Computer Networks 88 (2015) 136–148

Contents lists available at ScienceDirect

Computer Networks

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

Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation

Wei Wang [a,b,*], Lei Chen [b,c], Qian Zhang [b,c]

[a] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei, China
[b] Fok Ying Tung Research Institute, Hong Kong University of Science and Technology, Kowloon, Hong Kong
[c] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 19 March 2015
Revised 15 May 2015
Accepted 25 June 2015
Available online 30 June 2015

Keywords:
Healthcare data
Privacy
Hybrid cloud

ABSTRACT

According to the recent rule released by Health and Human Services (HHS), healthcare data can be outsourced to cloud computing services for medical studies. A major concern about outsourcing healthcare data is its associated privacy issues. However, previous solutions have focused on cryptographic techniques which introduce significant cost when applied to healthcare data with high-dimensional sensitive attributes. To address these challenges, we propose a privacy-preserving framework to transit insensitive data to commercial public cloud and the rest to trusted private cloud. Under the framework, we design two protocols to provide personalized privacy protections and defend against potential collusion between the public cloud service provider and the data users. We derive provable privacy guarantees and bounded data distortion to validate the proposed protocols. Extensive experiments over real-world datasets are conducted to demonstrate that the proposed protocols maintain high usability and scale well to large datasets.

1. Introduction

Gaining access to healthcare data is a vital requirement for medical practitioners and pharmaceutical researchers to study characteristics of diseases. In recent years, the proliferation of cloud computing services enables hospitals and institutions to transit their healthcare data to the cloud, which provides ubiquitous data access and on-demand high quality services at a low cost. On January 25, 2013, the U.S. Department of Health and Human Services (HHS) released the Omnibus Rule [1], which defines cloud service providers (CSPs) as business associates for healthcare data. Currently, many CSPs, including Box, Microsoft, Verizon and Dell, have announced their support for this business associate agreement.

Despite the benefits of healthcare cloud services, the associated privacy issues are widely concerned by individuals and governments [1,2]. Privacy risks rise when outsourcing personal healthcare records to cloud due to the sensitive nature of health information and the social and legal implications for its disclosure. A natural method is to encrypt healthcare data before transiting them to cloud [3–5]. However, processing encrypted data are not efficient and is limited to specific operations, and thus is not suitable for healthcare data with versatile usages. An alternative solution is applying existing privacy-preserving data publishing (PPDP) techniques, such as partition-based anonymization [6–8], and differential privacy [9–12], to the outsourced healthcare data. However, as we show below, when the following practical requirements are considered, the existing works are not applicable in the context of healthcare data outsourcing.

* Corresponding author at: School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei, China. Tel.:+852 23588766.
E-mail addresses: gswwang@cse.ust.hk, wangwei200606@gmail.com (W. Wang), qianzh@cse.ust.hk (Q. Zhang).

http://dx.doi.org/10.1016/j.comnet.2015.06.014
1389-1286/

- **High-dimensional sensitive healthcare attributes.** In real-world scenario, hospitals and healthcare institutions often collect and maintain many different healthcare attributes (e.g., blood pressure, heart rate) of their patients. We investigate two real-life healthcare datasets owned by an anonymous hospital in Shenzhen, China, and both contain more than 100 attributes. However, due to limited access of high-dimensional healthcare data, most previous PPDP works have focused on low-dimensional datasets, while the case of high-dimensional data has been overlooked. Partition-based anonymization techniques [6–8] usually assume that the data only contain a single sensitive attribute, or support only low-dimensional data due to the *curse of dimensionality*. Differential privacy algorithms [9–12] are designed for data with limited dimensions of sensitive attributes. In the case of high-dimensional sensitive data, differential privacy techniques will inject a huge amount of noise to results, thus, makes the results useless.
- **Personalized protection at attribute level.** Different individuals may have different privacy preferences. For example, some individuals are sensitive about their blood related records, while others may care about skin related records. Existing personalized protection techniques have focused on personalized access control (e.g., attribute-based encryption [13]) or personalized sensitivities of a single dimension [8], while none has investigated personalized sensitivities over multiple data dimensions.
- **Collusion resistance.** In practice, the outsourced CSP and the data users (DUs), e.g., medical practitioners and pharmaceutical researchers, may collude together due to various incentives [4,14]. Under such collusion, the whole dataset stored in the cloud as well as the adopted privacy-preserving scheme will be disclosed. Nonetheless, most existing PPDP approaches [6–12,15] do not consider this collusion.

To satisfy the above practical requirements, we propose a privacy-preserving framework to outsource healthcare data to a *hybrid cloud*. The hybrid cloud [16] consists of a private cloud that keeps sensitive data within hospitals or institutions, and a public commercial cloud that handles the rest of the dataset. Based on this type of cloud, the proposed framework moves the attributes that are insensitive to any individual to the public cloud while keeping the rest on the private cloud. To answer DUs' queries, the private cloud sends sanitized data to the public cloud to compute results.

In this framework, we propose an optimal sanitization protocol to achieve personalized privacy protections for high-dimensional sensitive data with minimal data distortion. To support high-dimensional data, we exploit the merits of both partition-based anonymization and differential privacy. Instead of directly enforcing differential privacy conditions on each sensitive attribute, we first inject differential privacy in the process of partitioning, and then provide attribute-level protection via partition-based anonymization. As such, the randomness injected to ensure differential privacy is only related to the partition process, which is independent of attribute dimensionality. To achieve such hybrid privacy protection, a partition algorithm with minimal distortion is proposed. In the sanitization protocol, the clouds perform data partition to anonymize the whole dataset based on the amount of distortion required for privacy preservation, which is computed on the private cloud. As such, the public cloud does not have access to the sensitive data stored on the private cloud, but obtains the optimal partition results based on the distortion information. To resist collusion, partition selections are randomized to prevent the public CSP and the DUs from gaining extra knowledge from the partition results. Furthermore, we also propose a greedy protocol to reduce the computational cost. We validate the proposed protocols by formal privacy and usability analyses and evaluate their performance using real-world healthcare datasets.

The main contributions of this paper are threefold. First, we propose a privacy-preserving framework for high-dimensional healthcare data outsourcing. To the best of our knowledge, this is the first framework considering high-dimensional sensitive attributes and personalized privacy requirements over different attributes. Second, through formal analytic study, we derive provable privacy guarantees and bounded data distortion achieved by the proposed framework. We show that the proposed framework can defend against the collusion between the public cloud and the DUs while still retaining high usability. Finally, for the first time, we conduct experiments on real-world healthcare datasets with high-dimensional sensitive attributes to validate the proposed framework.

The rest of the paper is organized as follows. Section 3 defines the problem. Section 2 reviews related work. Section 4 overviews the privacy-preserving outsourcing framework and two sanitization protocols, followed by privacy and usability analysis in Section 5. Experimental evaluations are reported in Section 6. Section 7 concludes the paper.

2. Related work

Previous works on privacy-preserving data outsourcing mainly adopt encryption techniques to protect sensitive data [3–5,17]. Yuan and Yu [4] encrypt the biometric database before outsourcing it to the cloud, which can perform *kNN* search in the encrypted database. Li et al. [17] leverage Hierarchical Predicate Encryption to establish a scalable framework for authorized private keyword search on cloud data. Cao et al. [3] enable privacy-preserving multi-keyword ranked search over encrypted cloud data. Nonetheless, these solutions are limited to specific operations, which is not suitable for healthcare data outsourcing that supports a variety of queries. Besides, encryption leads to large overhead when answering queries.

Another brand of privacy-preserving approaches are PPDP techniques. Basically, the works on privacy protection in data publishing can be divided into two categories, partition-based approaches and differential privacy. Many partition-based privacy models are proposed to tackle different privacy concerns. k-anonymity [18] is developed to prevent adversaries with Qausi-Identifier (QI) background knowledge from re-identifying an individual with a probability higher than $\frac{1}{k}$. Fragmentation is used in [15] to break sensitive associations among attributes. Other privacy models consider the privacy attack where adversaries associate an individual with a particular sensitive value. ℓ-diversity [6,19] aims to bound this inference confidence to be no larger than $\frac{1}{\ell}$. (α, k)-anonymity