ARTICLE IN PRESS

Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Aggregating privatized medical data for secure querying applications

Kalpana Singh *,[1], Lynn Batten

Deakin University, 221 Burwood Highway Burwood, VIC 3125, Australia

HIGHLIGHTS

- Proposes solutions for the aggregation and data querying of sensitive data.
- Delineates applications of aggregation of sensitive medical data.
- Introduces an efficient diagonal data aggregation method.
- Presents a method for privately and efficiently querying the aggregated data.
- Data service manager is untrusted beyond seeing the privatized contributed data.

ARTICLE INFO

Article history:
Received 14 November 2015
Received in revised form
12 November 2016
Accepted 22 November 2016
Available online xxxx

Keywords:
Data privacy
Data aggregation
Data querying

ABSTRACT

Public and private organizations generate large amounts of data which they are happy to allow others to query as long as it is privatized. (One example is that of medical data which can be used for research purposes.) Aggregation of such data on a cloud provides an opportunity for querying over rich data. This paper provides a solution for sharing sensitive data where large numbers of data contributors publish their privatized data sets which are then aggregated by a cloud manager on a cloud so that data can be made available to anyone who wants to query it. Additionally, our solution determines how aggregated data can be efficiently and effectively queried, while retaining privacy not only of the data, but also of the original data owner, the query and the person querying. We introduce a non-standard diagonal data aggregation method and, by experimental testing, demonstrate that our data querying procedure is efficient, maintains acceptable data privacy and acceptable data utility, along with practical computation and storage costs. Our solution also accepts a number of varied queries including join, aggregate, range, nested, ordered by and pattern matching. Finally, we discuss four potential threats posed by our cloud manager against which our scheme is resistant.

1. Introduction

The importance of data sharing in the international context of global issues such as health, environmental change, and food production, is amplified by projects such as the International Cancer Genome Consortium (https://icgc.org/icgc/goals-structure-policies-guidelines/b-consortium-goals).

The prodigious amount of data accumulated by science and business needs to be aggregated in order to extract information and gain knowledge. Such data sets are often a result of the systematic collection of published data from multiple sources and are eventually transmitted to a cloud server on which efficient processing is required to produce high-level, high-quality information. Given the sensitive nature of much data and the varying social and legal implications for its disclosure, privacy is a major concern when sharing data. In order to prevent disclosure of individually identifiable information, usually only de-identified data sets are shared. De-identification is implemented by means of privacy preserving data mining methods [1]. Several major challenges face those wishing to aggregate data for the purposes of data sharing and querying. One challenge is to obtain an aggregated data set which achieves acceptable privacy and utility levels; a second is maintaining practical storage and computation costs. As public data sets are not under the data provider's control, data confidentiality and integrity are of concern in outsourced databases. In order to protect sensitive data sets, the primary way to make these secure is to privatize the data before sharing [2–4]. Once the data have been appropriately aggregated, a cloud service manager hosts the data of its clients and provides a variety of data

* Corresponding author.
E-mail addresses: kalpana.singh@cea.fr (K. Singh), lynn.batten@deakin.edu.au (L. Batten).
[1] DRT/LIST/DACLE/SCSN/L3S, Commissariat à l'Energie Atomique, NanoInnov Centre de Saclay, 91191 Gif-sur-Yvette Cedex, France.

http://dx.doi.org/10.1016/j.future.2016.11.028
0167-739X/

Please cite this article in press as: K. Singh, L. Batten, Aggregating privatized medical data for secure querying applications, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.11.028

management functionalities, including modifications and queries. A third challenge is ensuring that the server implements querying operations correctly while not having access to the identifiable information.

Requirements of Privacy Preserving Querying Services: On a daily basis, people query public online data services such as search engines, social network sites and news portals. While users need such public data services, they may also be concerned that their personal information could be disclosed or compromised. User queries can be revealed intentionally to advertisers (in some cases without the user's knowledge) such as in some of the Google and Facebook applications (http://news.cnet.com/2702-1009_3-986.html). In our proposed data aggregation system, we enable private querying on public data services so that the contents of user queries and their replies are hidden from the service manager.

Private Data Querying Methods: A great deal of work has been done in the area of private query processing [5,4,6]. We review existing privacy preserving data querying methods in Section 2.2. Query processing protocols on encrypted data sets stored on a cloud have been extensively studied (eg. [7]), while query processing that preserves both the data privacy of the data providers along with the query privacy of the data requester is a relatively new research area.

Contributions of this paper. In this paper, we propose an entirely new approach to aggregate data, which is known as the "diagonal data aggregation" method. We experimentally and comparatively analysed our proposed diagonal data aggregation method with the most popular horizontal and vertical methods in Section 5.1, and we find that our proposed method provides better efficiency in terms of data modification operations with acceptable data privacy and acceptable data utility. We propose a method of data aggregation and data querying which achieves high levels of privacy and utility with acceptable cost, and performs data modification efficiently. However, there is a trade-off for this improvement: increased data storage costs. The results of our work are presented in Tables 2–6 of Section 5. We show that the proposed solution provides privacy preserving data querying processes, for several query types, with low computation and communication cost in Tables 5 and 6 of Section 5.2. Our proposal also supports data modification and credential revocation processes.

Our main contributions are (i) computationally efficient data aggregation and data querying procedures (Tables 3 and 6) and (ii) aggregation and querying procedures in which the cloud service manager has no access to the original data. Additionally, our system supports several types of queries (Section 4.4) over public data sets as well as managing data updates and credential revocation.

The paper is designed as follows. In Section 2, we review the current literature relative to the challenges mentioned in this Introduction; in particular, we list the recent work on private data aggregation methods and querying methods in Sections 2.1 and 2.2 respectively. We identify the gaps for our work in Section 2.3 and provide solutions to fill those gaps in Section 2.4. Sections 3 and 4 provide descriptions of the workflow processes between the components of our architecture. In the experimental Section 5, we aggregate four data sets using our method and then demonstrate the querying process on the result. A comparison with other research work indicates that our aggregation method is more computationally efficient than others (Table 3) and that our data querying method for varied queries is also more computationally efficient than others (Table 6). Tables 2 and 5 present time needed in communication of a DC and the DSM, and a DR and the DSM respectively. In Section 6, we demonstrate the prevention of several insider attacks potentially made from the data server. Table 7 indicates that our architecture provides better protection from insider threats than do several recent papers. Section 7 summarizes and presents conclusions.

2. Current solutions on privacy preserving data aggregation and data querying

This section presents previous work on sharing public data sets, and is divided in two Subsections. The first Subsection presents the literature on privacy preserving data aggregation methods on distributed data sets. The second Subsection provides the existing work on privately data querying on public data sets.

2.1. Privacy preserving data aggregation

In the recent literature, aggregation protocols are proposed based on cryptographic solutions and partitioning methods. The most common cryptography protocols are homomorphic encryption, identity-based proxy re-encryption, symmetric searchable encryption and an asymmetric searchable encryption scheme following [8,9], along with improved approaches [10–12]. Due to high communication costs, long delays, high computation and storage costs, these solutions are impractical, in particular, increasing the costs of storing and transmitting ciphertexts, in situations such as shared cloud storage. There is other existing work (e.g. [13]), which focuses on security and privacy preserving data aggregation, but most of it assumes a trusted aggregator and cannot protect user privacy against untrusted aggregators. Most of the work is based on the partitioning of a table, vertically [14,15] or horizontally [14,16,17] into several parts for aggregation. These methods cannot provide data modification operations efficiently.

We note here, that in this paper, we examine for the first time a method of diagonal data aggregation (Section 5) and demonstrate that it is a major factor in improved computational efficiency both for aggregation and querying.

2.2. Privacy preserving data querying

The problem of private query processing addressed by the research on Private Information Retrieval (PIR) [18] is to provide a user with the means to retrieve data from a database without the database learning any information about the particular item that was retrieved. Private Block Retrieval (PBR) is a natural and practical extension of PIR in which, instead of retrieving only a single bit, the user retrieves a block of bits from a single database. Since existing PIR methods are criticized as impractical for their expensive computation and communication costs in [19,6] in 2005, Gentry and Ramzan [19] extended the single-database PIR to a more efficient PBR. In our approach, we also use a PBR method, but choose a relatively low-cost adaptation of the one in [20].

Several alternative methods have been proposed in the literature to execute data querying privately such as query anonymization based on *k*-Anonymity [21,22] and Plausibly Deniable Search (PDS) [23]. However, according to [24,25,6], they do not provide acceptable data privacy and practical performance of data querying. The bucket based method [5] was proposed for efficient processing of the requested data set from encrypted data sets in the database. A fatal drawback of the bucket based index algorithm is the risk of data exposure, which is analysed in the paper [26]. From the presented literature on data aggregation and data querying, we identified research gaps for our work in Section 2.3.

2.3. Identified gaps for our work

We observe the possible research direction for our work from the identified shortcomings on data aggregation and data querying in the literature review in Sections 2.1 and 2.2 which respectively identify two main areas upon which we focus in this paper. These are: the fact that data modification operations are not efficiently

Please cite this article in press as: K. Singh, L. Batten, Aggregating privatized medical data for secure querying applications, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.11.028