} (Line 2). The predictor attribute $\mathcal{A}^{pr}$ can be either categorical or numerical, but the class attribute is required to be categorical. The value of a categorical attribute is denoted by v_c, whereas the value of a numerical attribute is denoted by v_d. Each data provider keeps a copy of the $\cup Mark_i$ and a generalized data table D_g. The algorithm first determines the split points for all numerical candidates $v_d \in \cup Mark_i$ by using the exponential mechanism (Line 4), then computes the scores for all candidates $v \in \cup Mark_i$ (Line 5). At each iteration the algorithm uses the secure distributed exponential mechanism (DistExp) as presented in Mohammed et al. (2014) (readers may refer to the details of DistExp algorithm) to select a winner candidate $w \in \cup Mark_i$ for specialization (Line 7). Different utility functions (e.g., information gain) can be used to calculate the score. If the winner candidate w is local to DP_1, DP_1 specializes w on D_g by splitting its records into child partitions, updates its local copy of $\cup Mark_i$, and instructs all the other participating data providers to specialize and update their local copy of $\cup Mark_i$ (Lines 8–11). DP_1 further calculates the scores of the new candidates as a result of the specialization (Line 13). If the winner w is not one of DP_1's candidates, DP_1 waits for instructions from the other winner data provider to specialize w and to update its local copy of $\cup Mark_i$ (Lines 15 and 16). This process is iterated until the specified number of the specializations h is reached. Finally, the algorithm perturbs the output by adding the noisy count at each leaf node (Line 19) using the Laplace mechanism.

Algorithm 2. DiffGen for multiple data providers (Mohammed et al. 2014).

1: Initialize D_g with one record containing topmost generalized values;
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 $v_d \in \cup Mark_i$ with probability $\propto \exp(\frac{\epsilon'}{2\Delta u} u(D, v_d))$;
5: Compute the *Score* for $\forall v \in \cup Mark_i$;
6: **for** *iter* = 1 to h **do**
7: Determine the winner candidate w by using the DistExp Algorithm (Mohammed et al. 2014);
8: **if** w is local **then**
9: Specialize w on D_g;
10: Replace w with $child(w)$ in the local copy of $\cup Mark_i$;
11: Instruct all the other participating data providers to specialize and update $\cup Mark_i$;
12: Determine the split value for each new $v_d \in \cup Mark_i$ with probability $\propto \exp(\frac{\epsilon'}{2\Delta u} u(D, v_d))$;
13: Compute the *Score* for each new $v \in \cup Mark_i$;
14: **else**
15: Wait for the instruction from the winner data provider;
16: Specialize w and update $\cup Mark_i$ using the instruction;
17: **end if**
18: **end for**
19: **return** each leaf node with count $(C + \mathrm{Lap}(2/\epsilon))$

5. Proposed solution

In this section, we present a privacy-preserving solution for the business enterprises that seek to adopt an appropriate approach to manage the challenges of the e-market for trading person-specific information. Section 3.1 discusses the challenges of integrating data from multiple data providers, where each data provider owns a different set of attributes. We assume that every data provider intends to maximize the data utility, which in turn maximizes their profits, without violating the mutually agreed-upon privacy requirement. In this paper, we focus on analyzing the problem of preventing the disclosure of sensitive information during data mashup and on the final release of mashup data. We employ anonymization algorithms, namely *Top-Down Specialization*(*TDS*) (Fung et al. 2012) and *Diff*erentially private anonymization based on *Gen*eralization (*DiffGen*) (Mohammed et al. 2014), for relational data mashup from multiple data providers. Our model quantifies the costs and benefits of privacy-preserving data publishing for the contributing data providers in terms of monetary value.

In our model, customers, data providers, and data consumers are the main stakeholders. For these stakeholders we identify the most relevant factors, as illustrated in Fig. 3, to reflect the customers' requirements on data privacy, the data consumers' requirements on data utility, and the data providers' requirements on properly balancing privacy and utility with the goal of releasing the integrated data for profit. One of the limitations of our model is the lack of a standard method to monetize the value of personal data, especially when several parties are involved in collecting person-specific information from the same population. Currently, many companies actively collect personal information by providing monetary rewards to their customers or respondents. There is no standard price for a specific piece of personal information,

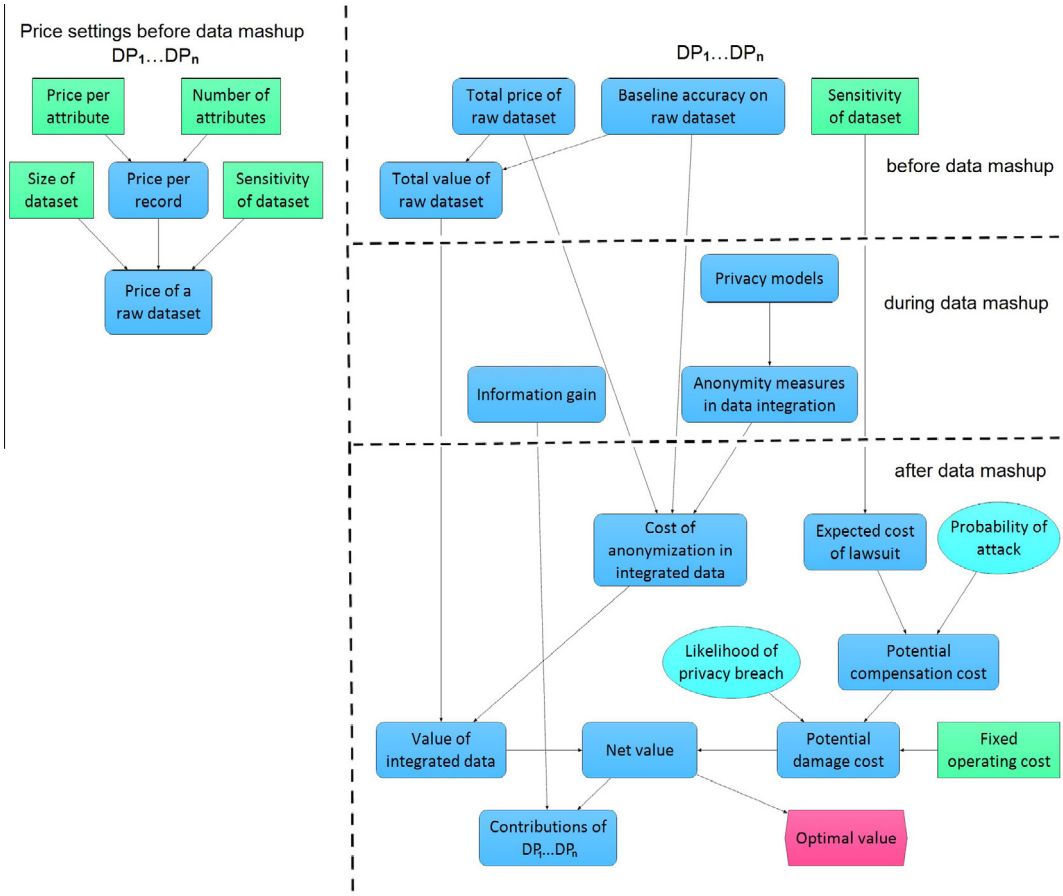


Fig. 3. Business model for privacy-preserving data mashup.

but some market estimates are available in OECD (2013) and Gates and Matthews (2014). It is also pointed out in OECD (2013) that there is no commonly accepted methodology for estimating the monetary value of personal data. Person-specific data contains sensitive and non-sensitive information. It is the utmost responsibility of data providers to take preventive measures when dealing with the sensitive information of individuals. Indeed, sensitive data is qualitative by nature. We set the sensitivity level of a dataset on the scale of 1–5 to indicate its significance for privacy protection. Another limitation of our model is the inconsistency of the expected cost of a lawsuit. The expected cost of a lawsuit depends on the sensitivity of data and can be estimated from the historical trends of privacy breach incidents. An individual may file a lawsuit against a data provider when his or her sensitive information is disclosed to a third party or made public without his or her consent. Although there is no fixed cost related to privacy breach cases, regulatory agencies such as the *Federal Trade Commission* (*FTC*) and the *Securities and Exchange Commission* (*SEC*) have imposed monetary fines and penalties subject to the nature of privacy breaches (Romanosky and Acquisti 2014). According to the revised *HITECH* penalty scheme (Department of Health and Human Services 2013), the penalty for a violation due to *reasonable cause* and not to *willful neglect* is between $1000 and $50,000 for each violation.

Section 5.1 presents the business model for privacy-preserving data mashup. Section 5.2 discusses the key business factors for determining the value of integrated data and the factors that contribute to the potential damage cost. Section 5.3 discusses the implicit and explicit risk measures for privacy attacks.

5.1. Business model for privacy-preserving data mashup

Our proposed privacy-preserving data mashup business model allows the collaboration of multiple data providers to mashup their data over the cloud and to quantify the costs and benefits of releasing anonymized person-specific information in terms of *monetary value*. Fig. 3 provides an overview of the proposed model; key factors are organized into three phases: before data mashup, during data mashup, and after data mashup. The left pane of the model depicts the decision factors held by each data provider, who registers its available data before the data mashup. For example, *Price per attribute*, *Number of attributes*, and *Size of dataset* are the decision factors that depend on the market value and consumer demand. Data providers can set their own decision factors. These decision factors contribute to finding the *Price of a raw dataset* for every data provider. In the presented model, nodes represent different types of factors, and arrows indicate the influences or dependencies between different factors. For example, an arrow pointing from the *Baseline accuracy on raw dataset* to the *Total value of raw dataset* in the model indicates the influence of the *Baseline accuracy on raw dataset* on the *Total value of raw dataset*.

The objective of maximizing the profit can be achieved by balancing the two important factors: maximizing the *Value of integrated data*, and minimizing the *Potential damage cost*. The *Value of integrated data* depends upon the *Total value of raw dataset* and *Cost of anonymization in integrated data*. The *Cost of anonymization in integrated data* is computed on the data integration of contributing data providers with respect to the classification analysis (data mining) task. Each data provider can compare his or her benefits and costs *before* and *after* participation in the mashup process.

For classification analysis, a data provider can estimate the classification analysis on the anonymized data of his or her own data, and then estimate the classification analysis on the integrated data. On one side, trading person-specific information has a high value in the e-market, but on the other side data providers who collaborate in sharing person-specific information need to be cautious of the risk of privacy breaches and cost of potential damages when integrating data. Our business model allows the participating data providers to: (1) set up their joint privacy requirements during data mashup by choosing the privacy model along with the anonymization algorithm and privacy parameters, and (2) analyze the impact of anonymization on information utility for classification requirement in terms of monetary value after data mashup. The aforementioned business factors can help the data providers in defining the overall objective of maximizing *Net value*. Further, in the data mashup process the contribution of each data provider is derived from the achieved *Net value* by fairly computing the information gain on the anonymized data. Accordingly, the data provider whose data provides a larger information gain for classification can get the larger share of the monetary net value.

The companies that face similar challenges, and whose business models are primarily based on sharing person-specific information, can be our potential audiences. There are quite a few companies for which our research problem can be generalized. Some of them are Acxiom, AdAdvisor, AnalyticsIQ, BlueKai, comScore, Datacratic, Dataline, eXelate, Lotame, etc. that aggregate information from various sources for a variety of purposes (Data Partners 2014).

5.2. Key factors for business model

The selection of key factors and their valuations is crucial in developing the cost-benefit business model. We learn and identify key factors from different sources (OECD 2013, Hirshleifer et al. 2005). These factors are broadly classified into two categories: factors that contribute to estimate the *Value of integrated data* and other factors associated with the *Potential damage cost*. We further divide the factors by organizing a set of factors that are involved before the start of the data mashup process, during the data mashup process, and after the data mashup process.

5.2.1. Before data mashup

In this subsection we discuss the factors that are considered as essential prior to performing the cost-benefit analysis. The data providers can set up the market prices on their available data (Gates and Matthews 2014) (e.g., set of attributes) before the data mashup process. Let us assume there are n data providers $DP_1, \ldots, DP_n$, and DP_i denotes the identity of the data provider.

5.2.1.1. Price per attribute. The price per attribute $Price_{attr_i}$ of a data provider DP_i represents the cost of collecting one successful questionnaire for an attribute. Each DP_i can set a price on their data attributes based upon prior knowledge about market pricing offered by other competing data providers (OECD 2013). There is no definite price for the personal identifying attributes, such as name, address, email, birthdate, phone number, etc. But the values can be inferred from cases where personal identity is being sold at a low pricing, as highlighted in the current literature (Gates and Matthews 2014). In our empirical study, we assume the monetary value for $Price_{attr_i}$.

5.2.1.2. Number of attributes. The attribute count $Count_{attr_i}$ of a data provider DP_i represents the number of attributes in a single record. Each DP_i owns a different set of attributes.

5.2.1.3. Price per record. The price per record $Price_{rec_i}$ of a data provider DP_i represents the unit price of a record. Naturally, it is the product of the price per attribute $Price_{attr_i}$ and the attribute count $Count_{attr_i}$ in a single record. That is,

$$Price_{rec_i} = Price_{attr_i} \times Count_{attr_i} \tag{4}$$

The price of a raw dataset of the data provider DP_i increases as the unit *price per record* increases.

5.2.1.4. Size of dataset. The size of a dataset $Size_{ds_i}$ represents the total number of records in the DP_i dataset. $Size_{ds_i}$ increases as the number of records in the dataset increases. Each record has an associated price. As the number of records increases, the overall pricing of a raw dataset also increases.

5.2.1.5. Sensitivity of dataset. The sensitivity of a dataset Sen_{ds_i} indicates that a dataset contains sensitive or personally significant information. It is a given qualitative factor and every data provider should consider this factor for privacy risk assessment. The sensitivity level signifies the importance of data privacy for each data provider DP_i. Intuitively, a higher sensitivity level implies a higher price of a raw dataset and a higher impact on the lawsuit and compensation cost.

5.2.1.6. Price of a raw dataset. The price of a raw dataset $Price_{rd_i}$ represents the data provider DP_i's selling price of a raw dataset in the e-market. It is the product of the sensitivity of the dataset Sen_{ds_i}, the size of the dataset $Size_{ds_i}$, and the price per record $Price_{rec_i}$, which is formulated as follows.

$$Price_{rd_i} = Sen_{ds_i} \times Size_{ds_i} \times Price_{rec_i} \tag{5}$$

5.2.1.7. Total price of raw dataset. The total price of the raw dataset $TPrice_{rd}$ is the sum of the prices of all contributing data providers' raw datasets, which is formulated as follows.

$$TPrice_{rd} = \sum_{i=1}^{n} Price_{rd_i} \tag{6}$$

5.2.1.8. Baseline accuracy on raw dataset. Baseline accuracy on raw dataset BA is determined by considering the classification task as the utility function to evaluate the information utility on the raw datasets of contributing data providers. Data providers can compute the baseline accuracy (BA) using the secure multiple party classifier (Du and Zhan 2002) without sharing their raw data.

5.2.1.9. Total value of raw dataset. The total value of the raw dataset $TValue_{rd}$ represents the monetary value of a raw dataset that the data providers derive from the information utility. It is the product of the total price of the raw dataset $TPrice_{rd}$ and the baseline accuracy of the raw dataset BA, which is formulated as follows.

$$TValue_{rd} = TPrice_{rd} \times BA \tag{7}$$

5.2.2. During data mashup

In this subsection, we discuss the factors that are involved during the data mashup process.

5.2.2.1. Privacy models. The participating data providers DP_n can mutually choose the privacy model (refer to Section 4.1 for details), such as *K-anonymity*, *LKC-privacy*, and *ϵ-differential privacy*, prior to integrating their data.

5.2.2.2. Anonymity measures in data integration. The participating data providers DP_n can jointly set up the data mashup

anonymization algorithm (refer to Section 4.3 for details), such as multi-party *TDS* (Algorithm 1) and *DiffGen* (Algorithm 2), along with the anonymity thresholds, such as K, L, C, for *K-anonymity* and *LKC-privacy* models and ϵ, and h for a ϵ-*differential privacy* model.

5.2.2.3. Information gain. The information gain is employed to determine the usefulness of classification. It computes the reduction of entropy by specializing node v into $c \in child(v)$ as discussed in Section 4.2. Each data provider owns a different set of attributes, but for the same set of records. Each data provider DP_i computes the information gain or $Score(x)$ locally for each candidate and picks the candidate x with the highest value of $Score(x)$. Then each data provider DP_i communicates $Score(x)$ with the n collaborating data providers for determining the global winner w. The winner w data provider performs specialization $w \in child(w)$ on its own copy locally. The winner w data provider then instructs other n collaborating data providers how to perform specialization (further explanation of this process can be seen in Section 4.3). This process is iterative and it runs until no candidate is left in the mark. The information gain $Score(x)$ of winner candidate w data provider accumulates under the relevant winner w data provider.

5.2.3. After data mashup

In this subsection we discuss the factors that are applied after the data mashup process. These factors help in determining the optimal value and the contribution of each data provider.

5.2.3.1. Cost of anonymization in integrated data. To determine the cost of anonymization in integrated data $Cost_{intgdata}$, we make use of the difference between baseline accuracy (BA) and classification accuracy (CA). BA measures the accuracy of classification analysis on raw data while CA measures the accuracy on anonymized integrated data. Therefore, $Cost_{intgdata}$ becomes:

$$Cost_{intgdata} = TPrice_{rd} \times (BA - CA) \tag{8}$$

5.2.3.2. Value of integrated data. The value of integrated data $Val_{intgdata}$ is the difference between the total value of raw dataset $TValue_{rd}$ and the cost of anonymization in integrated data $Cost_{intgdata}$. It is the benefit that the data providers can earn from the information utility of classification analysis by trading their integrated data. Formally, $Val_{intgdata}$ is defined as:

$$Val_{intgdata} = TValue_{rd} - Cost_{intgdata} \tag{9}$$

5.2.3.3. Probability of attack. The probability of attack $Prob_{atk}$ is employed to determine the implicit weaknesses in privacy protection methods. The data providers can carefully consider and plan concerning an adversary's attempt to assess the probability of occurrence of a sensitive attribute value in the anonymized integrated dataset using precision and recall measures (refer to Section 5.3 for details). The probability of occurrence changes with respect to the chosen privacy model and its level of privacy protection. $Prob_{atk}$ is calculated using F-measure on the sensitive attribute value Sen_{val}. F-measure is a weighted harmonic mean of precision and recall. Formally, $Prob_{atk}$ is defined as:

$$Prob_{atk} = \frac{2 \times (Precision\ on\ Sen_{val} \times Recall\ on\ Sen_{val})}{Precision\ on\ Sen_{val} + Recall\ on\ Sen_{val}} \tag{10}$$

5.2.3.4. Expected cost of lawsuit. The expected cost of lawsuit $Ecost_{lwst}$ is enforced subject to the nature of a privacy breach and the sensitivity of data. It increases as the level of data sensitivity increases. $Ecost_{lwst}$ enables business enterprises in predicting the potential cost of privacy breach incidents. The monetary costs can be estimated based on the historical trends of privacy breach incidents. The Federal Trade Commission Act (FTCA), Gramm–Leach–Bliley Act (GLBA), Fair Credit Reporting Act (FCRA), and Personal Data Privacy and Security Act regulate the collection, use, and protection of personal information and impose monetary fines and penalties subject to the nature of the data breach A Legal Guide to Privacy and Data Security 2014, Personal Data Privacy and Security Act 2011.

The lawsuit cost is not fixed and it varies with the applied anonymity measures on data mashup. For instance, an adversary may exploit the inherent weakness of the privacy protection method to infer sensitive information about a victim by using the precision and recall measures in the equation of the probability of attack.

5.2.3.5. Likelihood of privacy breach. The likelihood of a privacy breach L_{pb} measures an adversary's prowess in inferring the victim's sensitive value. This inference is measured using an attack model (refer to the Section 5.3 for details) by exploiting the background knowledge about a victim. We assume that the victim's record is in the integrated published dataset and the adversary knows the victim's QID. Formally, L_{pb} is defined as:

$$L_{pb} = \frac{Total\ records\ count\ on\ Sen_{val}}{Total\ records\ count\ on\ class\ label\ Sen_{attr}} \tag{11}$$

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

5.2.3.6. Potential compensation cost. The potential compensation cost PCC is a factor that can help data providers to determine the approximate cost of compensation prior to sharing the anonymized integrated dataset. It is impacted by the enforcement of privacy policies and privacy protection methods. The potential compensation cost would vary in the presence of a privacy attack and the associated risk of sensitive information disclosure. In general, more stringent privacy parameters impede the probability of a privacy attack. It is our rational hypothesis that privacy attacks would have an exponential impact on the compensation cost due to the substantial increase in the cost of litigation processes (Review of the Personal Data (Privacy) Ordinance 2009). There is no fixed monetary value for compensation cost in Review of the Personal Data (Privacy) Ordinance (2009), but in the e-market a customer who suffers monetary loss due to the disclosure of his or her sensitive information may claim against data providers (e.g., business enterprises) for compensation. Formally, PCC is defined as:

$$PCC = \exp(Prob_{atk}) \times Ecost_{lwst} \tag{12}$$

5.2.3.7. Fixed operating cost. The fixed operating cost F_{OpCost} indicates the fixed monthly cost that business enterprises would have to pay when adopting cloud-services for data integration. Business enterprises would gain more benefits with the adoption of cloud-services comparative to expenditures incurred on hardware and software purchase, setup and installation, licensing and upgrades, maintenance and support, power and utility, and allocation of physical space. F_{OpCost} is a quantitative factor, and its value is independent of the employed anonymity measures in the process of data mashup. It remains the same regardless of the changes in value of integrated data $Val_{intgdata}$.

5.2.3.8. Potential damage cost. The potential damage cost PDC indicates the cost that the data providers would suffer from data privacy breaches. An adversary may attempt to infer sensitive information about a victim from the anonymized integrated

dataset by using an explicit form of a privacy attack as discussed in Section 5.2.3.5. In case of a privacy breach, business enterprises (e.g., data providers) would face substantial costs because of the mandatory notification of data breach, handling of regulatory investigations, hiring of external auditors, facing class action litigation, and loss of business goodwill and customer relationships (Bevitt et al. 2012). As suggested by existing studies (Backman and Levin 2011, Acquisti et al. 2006, Gwebu et al. 2014), data breaches negatively impact business profitability. We postulate that the likelihood of a privacy breach would have an exponential impact on the potential damage cost because a plaintiff (e.g., customer) seeks redress for alleged harms such as actual monetary loss from the identity theft, emotional distress, sexual harassment, discrimination, or possible future losses (Romanosky et al. 2014). *PDC* is determined by the likelihood of a privacy breach L_{pb}, the potential compensation cost *PCC*, and the fixed operating cost F_{OpCost}. Formally, *PDC* is defined as:

$$PDC = \exp(L_{pb}) \times PCC + F_{OpCost} \tag{13}$$

5.2.3.9. Net value. The net value *NV* demonstrates due diligence in evaluating the key business factors on the trade-off between privacy and information utility. It is employed to quantify the difference between the value of integrated data and the potential damage cost on the applied anonymity measures in the mashup process. The net value changes with respect to the chosen privacy model along with the anonymization algorithm and privacy parameters. Formally, *NV* is calculated as follows.

$$NV = Val_{intgdata} - PDC \tag{14}$$

5.2.3.10. Optimal value. The optimal value Opt_{val} is achieved at the maximum of the net value *NV*. It changes with the variations of price settings and joint privacy requirements of data providers. *NV* is realized by the difference between the value of integrated data and the potential damage cost. Formally, Opt_{val} is defined as:

$$Opt_{val} = \max(NV) \tag{15}$$

5.2.3.11. Contributions of data providers. The contribution of each data provider DP_i is derived from the net value *NV* by fairly computing first the accumulative information gain $Score(x)$ of each data provider DP_i on the anonymized integrated dataset. Generally, the data provider whose data attributes result in greater information gain can get a significantly higher share of the monetary net value. Formally, $Cont_{DP_i}$ is defined as:

$$Cont_{DP_i} = \frac{Infogain_{DP_i}}{\sum_{i=1}^{n} Infogain_{DP_i}} \times NV \tag{16}$$

5.3. Risk measurement

In this section, we present an attack model to measure the risk associated with implicit weaknesses of privacy protection methods and the risk caused by explicit knowledge attack.

5.3.1. Attack model

Data providers participating in data integration express concern on two types of privacy threats: identity linkage and attribute linkage. Based on background knowledge, adversaries in identity linkage attacks can uniquely identify an individual, whereas adversaries in attribute linkage attacks can infer an individual's sensitive information with relatively high confidence. In this paper we employ classification analysis to quantify the potential privacy risks. Specifically, we build a C4.5 classifier by using the sensitive attribute as the class attribute, and we quantify the privacy risks by measuring the accuracy of predicting the sensitive values. There are many types of classification models, such as naive Bayesian, support vector machines, and so forth, that an adversary can employ to make predictions. Our proposed framework is flexible to adopt other classification methods to quantify the potential privacy risks.

Let D be the raw data, as shown in Table 1, and $D\prime$ be the anonymous integrated data from the mashup process of two data providers, as shown in Table 2. Recall that *Marital-status* is the sensitive attribute and *Loanapproval* is the class attribute. Let us assume that the data providers release their anonymized integrated data table D' to the data consumer (i.e., data recipient) with the classifier. A data recipient (or an adversary) can employ the C4.5 classification algorithm to infer sensitive records of individuals by setting the sensitive attribute *Divorced* as the class label. This approach is similar to Kifer (2009) in a way that a data recipient (or an adversary), instead of inferring new records on a class label, can predict the sensitive attribute value of a target victim who is a participant in the anonymized integrated training data.

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

Rec.ID	Data provider DP_1		Data provider DP_2		Sensitive	Class
	Age	Job	Sex	Education	Marital-status	Loan approval
1	[39–99]	Blue-collar	Any	Secondary	Divorced	N
2	[39–99]	White-collar	Any	Post-secondary	Never-married	Y
3	[33–39]	Blue-collar	Any	Secondary	Divorced	Y
4	[39–99]	Blue-collar	Any	Secondary	Never-married	N
5	[39–99]	Blue-collar	Any	Elementary	Married-civ-spouse	Y
6	[39–99]	White-collar	Any	Post-secondary	Married-civ-spouse	Y
7	[33–39]	White-collar	Any	Post-secondary	Never-married	Y
8	[39–99]	White-collar	Any	Post-secondary	Married-civ-spouse	N
9	[1–33]	Blue-collar	Any	Secondary	Divorced	N
10	[39–99]	Blue-collar	Any	Post-secondary	Divorced	Y
11	[39–99]	Blue-collar	Any	Secondary	Divorced	Y
12	[39–99]	White-collar	Any	Post-secondary	Married-civ-spouse	Y
13	[1–33]	Blue-collar	Any	Secondary	Married-civ-spouse	N
14	[39–99]	Blue-collar	Any	Post-secondary	Divorced	N
15	[33–39]	White-collar	Any	Post-secondary	Never-married	Y
16	[1–33]	Blue-collar	Any	Secondary	Divorced	N
17	[33–39]	Blue-collar	Any	Secondary	Divorced	Y
18	[39–99]	Blue-collar	Any	Secondary	Divorced	Y
19	[39–99]	Blue-collar	Any	Elementary	Married-civ-spouse	N

Table 3
Confusion matrix.

		Predicted class		
		A	**B**	**C**
Actual class	Divorced (**A**)	4	0	0
	Married-civ-spouse (**B**)	1	0	0
	Never-married (**C**)	0	1	0

5.3.1.1. Implicit risk measure. Implicit risk is due to attribute linkage attack (Fung et al., 2010Fung et al. 2010): an adversary attempts to infer the sensitive attribute value in the released dataset using a C4.5 classifier. In this type of attack, an adversary can negatively use the precision and recall performance measures to identify a victim's sensitive value. *Precision* indicates the measure of exactness or quality, meaning the number of correctly classified positive elements divided by the total number of elements classified as positive. *Recall* indicates the measure of completeness or quantity, which means the number of correctly classified positive elements divided by the total number of actual positive elements. We measure the adversary's power of inferring sensitive values by calculating the F-measure according to Eq. (10), which is a weighted harmonic mean of precision and recall measures. F-measure represents the probability of attack $Prob_{atk}$. An adversary may use these performance measures to determine the success rate of a privacy attack. We elaborate this by the following example.

Example 2. Consider the anonymous integrated data D' in Table 2. Suppose an adversary sets the sensitive attribute *Marital-status* as a class on D'. This results in a new integrated data table T^*. The adversary performs the attack by using the classification model C4.5 on T^* to infer the sensitive attribute value of the victim. Table 3 shows the confusion matrix for the classification of three classes. Each instance (e.g., an individual) has an actual class and a predicted class. The rows represent actual classes of the raw records, and the columns represent predictions made by the model. The entries on the diagonal indicate the correct predictions; other entries show the errors. For the sensitive value *Divorced*, true positive $TP = 4$, false negative $FN = 0$, and false positive $FP = 1$. So, the values of performance measures are *Precision* $= 80\%$, *Recall* $= 100\%$, and *F-measure* = 88.8%. □

5.3.1.2. Explicit risk measure. Explicit risk is due to record linkage attack (Fung et al. 2010): an adversary applies his or her background knowledge on the integrated data table T^* to predict the sensitive value of a victim who is part of the anonymized integrated training data. In addition, we assume that an adversary knows that a victim has a record on the table and also has some knowledge about the victim. For example, an adversary knows that the victim is female, age is greater than 35, education level is secondary, and job is cleaning. By applying this external knowledge to the anonymized integrated training data, the adversary finds a total of 3 records on the sensitive value *Divorced* under the class attribute *Marital-status*. The likelihood of the privacy breach L_{pb} for this case becomes 3/4, which is calculated according to Eq. (11). This implies that the adversary has a 75% confidence of inferring the sensitive value of the victim. The likelihood of a privacy breach would increase if the data providers are semihonest (Lindell and Pinkas 2009, Yao 1982).

6. Limitations

In this section, we discuss some of the limitations of our proposed business model that are inherent problems related to the cost-benefit analysis. Our model provides the basic framework for analyzing the cost-benefit of data mashup. The data providers can add, remove, or adjust the cost factors according to their specific applications and scenarios. The common sources of errors are *omission errors* and *valuation errors*. *Omission error* refers to excluding relevant factors in the process of factor analysis. *Valuation error* refers to making an incorrect estimation of the value of the cost factors, especially in the presence of intangible assets such as person-specific information. These errors do not undermine the value of cost-benefit analysis, and they are expected to decline with the passage of time by the increase in domain knowledge and follow-up of ex-post analysis (Boardman et al. 2006).

The privacy protection, database, and data mining communities have identified many types of potential privacy attacks, such as record linkage attack, attribute linkage attack, table linkage attack, and probabilistic attack. Consequently, many privacy models and anonymization methods (Fung et al. 2010), such as *MinGen*, *K-Optimize*, *Bottom-Up Generalization*, *Top-Down Specialization*, *Anatomy*, and *ϵ-Differential Additive Noise*, have been proposed to thwart these attacks. The objective of this paper is *not* to address all these privacy attacks. Instead, we are presenting a framework with a flexible cost-benefit business model for multiple data providers to achieve optimal mutual benefits given an agreed privacy requirement. Any partition-based anonymization methods that result in equivalent classes with counts are applicable to our framework. To illustrate the effectiveness of our proposed framework and model, in our discussion we adopt two anonymization algorithms, namely *TDSmdpm* and *DiffGen*, that can anonymize vertically-partitioned relational data. *TDSmdpm* and *DiffGen* were chosen because they can achieve two commonly employed privacy models, *LKC-privacy* and *differential privacy*, respectively. We would like to emphasize that our model is not limited to these privacy models and anonymization algorithms. They can be replaced, depending on the consent of privacy protection among the data providers. The negotiation process for reaching the consent is beyond the scope of this article.

7. Empirical study

In this section, we analyze and compare the costs and benefits for each data provider before participation in the data mashup process on their own data and after participation in the data mashup process on the integrated data. We evaluate our business model with the assumption of having 3 data providers who mashup their data using a secure Privacy-Preserving High-Dimensional Data Mashup (PHDMashup) algorithm (Fung et al. 2012) in a cloud environment. This model is independent of the cloud platform.

We employ a real-life dataset *Adult*[1] in our experiments, which has been widely used for many empirical studies. It is also known as the *de facto* benchmark for comparing the performance of anonymization algorithms (Fung et al. 2007, Mohammed et al. 2011, Hore et al. 2007). After removal of records with missing values, the *Adult* dataset contains 45,222 records with 8 categorical attributes, 6 numerical attributes, and a binary class attribute *Income* with two levels, $\leqslant$50*K* or >50*K*. For a classification analysis task this dataset is split into 30,162, 15,060 records for the training and testing set, respectively. We vertically partition the *Adult* dataset into three partitions P_1, P_2, and P_3 for data providers DP_1, DP_2, and DP_3, respectively. Table 4 represents the attributes with their types of each data provider. Each data provider computes *Baseline Accuracy (BA)* and *Classification Accuracy (CA)* on its raw dataset and anonymized dataset, respectively, by using a C4.5 classifier. The *BA* is 81.8%, 82.5%, and 75.6% on DP_1, DP_2, and DP_3 datasets, respectively.

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

Table 4
Attributes hosted by each data provider.

Attribute	Type
DP_1	
Age	Numerical
Hours-per-week	Numerical
Workclass	Categorical
Capital-gain	Numerical
Income	Categorical
Marital-status	Categorical
DP_2	
Education	Categorical
Education-num	Numerical
Occupation	Categorical
Capital-loss	Numerical
Income	Categorical
Marital-status	Categorical
DP_3	
Sex	Categorical
Race	Categorical
Relationship	Categorical
Final-weight	Numerical
Native-country	Categorical
Income	Categorical
Marital-status	Categorical

Whereas, the baseline accuracy (BA) on the integrated data is 85.3% using the secure multiple party classifier (Du and Zhan 2002) without sharing their raw data. We consider *Income* as the class attribute and *Marital-status* as the sensitive attribute in each data provider's table. The remaining attributes in each data provider's table are the *QID* attributes. We consider *Married-civ-spouse* and *Divorced* in the attribute *Marital-status* as sensitive. In addition, a common unique ID is included in each table for joining the data provider's tables. All experiments were performed on an Intel Core i3-2350 M 2.3 GHz PC with 4 GB memory.

7.1. Cost of anonymization without data mashup

In this section, we analyze the cost of anonymization $Cost_{ad}$ to individual data providers without their participation in the data mashup process. Suppose the sensitivity of the dataset $Sen_{ds_i} = 2$ on the scale of 1–5, the price per attribute $Price_{attr_i} = \$0.1$, the size of dataset $Size_{ds_i} = 45,222$ for the data providers DP_1, DP_2, and DP_3 to fairly quantify and compare the cost of anonymization under different privacy models including *K-anonymity*, *LKC-privacy*, and ϵ-differential privacy.

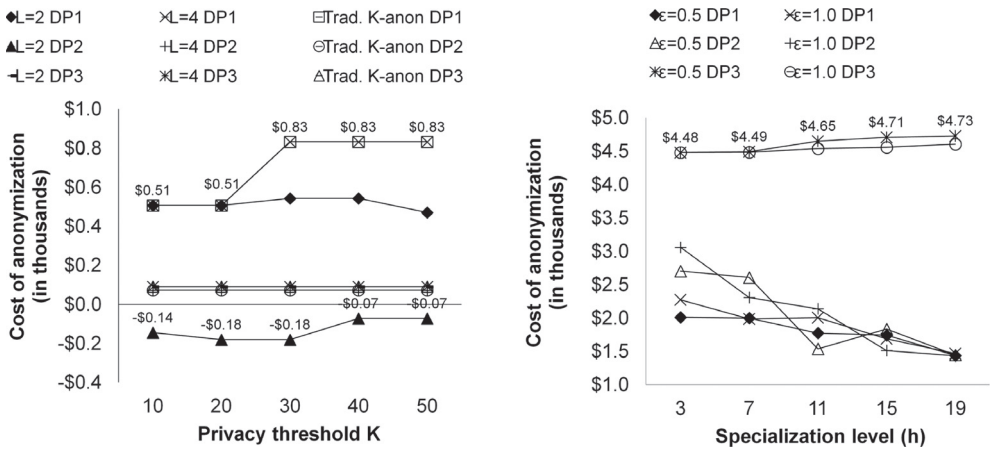
Fig. 4 depicts the cost of anonymization to each data provider without participating in the data mashup process. Fig. 4a depicts the cost of anonymization when privacy models *K-anonymity* and *LKC-privacy* are enforced with the anonymity threshold L, K, and C. $Cost_{ad}$ generally increases as K or L increases, but this monotonicity does not maintain for DP_1 and DP_2 with the increase of K. For example, $Cost_{ad}$ decreases by \$72.35 for DP_1, when K increases from 40 to 50 when $L = 2$. This is because of the better classification accuracy CA, which is increased from 80.3% to 80.5%. This anti-monotonic property of the algorithm helps in finding the sub-optimal anonymization cost. We observe that the DP_1 anonymization cost is higher than DP_2 and DP_3 because DP_1 holds 3 continuous numeric attributes (refer to Table 4) that require discretizing into intervals (categorical values) for anonymization. The classification analysis on new data would be less accurate than categorical attributes due to the chance of information loss. The $Cost_{ad}$ of *LKC-privacy* equals the $Cost_{ad}$ of the traditional *K*-anonymity when $L = 4$ for each data provider. $Cost_{ad}$ is also insensitive to the change of confidence threshold $10\% \leqslant C \leqslant 50\%$.

Fig. 4b depicts the cost of anonymization when ϵ-*differential privacy* is enforced with privacy parameters $\epsilon = 0.5$ and 1.0 and specialization levels $3 \leqslant h \leqslant 19$. We observe that $Cost_{ad}$ generally decreases when the specialization level h increases for DP_1 and DP_2 with the setting of a privacy budget to either $\epsilon = 0.5$ or 1.0. But this trend is quite different in relation to DP_3 where $Cost_{ad}$ increases monotonically with the increase in h; the random noise results in lower classification accuracy.

7.2. Cost of anonymization in integrated data

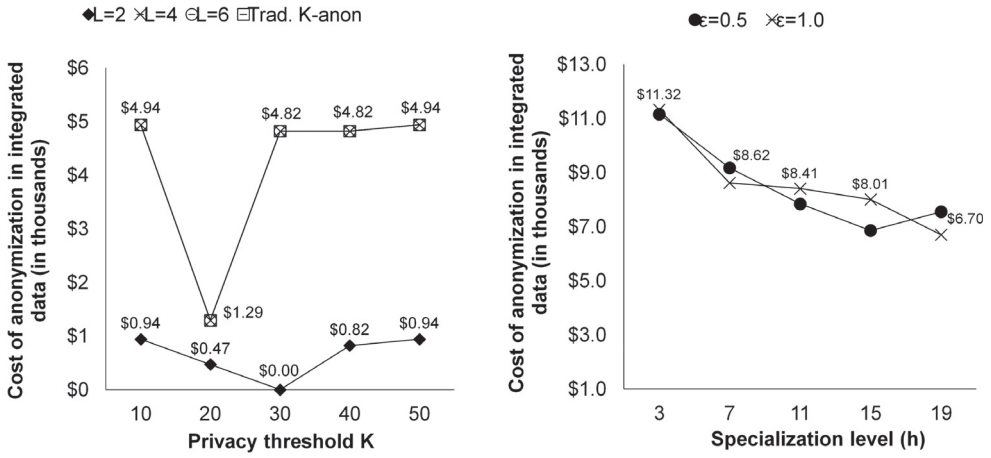
In this section, we analyze the cost of anonymization in integrated data $Cost_{intgdata}$ under the joint privacy settings of the three contributing data providers in the data mashup process. Suppose the sensitivity of the dataset $Sen_{ds_i} = 2$ on the scale of 1–5, the price per attribute $Price_{attr_i} = \$0.1$, the number of attributes $Count_{attr} = 13$ (sum of attributes of DP_1, DP_2, and DP_3) and the size of dataset $Size_{ds_i} = 45,222$ to quantify and compare the cost of anonymization in integrated data under different privacy models, including *K-anonymity*, *LKC-privacy*, and ϵ-*differential privacy*.

Fig. 5a depicts the cost of anonymization in integrated data when privacy models *K-anonymity* and *LKC-privacy* are enforced



(a) *K-anonymity* and *LKC-privacy* (b) ϵ-*differential privacy*

Fig. 4. Cost of anonymization to individual data provider without data mashup.



(a) *K-anonymity* and *LKC-privacy* (b) *ϵ-differential privacy*

Fig. 5. Cost of anonymization in integrated data.

with the anonymity threshold $10 \leqslant K \leqslant 50$, background knowledge $L = \langle 2, 4, 6 \rangle$, and confidence threshold $C = 50\%$. $Cost_{intgdata}$ generally increases as L increases, but does not exhibit obvious monotonicity with the increase of K. For example, $Cost_{intgdata}$ decreases by \$3644.89 when K increases from 10 to 20 when $L = 4$ and $L = 6$. This is because of improvement in classification accuracy *CA*, which increases by 3.1%. This helps in finding the sub-optimal anonymization cost. The $Cost_{intgdata}$ of *LKC*-privacy equals the $Cost_{intgdata}$ of traditional *K*-anonymity when $L = 4$ and $L = 6$. $Cost_{intgdata}$ is also insensitive to the change of confidence threshold $10\% \leqslant C \leqslant 50\%$.

Fig. 5b depicts the cost of anonymization in integrated data when ϵ-*differential privacy* is enforced with privacy parameters $\epsilon = 0.5$ and 1.0 and specialization levels $3 \leqslant h \leqslant 19$. We calculate the average accuracy on 10 runs. We observe that $Cost_{intgdata}$ generally decreases as the specialization level h increases, except an increase by \$693.71 when privacy budget $\epsilon = 0.5$ and the specialization level h increases from 15 to 19. When ϵ is small, having too many levels makes each specialization less accurate.

7.3. Implicit risk measure

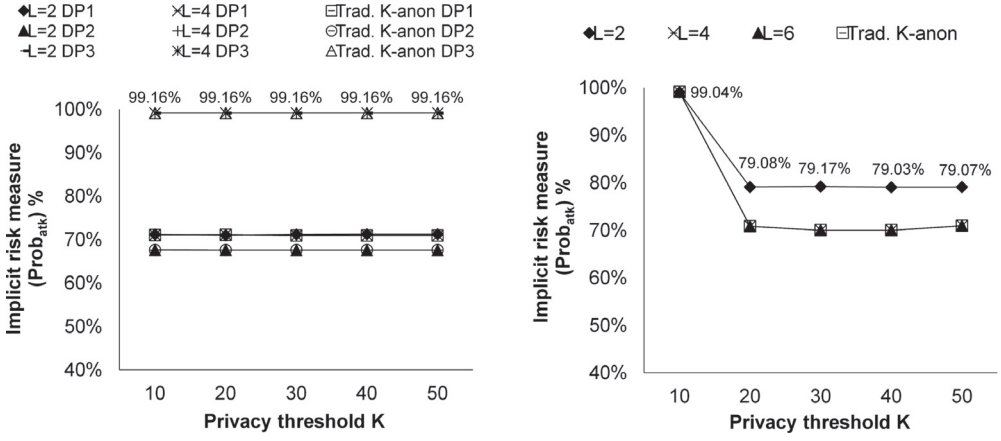
In this section, we analyze the implicit risk for each data provider before participation in the data mashup process on their own data and after participation in the data mashup process on the integrated data of the contributing data providers.

Fig. 6a depicts the probability of attack $Prob_{atk}$ on the sensitive value *Married-civ-spouse* to the data providers DP_1, DP_2, and DP_3 with privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = \langle 2, 4 \rangle$, and confidence threshold $C = 50\%$. We observe that the chance of inferring the sensitive attribute value is approximately 71%, 67%, and 99% on the anonymized dataset of DP_1, DP_2, and DP_3, respectively. DP_2 is comparatively better than DP_1 and DP_3 because it has less risk of inferring the sensitive attribute value.

Fig. 6b depicts the probability of attack $Prob_{atk}$ on the sensitive value *Married-civ-spouse* in the anonymized integrated dataset of contributing data providers DP_1, DP_2, and DP_3 under the joint privacy settings with the anonymity threshold $10 \leqslant K \leqslant 50$, background knowledge $L = \langle 2, 4, 6 \rangle$, and confidence threshold $C = 50\%$. We can observe the trend that $Prob_{atk}$ generally decreases as K or L increases, which also conforms to the theoretical analysis.

7.4. Explicit risk measure

In this section, we analyze the explicit risk for each data provider before participation in the data mashup process on their own data and after participation in the data mashup process on the integrated data of contributing data providers.



(a) without data mashup (b) with data mashup

Fig. 6. Implicit risk measure.

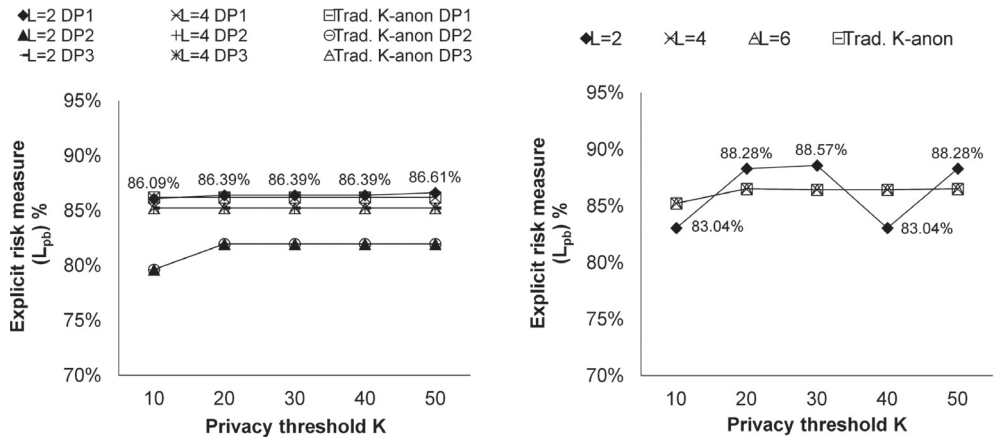


Fig. 7. Explicit risk measure.

Suppose an adversary has prior knowledge about a male victim, that his *age* is between 40 and 50, his *education* is *masters*, his *hours-per-week* is >40, and his *income* is ⩾50,000.

Fig. 7a depicts the likelihood of a privacy breach L_{pb} on the sensitive value *Married-civ-spouse* when the aforementioned external knowledge about the victim is linked to the data providers DP_1, DP_2, and DP_3 attributes, where privacy threshold $10 \leqslant K \leqslant 50$, background knowledge $L = \langle 2, 4 \rangle$, and confidence threshold $C = 50\%$. We observe that the L_{pb} is approximately 86%, 82%, and 85% on the anonymized dataset of DP_1, DP_2, and DP_3, respectively. DP_2 is comparatively better than DP_1 and DP_3 because it has less risk of a privacy breach.

Fig. 7b depicts the likelihood of a privacy breach L_{pb} on the sensitive value *Married-civ-spouse* when the aforementioned external knowledge about a victim is linked to the anonymized integrated dataset of contributing data providers DP_1, DP_2, and DP_3 under the joint privacy settings with the anonymity threshold $10 \leqslant K \leqslant 50$, background knowledge $L = \langle 2, 4, 6 \rangle$, and confidence threshold $C = 50\%$. Generally, L_{pb} decreases with the increase of L but this trend is not obvious with the increase of K. For example, L_{pb} is 86.44% when $K = 40$ and $L = \langle 4, 6 \rangle$, which is higher by 3.4% when $L = 2$. This anti-monotonic property of the TDS algorithm helps in identifying 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 attribute *Marital-status* remains unchanged with the increase of L. Though not shown in the figure, L_{pb} is insensitive to the change of the confidence threshold $10\% \leqslant C \leqslant 50\%$.

7.5. Impact of privacy requirements on net value

In this section, we analyze the impact of *K-anonymity*, *LKC-privacy*, and *ϵ-differential privacy* requirements on monetary value for each data provider before participation in the data mashup process and after participation on the integrated data of contributing data providers. Suppose the sensitivity of the dataset $Sen_{ds_i} = 2$ on the scale of 1–5, the price per attribute $Price_{attr_i} = \$0.1$, the expected cost of lawsuit $Ecost_{lwst} = \$1000$, the size of dataset $Size_{ds_i} = 45,222$, and the fixed operating cost $F_{OpCost} = \$300$.

Fig. 8 depicts the impact of *K*-anonymity and *LKC*-privacy requirements on DP_1's net value, where privacy threshold $10 \leqslant K \leqslant 50$, and confidence threshold $C = 50\%$. Fig. 8a depicts the impact on DP_1's net value when the threshold $L = 2$. We observe that DP_1's net value without data mashup (refer to the DP_1's attributes in Table 4) decreases slightly with the increase of K, but it does not maintain monotonicity when $K = 50$. On the other side, DP_1's net value with data mashup drops with the increase of K from 10 to 30, but the net value rises when $K > 30$. This change in trend depends on the information gain for classification analysis of the DP_1's attributes. Fig. 8b depicts the impact on DP_1's net value when the threshold $L = 4$. We observe that DP_1's net value without data mashup decreases slightly with the increase of K from 10 to 30, but it is insensitive to change when $K > 30$. On the other side, *DP1*'s net value with data mashup does not exhibit monotonicity with the increase of K because DP_1's attributes for classification analysis contribute different information gains at different privacy thresholds K on

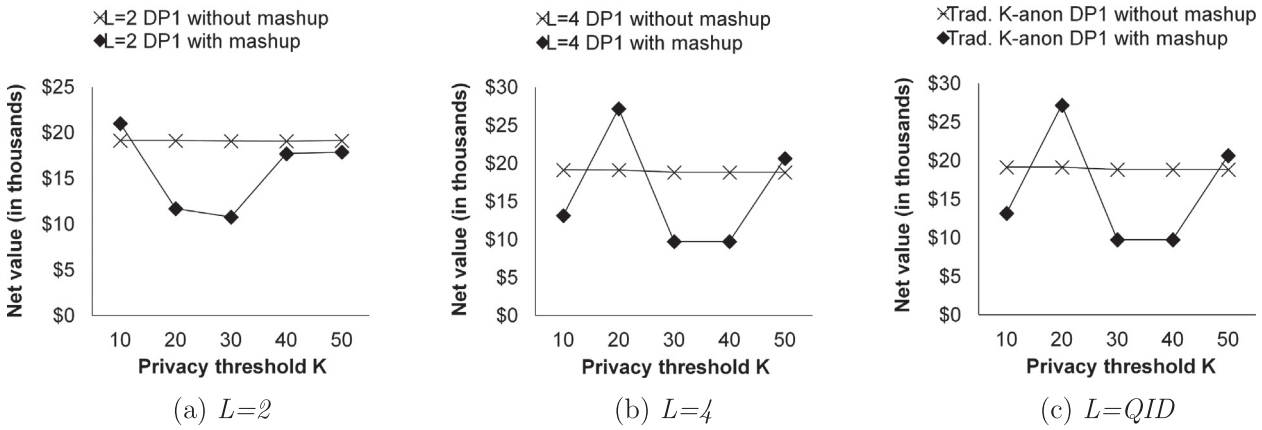


Fig. 8. Impact of *K*-anonymity and *LKC*-privacy requirements on DP_1's net value.

integrated data with collaborating data providers DP_2 and DP_3. Fig. 8c depicts the impact on DP_1's net value when the threshold $L = QID$. There are a total of 4 QID attributes in DP_1's dataset. DP_1's net value of traditional K-anonymity is equal to LKC-privacy when $L = 4$. Though not shown in Fig. 8, net value is insensitive to the change of the confidence threshold $10\% \leqslant C \leqslant 50\%$. The maximum net value achieved by the DP_1 is \$27,190.94 when $K = 20$ and $L = 4$.

Fig. 9 depicts the impact of K-anonymity and LKC-privacy requirements on DP_2's net value, where privacy threshold $10 \leqslant K \leqslant 50$, and confidence threshold $C = 50\%$. Fig. 9a depicts the impact on DP_2's net value when the threshold $L = 2$. We observe that DP_2's net value without data mashup (refer to the DP_2's attributes in the Table 4) decreases slightly with the increase of K except when $K = 30$. On the other side, DP_2's net value with data mashup increases with the increase of K from 10 to 30, but the net value drops when $K > 30$. This change in trend depends on the information gain for classification analysis of DP_2's attributes. Fig. 9b depicts the impact on DP_2's net value when the threshold $L = 4$. We observe that DP_2's net value without data mashup decreases slightly with the increase of K from 10 to 20, but it is insensitive to change when $K > 20$. On the other side, DP_2's net value with data mashup increases with the increase of K from 10 to 40, but it drops when $K = 50$. This drop in net value is due to the loss of information gain in classification analysis. Fig. 9c depicts the impact on DP_2's net value when the threshold $L = QID$. There are a total of 4 QID attributes in DP_2's dataset. DP_2's net value of traditional K-anonymity is equal to LKC-privacy when $L = 4$. Though not shown in Fig. 9, net value is insensitive to the change in the confidence threshold $10\% \leqslant C \leqslant 50\%$. The maximum net value achieved by DP_2 is \$68,060.37 when $K = 30$ and $K = 40$, and $L = 4$.

Fig. 10 depicts the impact of K-anonymity and LKC-privacy requirements on DP_3's net value, where privacy threshold $10 \leqslant K \leqslant 50$, and confidence threshold $C = 50\%$. Fig. 10a depicts the impact on DP_3's net value when the threshold $L = 2$. We observe that DP_3's net value without data mashup (refer to the DP_3's attributes in Table 4) is insensitive to change with the increase of K. On the other side, DP_3's net value with data mashup drops with the increase of K from 10 to 20, but the net value gradually rises when $K > 20$. This change in trend depends on the information gain for classification analysis of DP_3's attributes. Fig. 10b depicts the impact on DP_3's net value when the threshold $L = 4$. We observe that DP_3's net value without data mashup is insensitive with the increase of K. On the other side, DP_3's net value with data mashup drops with the increase of K except when $K = 50$. This fall in net value is due to the loss of information gain in classification analysis. Fig. 10c depicts the impact on DP_3's net value when the threshold $L = QID$. There are a total of 5 QID attributes in DP_3's dataset. DP_3's net value of traditional K-anonymity is equal to LKC-privacy when $L = 4$. Though not shown in Fig. 10, net value is insensitive to the change of the confidence threshold $10\% \leqslant C \leqslant 50\%$. The maximum net value achieved by DP_3 is \$34,522.01 when $K = 10$ and $L = 4$.

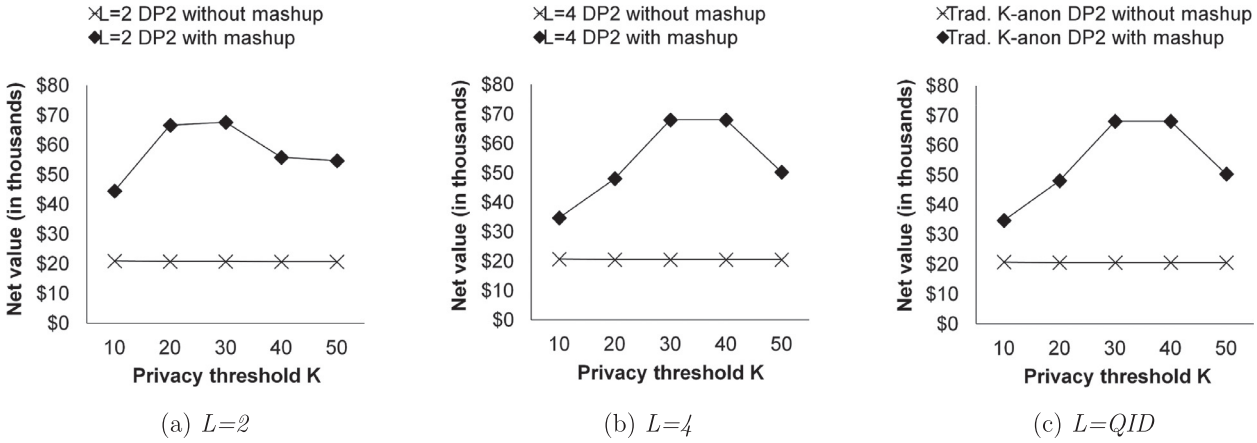


Fig. 9. Impact of K-anonymity and LKC-privacy requirements on DP_2's net value.

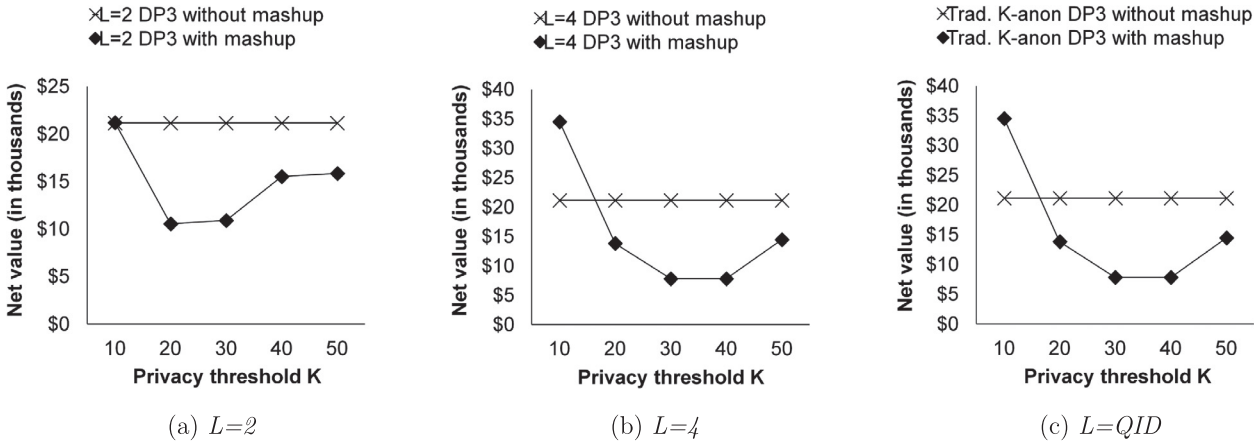


Fig. 10. Impact of K-anonymity and LKC-privacy requirements on DP_3's net value.

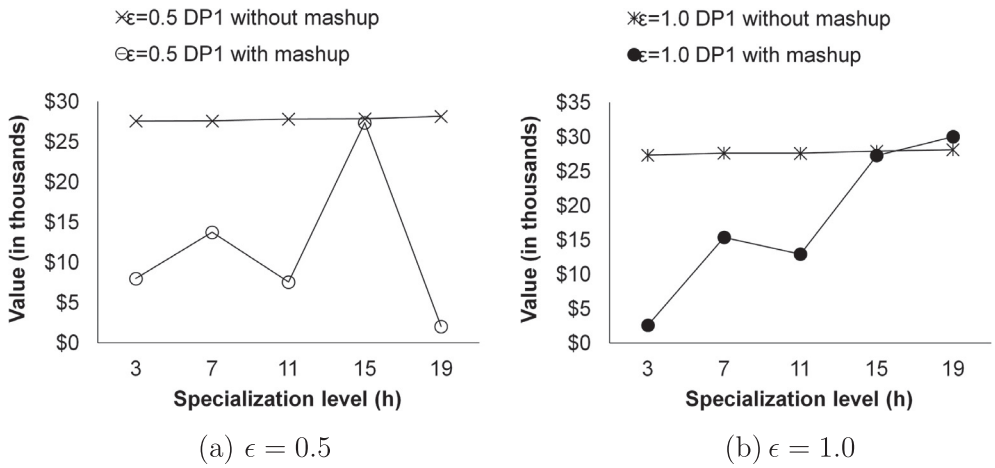


Fig. 11. Impact of ϵ-*differential privacy* requirements on DP_1's monetary value.

Fig. 11 depicts the impact on DP_1's monetary value when ϵ-*differential privacy* is enforced with privacy parameters ϵ=0.5 and 1.0 and specialization levels $3 \leqslant h \leqslant 19$. Fig. 11a depicts the impact on DP_1's monetary value when the threshold $\epsilon = 0.5$. We observe that DP_1's monetary value without data mashup (refer to the DP_1's attributes in Table 4) increases monotonically as the increase in specialization level h. On the other side, DP_1's monetary value with data mashup increases when specialization level h increases from 3 to 7 and 11 to 15, but the value drops due to the loss of data utility when $h = 11$ and $h = 19$. Fig. 11b depicts the impact on DP_1's monetary value when the threshold $\epsilon = 1.0$. We observe that DP_1's monetary value without data mashup increases slightly with the increase in the specialization level h except when $h = 11$. DP_1's net value with data mashup generally increases with the increase in h, but it does not maintain monotonicity when $h = 11$ due to the provision of less data utility in classification analysis with collaborating data providers DP_2 and DP_3. The benefits to DP_1 of doing data mashup is higher than going without data mashup by gaining the maximum net value \$30,0187.37 when $\epsilon = 1.0$ and $h = 19$.

Fig. 12 depicts the impact on DP_2's monetary value when ϵ-*differential privacy* is enforced with privacy parameters ϵ = 0.5 and 1.0 and specialization levels $3 \leqslant h \leqslant 19$. Fig. 12a depicts the impact on DP_2's monetary value when the threshold $\epsilon = 0.5$. We observe that DP_2's monetary value without data mashup (refer to the DP_2's attributes in Table 4) generally increases as the increase in specialization level h except when $h = 15$. DP_2's monetary value with data mashup does not exhibit monotonicity with the increase in the specialization level h due to the loss of data utility in classification analysis when $h = 11$ and the provision of less data utility in comparison to the other collaborating data providers DP_1 and DP_3 when $h = 15$. Fig. 12b depicts the impact on DP_2's monetary value when the threshold $\epsilon = 1.0$. We observe that DP_2's monetary value without data mashup increases monotonically with the increase in the specialization level h. On the other side, DP_2's monetary value with data mashup does not exhibit monotonicity with the increase in the specialization level h due to the loss of data utility in classification analysis when $h = 7$ and the provision of less data utility in comparison to other collaborating data providers DP_1 and DP_3 when $h = 19$. The benefits of doing data mashup are higher than doing without data mashup to DP_2 by gaining the maximum net value \$29,971.26 when $\epsilon = 0.5$ and $h = 7$.

Fig. 13 depicts the impact on DP_3's monetary value when ϵ-*differential privacy* is enforced with privacy parameters ϵ = 0.5 and 1.0 and specialization levels $3 \leqslant h \leqslant 19$. Fig. 13a depicts the impact on DP_3's monetary value when the threshold $\epsilon = 0.5$. We observe that DP_3's monetary value without data mashup (refer to the DP_3's attributes in Table 4) decreases slightly as the specialization level h increases. DP_3's monetary value with data mashup does not exhibit monotonicity with the increase in the specialization level h, but DP_3's monetary value is greater than DP_1 and DP_2 at specialization levels 3–19. Fig. 13b depicts the impact on DP_3's

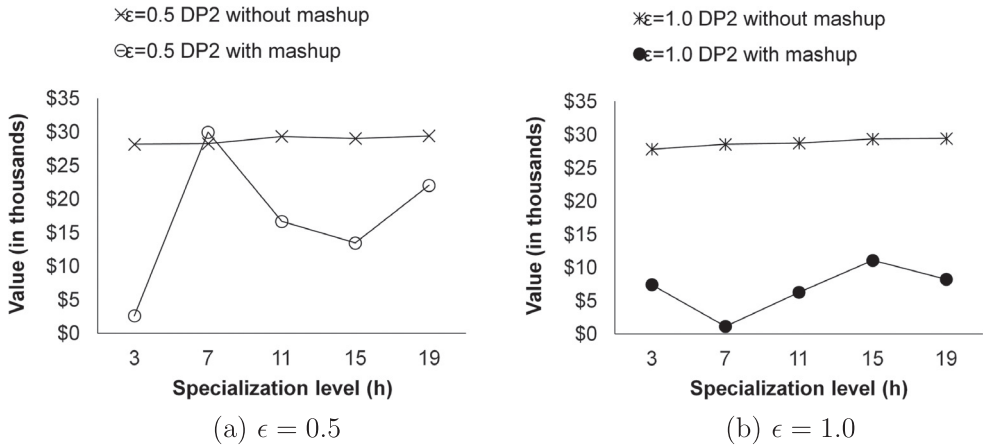


Fig. 12. Impact of ϵ-*differential privacy* requirements on DP_2's monetary value.

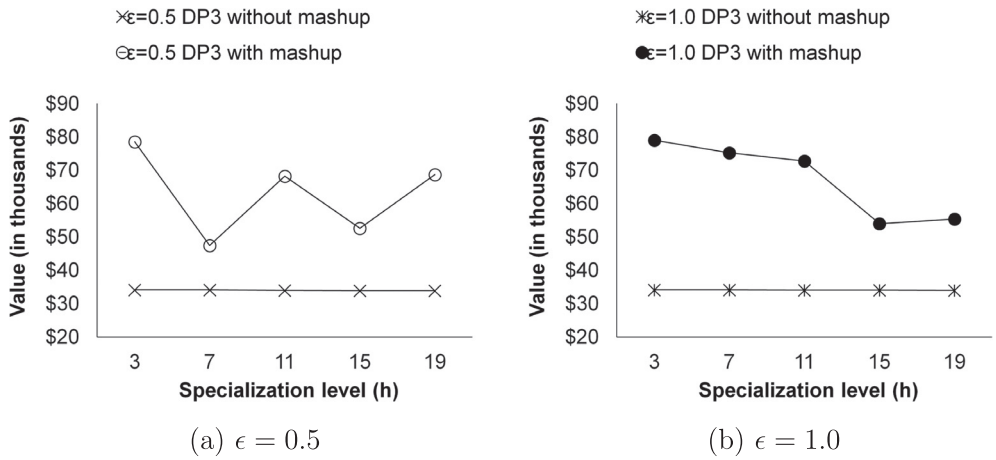


Fig. 13. Impact of *ϵ-differential privacy* requirements on DP_3's monetary value.

monetary value when the threshold $\epsilon = 1.0$. We observe that DP_3's monetary value without data mashup decreases slightly as the specialization level h increases. On the other side, DP_3's monetary value with data mashup decreases with the increase in the specialization level h except when $h = 19$. The benefits of doing data mashup is higher than going without data mashup to DP_3 by gaining the maximum net value $78,993.45 when $\epsilon = 1.0$ and $h = 3$.

8. Conclusion

We have proposed a business model to quantify and compare the costs and benefits for releasing integrated anonymized data of multiple providers over an individual data provider when trading person-specific information in the e-market. Our model enables data providers to set up their joint privacy requirements for classification analysis on mashup data. The data mashup process is implemented fairly that allows data providers to integrate their data subject to the given privacy requirements. During the data mashup process every data provider competes with the other participating data providers to generate more profit from their own data. The data provider whose data provides more information gain will get a significantly higher share in terms of monetary value from the distribution of the achieved net value. We have incorporated relevant factors that are associated with the revenue and costs to determine the net value. Our model helps data providers in finding the optimal value by evaluating the benefits of data mashup and impacts of data anonymization based on the choices of privacy models and data mashup anonymization algorithms.

Acknowledgments

This research is supported in part by the Discovery Grants (356065-2013) from the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs Program (950-230623), and the Research Incentive Fund (RIF) #R14033 from Zayed University.

References

A Legal Guide to Privacy and Data Security, 2014. Minnesota Department of Employment and Economic Development, Gray Plant Mooty.

Acquisti, A., Friedman, A., Telang, R., 2006. Is there a cost to privacy breaches? An event study. In: Proceedings of the 27th International Conference on Information.

Aljafer, H., Malik, Z., Alodib, M., Rezgui, A., 2014. A brief overview and an experimental evaluation of data confidentiality measures on the cloud. Journal of Innovation in Digital Ecosystems 1 (1–2), 1–11.

Arafati, M., Dagher, G.G., Fung, B.C.M., Hung, P.C., 2014. D-mash: a framework for privacy-preserving data-as-a-service mashups. In: Proceedings of the 7th IEEE International Conference on Cloud Computing. IEEE Computer Society, pp. 498–505.

Backman, P., Levin, K., 2011. Privacy Breaches – Impact, Notification and Strategic Plans. Aird and Berlis LLP.

Bevitt, A., Retzer, K., Lopatowska, J., 2012. Dealing with Data Breaches in Europe and Beyond. Practical Law Company.

Boardman, A.E., Greenberg, D.H., Vining, A.R., Weimer, D.L., 2006. Cost-Benefit Analysis: Concepts and Practice. Pearson Prentice Hall.

Currie, W.L., Seddon, J.J.M., 2014. A cross-country study of cloud computing policy and regulation in healthcare. In: Proceedings of the 22nd European Conference on Information Systems.

Dalenius, T., Reiss, S.P., 1982. Data-swapping: a technique for disclosure control. Journal of Statistical Planning and Inference 6 (1), 73–85.

Data Management Platforms Buyer's Guide, 2013. Econsultancy Digital Marketing Excellence.

Data Partners, 2014. Seventh Point. Last accessed: June 11, 2015; URL http://www.seventhpoint.com/whitepaper/data-partners/.

Department of Health and Human Services, 2013. 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 (78 FR 5565), pp. 5565–5702.

Domingo-Ferrer, J., Mateo-Sanz, J.M., 2002. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14 (1), 189–201.

Du, W., Zhan, Z., 2002. Building decision tree classifier on private data. Proceedings of the IEEE International Conference on Privacy, Security and Data Mining, vol. 14. Australian Computer Society, Inc., pp. 1–8.

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006. Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Conference on Theory of Cryptography. Springer.

Fung, B.C.M., Wang, K., Yu, P.S., 2007. Anonymizing classification data for privacy preservation. IEEE Transactions on Knowledge Data Engineering 19 (5), 711–725.

Fung, B.C.M., Wang, K., Chen, R., Yu, P.S., 2010. Privacy-preserving data publishing: a survey of recent developments. ACM Computing Survey 42 (4), 14:1–14:53.

Fung, B.C.M., Trojer, T., Hung, P.C., Xiong, L., Al-Hussaeni, K., Dssouli, R., 2012. Service-oriented architecture for high-dimensional private data mashup. IEEE Transactions on Services Computing 5 (3), 373–386.

Gates, C., Matthews, P., 2014. Data is the new currency. In: Proceedings of the 2014 Workshop on New Security Paradigms Workshop. ACM, pp. 105–116.

Gehrke, J., 2010. Programming with differential privacy: technical perspective. Communications of the ACM 53 (9).

Gwebu, K.L., Wang, J., Xie, W., 2014. Understanding the cost associated with data security breaches. In: Proceedings of the 18th Pacific Asia Conference on Information Systems.

Hore, B., Jammalamadaka, R.C., Mehrotra, S., 2007. Flexible anonymization for privacy preserving data publishing: a systematic search based approach. In: Proceedings of the Seventh SIAM International Conference on Data Mining.

Hirshleifer, J., Glazer, A., Hirshleifer, D., 2005. Price Theory and Applications: Decisions, Markets, and Information, 7th ed. Cambridge University Pres.

Hu, Y.J., Wu, W.N., Cheng, D.R., 2012. Towards law-aware semantic cloud policies with exceptions for data integration and protection. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. ACM, pp. 26:1–26:12.

Jurczyk, P., Xiong, L., 2009. Distributed anonymization: achieving privacy for both data subjects and data providers. In: Proceedings of the 23rd Annual IFIP WG 11.3 Working Conference on Data and Applications Security. Springer-Verlag, pp. 191–207.

Khokhar, R.H., Chen, R., Fung, B.C.M., Lui, S.M., 2014. Quantifying the costs and benefits of privacy-preserving health data publishing. Journal of Biomedical Informatics 50 (0), 107–121 (Special Issue on Informatics Methods in Medical Privacy).

Kifer, D., 2009. Attacks on privacy and deFinetti's theoerm. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM, pp. 127–138.

Kim, J., 1986. A method for limiting disclosure in microdata based on random noise and transformation. In: Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 303–308.

Kooiman, P., Willenborg, L., Gouweleeuw, J., 1997. PRAM: a method for disclosure limitation of microdata. No. 9705 in research paper. CBS.

Kuner, C., 2011. Regulation of Transborder Data Flows under Data Protection and Privacy Law: Past Present and Future. OECD Publishing, p. 187.

LeFevre, K., DeWit, D.J., Ramakrishnan, R., 2006. Workload-aware anonymization. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 277–286.

Li, C., Li, D.Y., Miklau, G., Suciu, D., 2014. A theory of pricing private data. ACM Transactions on Database Systems 39 (4), 34:1–34:28.

Lindell, Y., Pinkas, B., 2009. Secure multiparty computation for privacy-preserving data mining. Journal of Privacy and Confidentiality 1 (1), 59–98.

Little, R.J., 1993. Statistical analysis of masked data. Journal of Official Statistics 9 (2), 407–426.

Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M., 2007. ℓ-Diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data 1 (1).

McSherry, F., Talwar, K., 2007. Mechanism design via differential privacy. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. IEEE Computer Society, pp. 94–103.

Mohammed, N., Fung, B.C.M., Hung, P.C., Lee, C.K., 2009. 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. ACM, pp. 1285–1294.

Mohammed, N., Fung, B.C.M., Hung, P.C.K., Lee, C.K., 2010. Centralized and distributed anonymization for high-dimensional healthcare data. ACM Transactions on Knowledge Discovery from Data 4 (4), 18:1–18:33.

Mohammed, N., Chen, R., Fung, B.C.M., Yu, P.S., 2011. Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 493–501.

Mohammed, N., Alhadidi, D., Fung, B.C.M., Debbabi, M., 2014. Secure two-party differentially private data release for vertically partitioned data. IEEE Transactions on Dependable and Secure Computing 11 (1), 59–71.

OECD, 2013. Exploring the economics of personal data: a survey of methodologies for measuring monetary value. OECD Digital Economy Papers 220.

Personal Data Privacy and Security Act, 2011. Bill S.1151 – 112th Congress in the Senate of the United States.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

Review of the Personal Data (Privacy) Ordinance, 2009. Office of the Privacy Commissioner for Personal Data, Hong Kong.

Riederer, C., Erramilli, V., Chaintreau, A., Krishnamurthy, B., Rodriguez, P., 2011. For sale: your data. In: Proceedings of the 10th ACM Workshop on Hot Topics in Networks. ACM, pp. 1–6.

Roebuck, K., 2012. Enterprise Mashups: High-impact Strategies – What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. Emereo Publishing.

Romanosky, S., Acquisti, A., 2014. Privacy costs and personal data protection: economic and legal perspectives. Berkeley Technology Law Journal 24 (4).

Romanosky, S., Hoffman, D., Acquisti, A., 2014. Empirical analysis of data breach litigation. Journal of Empirical Legal Studies 11 (1), 74–104.

Samarati, P., Sweeney, L., 2001. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. IEEE Transactions on Knowledge and Data Engineering.

Skinner, C., Marsh, C., Openshaw, S., Wymer, C., 1994. Disclosure control for census microdata. Journal of Official Statistics 10 (1), 31–51.

Takemura, A., 1999. Local Recoding by Maximum Weight Matching for Disclosure Control of Microdata Sets. CIRJE F-Series 40. Faculty of Economics, University of Tokyo.

Truta, T.M., Fotouhi, F., Barth-Jones, D., 2003. Privacy and confidentiality management for the microaggregation disclosure control method: disclosure risk and information loss measures. In: Proceedings of the 2003 ACM Workshop on Privacy in the Electronic Society. ACM, pp. 21–30.

Waal, A.D., Willenborg, L.C., 1998. Optimal local suppression in microdata. Journal of Official Statistics 14 (4), 421–435.

Waal, T.D., Willenborg, L., 1999. Information loss through global recoding and local suppression. Netherlands Official Statistics 14, 17–20 (Special Issue on Statistical Disclosure Control).

Wixom, B.H., Buff, A., Tallon, P., 2015. Six Sources of Value for Information Businesses. MIT Sloan Center for Information Systems Research and SAS Institute Inc..

Wixom, B.H., Markus, M.L., 2015. Data Value Assessment: Recognizing Data as an Enterprise Asset. MIT Sloan Center for Information Systems Research.

Yao, A.C., 1982. Protocols for secure computations. In: Proceedings of the 23rd Annual Symposium on Foundations of Computer Science. IEEE Computer Society, pp. 160–164.