Future Generation Computer Systems 43–44 (2015) 74–86

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A hybrid solution for privacy preserving medical data sharing in the cloud environment

Ji-Jiang Yang [a,b,*], Jian-Qiang Li [c], Yu Niu [b]

[a] Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, 100084, PR China
[b] Research Institute of Information Technology, Tsinghua University, 100084, PR China
[c] School of Software Engineering, Beijing University of Technology, 100022, PR China

HIGHLIGHTS

- Proposed a hybrid solution for privacy preserving data sharing in cloud environment.
- Different methods are innovatively combined to support multiple paradigms of medical data sharing with different privacy strengths.
- The experimental evaluations are reported based on the implementation of four basic components and a real world case study.

ARTICLE INFO

Article history:
Received 1 January 2014
Received in revised form
25 March 2014
Accepted 1 June 2014
Available online 10 June 2014

Keywords:
Privacy protection
Cloud storage
Integrity check
Medical data sharing

ABSTRACT

Storing and sharing of medical data in the cloud environment, where computing resources including storage is provided by a third party service provider, raise serious concern of individual privacy for the adoption of cloud computing technologies. Existing privacy protection researches can be classified into three categories, i.e., privacy by policy, privacy by statistics, and privacy by cryptography. However, the privacy concerns and data utilization requirements on different parts of the medical data may be quite different. The solution for medical dataset sharing in the cloud should support multiple data accessing paradigms with different privacy strengths. The statistics or cryptography technology alone cannot enforce the multiple privacy demands, which blocks their application in the real-world cloud. This paper proposes a practical solution for privacy preserving medical record sharing for cloud computing. Based on the classification of the attributes of medical records, we use vertical partition of medical dataset to achieve the consideration of different parts of medical data with different privacy concerns. It mainly includes four components, i.e., (1) vertical data partition for medical data publishing, (2) data merging for medical dataset accessing, (3) integrity checking, and (4) hybrid search across plaintext and ciphertext, where the statistical analysis and cryptography are innovatively combined together to provide multiple paradigms of balance between medical data utilization and privacy protection. A prototype system for the large scale medical data access and sharing is implemented. Extensive experiments show the effectiveness of our proposed solution.

1. Introduction

Healthcare covers complex processes of the diagnosis, treatment, and prevention of disease, injury, and other physical and mental impairments in humans. The patients' consumption of products and services provided by hospitals and other institutions forms the healthcare industry, which is one of the largest and fastest-growing part of a country's economy. It is widely accepted that the high quality healthcare services lie in the effectiveness and efficiency of health problem detection, innovative solution identification, and medical resource allocation [1,2], which in turn depend heavily on the proper collection, management and utilization of health information [3]. Considering the fact that the health information collection and utilization may be distributed in multiple organizations, the medical data sharing plays the critical role in enabling the medical information flow across these organizations and then improving the quality of healthcare services.

* Corresponding author at: Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, 100084, PR China. Tel.: +86 1062788788 13.
E-mail addresses: yangjijiang@tsinghua.org.cn, cemon_yang@163.com (J.-J. Yang), lijianqiang@tsinghua.org.cn (J.-Q. Li).

http://dx.doi.org/10.1016/j.future.2014.06.004
0167-739X/

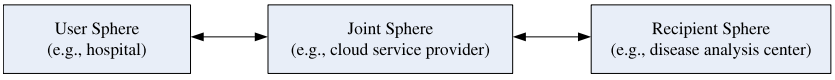

Fig. 1. The usage scenario of remote data storage for sharing.

With the emergence of cloud computing technologies, the connectivity allowed by the Internet is exploited to make users have the ability to utilize scalable, distributed computing environments. In the cloud, computing resources including storage is provided by a third party service provider [4,5]. Since data in the cloud typically resides in a shared environment, users will know neither the exact location of their data nor the other sources of the data collectively stored with theirs [6–8]. Therefore, in this Internet-based computing paradigm, users are universally required to accept the underlying premise of trust [7,2].

With healthcare providers looking at automating processes of health information manipulation at lower cost and higher gains, cloud computing has been viewed as an appropriate platform to deploy standard medical information systems for its scalable and cost-effective services delivered by cloud service providers [1,2]. However, despite the increased use of cloud-based data sharing platforms, the privacy related problems have prevented their adoption in the healthcare domain [6,7].

The diverse security and privacy concerns surrounding medical data has been studied widely over the last few years. Many organizations have published their reports [9–11] on the security and privacy issues for the manipulation of medical data in the networked systems. According to [12], the most widely used regulations are the Health Insurance Portability and Accountability Act (HIPAA) and the European Data Protection Directive 95/46/EC. In these regulations, there are two fundamental issues for the privacy of medical data sharing, i.e., privacy protection during transmission and privacy protection of the stored data. The former has been studied widely and addressed by the Secure Socket Layer protocol (SSL) [13] and Transport Layer Security (TSL) protocol [14]. The latter is less studied and of greater relevance to storage as a service [7] in the cloud computing paradigm, where the outsourced data is stored on the site of the cloud service provider.

As a typical usage scenario shown in Fig. 1, the privacy protection of the stored data involves three spheres [15], i.e., user sphere, joint sphere, and recipient sphere. Here, a hospital outsources the storage of the shared medical records to the cloud service provider and publishes these datasets to a medical research center. In this example, the hospital serves as the data owner, the medical research center is the data recipient, and the joint sphere encompasses the hardware and software in the site of cloud service provider to supply the outsourced storage service of the shared medical records.

This paper considers privacy protection problem of the stored medical data in the joined sphere and proposes a practical solution to the storage service provider for preserving the privacy of the user data.

In the research literature on privacy protection, there are three fields of work with seemingly very different goals. The first field is about privacy by policy [16–22,15], by assuming policies and regulations are generally enforceable and the role of technology is to aid enforcement, these researches aims at protecting user data from accidental disclosure or misuse and facilitating informed choice options. Since the proposed solutions cannot guarantee the enforcement of the policies, they only provide limited capabilities of the privacy protection [23–26]. For example, in the case of medical data sharing in the cloud environment, the employees in the cloud service provider can easily obtain the assignment of a specific role with the authorization to access the medical data beyond their privileges. The second field is about privacy by statistics. The researchers in this field are adopting analytical/statistical technologies to construct finely tuned information disclosure mechanisms. The proposed solutions generally assume one or several specific attacked models and limited data mining behaviors and has strong guarantee only regarding to the assumed attack model (it is vulnerable for a practical solution, since the adversary could have public information and background knowledge). The third field is about privacy by cryptography [27–34]. Under the assumption that sophisticated adversaries will not be deterred by policies or regulations, which might be only theoretically doable, the researchers in this field aimed at developing cryptographic privacy protections and systems with provable privacy guarantees. Through wisely applied combinations of cryptography and data security protocols, the proposed solution can provide strong guarantee on privacy protection, but achieving this goal generally undermines usability of the target data and then of the solutions [15].

In this paper, we describe a hybrid solution for privacy preserving medical data sharing in the cloud, where the technologies of statistical analysis and cryptography are innovatively combined together to support multiple paradigms of medical data sharing with multiple levels of privacy strength. More specially, multiple enforceable mechanisms are supplied to prevent the adversary from attacking the privacy of the shared medical dataset when role-based access models are breached. Considering the fact that the privacy concerns and data utilization requirements on different parts of the medical data may be quite different, the attributes of medical records are classified into multiple categories. Accordingly, the original medical dataset is vertically partitioned into three parts for the remote storage in the cloud. The part that can be used for patient identification is stored in ciphertext. The other parts that will be used for medical analysis is stored in plaintext. The functionalities of data merging and integrity checking are provided for the user to retrieve the medical dataset and correctness verification, respectively. In addition, a hybrid approach to search across the ciphertext and plaintext is given to support the record-level utilization of the medical data. The integration of multiple technologies on privacy protection and cloud data manipulation makes the proposed solution applicable in real-world cloud for the better balance of the information utilization and privacy protection.

The contribution of this paper can be summarized as follows: (1) a pilot study is described for the privacy preserving medical data sharing in the cloud environment, where the advantages of statistics and cryptography technologies are combined together for the better balance between information utilization and privacy protection. (2) By borrowing the ideas of existing technologies, three novel approaches, respectively on vertical data partition handling anonymization of EMR data with text-type sensitive attributes, probabilistic remote integrity checking with high efficiency, and the query processing across both plaintext and ciphertext EMR data, are proposed. (3) The empirical experiments of privacy preserving medical data sharing on a real world regional healthcare collaboration platform are reported, which demonstrate the effectiveness of the proposed approaches. We believe that this pilot study can serve as an important reference for the development of similar real world systems.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 describes first the high level framework of the proposed practical solution for privacy preserving medical data sharing, and then the technical details of the four basic components on vertical data partition, data merging, integrity checking, and hybrid search, respectively. In Section 4, a case study on the implementation and performance evaluation of the proposed solution on a real-world cloud of regional healthcare collaboration platform is reported. Section 5 concludes the paper.