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/© 2016 Elsevier B.V. All rights reserved.

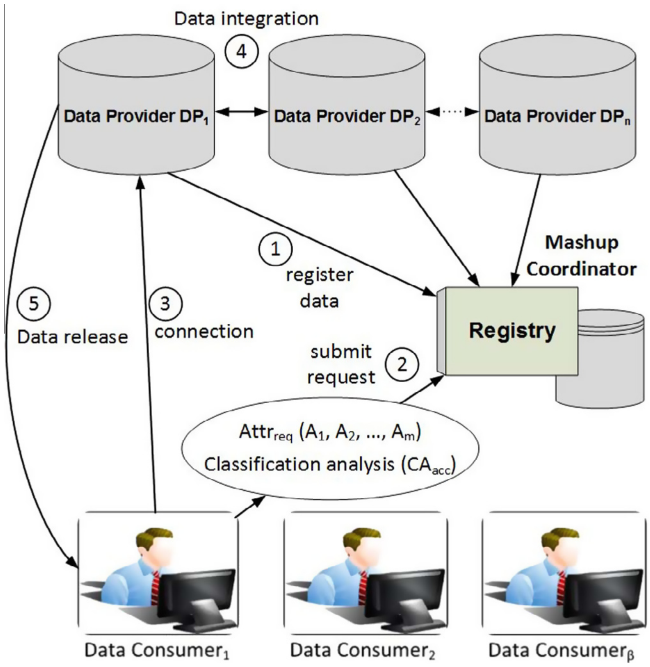

Fig. 1. Privacy-preserving data mashup architecture for trading person-specific information.

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