Information Sciences 436–437 (2018) 320–342

Contents lists available at ScienceDirect

Information Sciences

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

Outsourcing scalar products and matrix products on privacy-protected unencrypted data stored in untrusted clouds

Josep Domingo-Ferrer*, Sara Ricci, Carles Domingo-Enrich

Universitat Rovira i Virgili, Department of Computer Science and Mathematics, UNESCO Chair in Data Privacy, Av. Països Catalans 26, 43007 Tarragona, Republic of Catalonia

ARTICLE INFO

Article history:
Received 22 December 2016
Revised 16 December 2017
Accepted 13 January 2018

Keywords:
Privacy
Cloud computing
Data splitting
Scalar product
Matrix product
Honest-but-curious clouds

ABSTRACT

Data controllers accumulate more and more data on people, of which a substantial proportion are personally identifiable and hence sensitive data. Storing and processing those data in local premises is increasingly inconvenient, but resorting to cloud storage and processing raises security and privacy issues. We show how to use untrusted clouds to compute scalar products and matrix products on privacy-protected data stored in them. These operations are useful in statistics, linear algebra, data analysis and engineering. In our solutions, the privacy-protected sensitive data stored in the clouds are not encrypted, but preserve some utility (that is, some statistical properties) of the original data. We consider two variants of honest-but-curious clouds: clouds that do not share information with each other and clouds that may collude by sharing information with each other. In addition to analyzing the security of the proposed protocols, we also evaluate their performance against a baseline consisting of downloading plus local computation.

1. Introduction

With the advancement and spread of computation and communication technologies, the amount of data collected and stored by private and public sectors is constantly increasing. Storing and processing such huge amounts of information in local premises becomes very problematic, due to soaring costs of software, hardware, energy and maintenance. In this context, the need emerges to find a fast and cost-effective alternative. An attractive possibility for a data controller is to outsource storage and processing to a cloud [2]. Such outsourcing brings several benefits like elimination of infrastructure costs (no software/hardware investments needed), flexibility (storage and computing power can scale depending on business growth) and energy savings.

Unfortunately, storing and processing data in the cloud has also downsides related to security and privacy. A lot of the information being collected is personally identifiable and therefore sensitive. Neither the data controller nor the subjects to whom the data refer want the cloud service provider (CSP) to read, use or sell their data.

In this article we discuss several procedures to store *and process* sensitive data in a privacy-preserving way in untrusted clouds, where processing consists of two basic operations: scalar products and matrix products. These operations are useful in statistics and data analysis (to compute correlations between attributes, contingency tables, etc.), and also in engineering

* Corresponding author.
E-mail addresses: josep.domingo@urv.cat (J. Domingo-Ferrer), sara.ricci@urv.cat (S. Ricci), carles.domingo@estudiant.upc.edu (C. Domingo-Enrich).

https://doi.org/10.1016/j.ins.2018.01.024
0020-0255/

(image encryption, 3D graphics simulation, etc.); see Section 1.2 of [22] and references therein. In our solutions, *the privacy-protected sensitive data stored in the clouds are not encrypted* and *preserve some of the utility (that is, some statistical features) of the original data*. This allows making the most of the outsourced data, while ensuring that no original records can be re-created from the outsourced records. The outsourced data can be used for purposes other than computing scalar products or matrix products. This is a relevant difference with respect to related work on algebraic computation outsourcing (see Section 3.2).

Specifically, the outsourced data preserve means and standard deviations of attributes for the entire data set and even in subdomains of it. In some of our protocols, we use vertical splitting, so that each cloud stores a cleartext fragment on which any statistical analysis can be directly performed by the cloud. In the rest of our protocols, the cloud stores synthetic or anonymized versions of the original data set that preserve the above mentioned statistics. The goal is that the outsourced version retains some utility for exploratory analysis by any user with direct access to the cloud-stored data (who should however be unable to reconstruct the original data). Note that there are many organizations interested in releasing privacy-protected/anonymized data for secondary analysis, including but not limited to official statistics [19].

Following the architecture defined in the "CLARUS" European H2020 project [9] (within which this work has been carried out), we will assume a proxy located in a domain trusted by the data controller (e.g., a server in her company's intranet or a plug-in in her device) that implements security and privacy-enabling features towards the cloud service providers. We will call this trusted proxy CLARUS.

1.1. Contribution and plan of this paper

In [5], we evaluated some non-cryptographic proposals for secure scalar product and secure matrix product on vertically split data presented in the literature. We then proposed and enhanced some protocols adapted to the CLARUS scenario. The CSPs were assumed to be honest-but-curious and not sharing information with each other.

Here, we start from the conclusions of that previous work and break new ground by: (i) exploring new non-cryptographic protocols for the scalar product on split data; (ii) considering also cryptographic protocols to compute on split data; and (iii) relaxing the non-sharing assumption. In the cryptographic protocols we use encryption only in the communication between clouds; however, the sensitive data stored in the clouds are protected by splitting, not by encryption. Regarding the sharing assumptions, we first assume that the CSPs do not pool their fragments to reconstruct the original data set; we start from two existing protocols for this setting, one without cryptography and one using cryptography, and we present two new non-cryptographic protocols and two variants of the cryptographic protocol. We then relax the non-sharing assumption and present two additional non-cryptographic protocols that are sharing-resistant, even though they require substantial cloud storage (because they rely on data replication rather than splitting).

This paper is organized as follows:

- Section 2 presents the CLARUS architecture and the security models considered in the rest of the paper for the untrusted CSPs: (i) honest-but-curious CSPs not sharing information with each other; (ii) honest-but-curious CSPs that may collude by sharing information with each other, but that lack side knowledge on the original data set; (iii) honest-but-curious CSPs sharing information with each other and having side knowledge on the original data set.
- Section 3 reviews background on data splitting and related work on algebraic computation outsourcing and secure scalar products.
- Section 4 focuses on secure scalar products on vertically partitioned data when CSPs are honest-but-curious and do not share information: a non-cryptographic protocol and a cryptographic protocol are reviewed, and then two new non-cryptographic protocols and two variants of the cryptographic protocol are presented. After that, we show how secure scalar products between pairs of clouds can be combined with computations involving a single cloud to perform data analyses such as correlations and contingency tables.
- Section 5 presents a new protocol that can resist information sharing between CSPs but assumes the CSPs have no side knowledge about the original data set; rather than computing individual scalar products, this protocol computes a matrix product $\mathbf{X}^T\mathbf{X}$ of an original data set $\mathbf{X}$ of which the cloud only knows a masked version $\mathbf{Y}$. Just like the scalar products allowed computing the data set correlation matrix, so does the above matrix product; furthermore, the clouds can also be used to compute the means and the standard deviations of attributes in $\mathbf{X}$.
- Section 6 proposes another new sharing-resistant protocol to compute the matrix product $\mathbf{X}^T\mathbf{X}$ that involves heavier computation but which stays safe even if the CSPs have information on the statistical structure of the original data set $\mathbf{X}$.
- In Section 7, the computation and communication costs of all protocols described in Sections 4–6 are assessed and compared against a benchmark protocol consisting of the CLARUS proxy downloading the entire data set and locally computing on the downloaded data set.
- Section 8 presents the experimental results obtained by implementing the proposed protocols in a multi-cloud scenario.
- Finally, Section 9 lists some conclusions and future research lines.

The Appendix contains mathematical technicalities related to the cryptographic protocol in Section 4 and the sharing-resistant protocol in Section 5.