Information Sciences 265 (2014) 176–188

Contents lists available at ScienceDirect

Information Sciences

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

Search pattern leakage in searchable encryption: Attacks and new construction

Chang Liu, Liehuang Zhu *, Mingzhong Wang, Yu-an Tan

Beijing Engineering Research Center of Massive Language Information Processing and Cloud Computing Application, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Article history:
Received 6 August 2013
Received in revised form 4 November 2013
Accepted 19 November 2013
Available online 27 November 2013

Keywords:
Search pattern
Searchable encryption
Cloud computing
Fake query
Grouping-based construction

ABSTRACT

Searching on remote encrypted data (commonly known as *searchable encryption*) has become an important issue in secure data outsourcing, since it allows users to outsource encrypted data to an untrusted third party while maintains the capability of keyword search on the data.

Searchable encryption can be achieved using the classical method called oblivious RAM, but the resultant schemes are too inefficient to be applied in the real-world scenarios (e.g., cloud computing). Recently, a number of efficient searchable encryption schemes have been proposed under weaker security guarantees. Such schemes, however, still leak statistical information about the users' search pattern.

In this paper, we first present two concrete attack methods to show that the search pattern leakage will result in such a situation: an adversary who has some auxiliary knowledge can uncover the underlying keywords of user queries. To address this issue, we then develop a grouping-based construction (GBC) to transform an existing searchable encryption scheme to a new scheme hiding the search pattern. Finally, experiments based on the real-world dataset demonstrate the effectiveness of our attack methods and the feasibility of our construction.

1. Introduction

Cloud computing has been increasingly applied to outsource data by cloud tenants to diminish the high overhead of data storage and management. In many cases, cloud tenants need to encrypt data before outsourcing them to prevent cloud administrators from accessing sensitive information, such as government documents, hospital records and personal emails. However, data encryption invalidates many data query functions, which will lead to inefficient data utilization. For instance, the cloud service provider cannot respond keyword search query over encrypted data. The purpose of searchable encryption is allowing a user to outsource data to a third party in a secure manner, and then retrieve documents containing the queried keyword. Therefore, searchable encryption plays an important role in cloud computing scenario [16].

Index, an auxiliary structure that accelerates the search process, has been widely studied in information retrieval fields. In general, each entry of the index is formed as a ⟨keyword, document identifiers⟩ tuple, so that all the documents containing the queried keyword can be easily located. We describe the general scenario of searchable encryption as follows: a data owner (e.g., Alice) has a set of documents to outsource to an untrusted third party (e.g., Carol). Alice first builds an index of all the keywords appeared in the documents, and then encrypts both the documents and the index. After that, she outsources the encrypted documents and index to Carol. An authorized data user (e.g., Bob) has a secret key, so that he is able to generate keyword search queries (a.k.a. trapdoors or tokens) by calling a trapdoor function. Once Carol receives a query

* Corresponding author.
E-mail addresses: changliu.bit@gmail.com (C. Liu), liehuangz@bit.edu.cn (L. Zhu), wangmz@bit.edu.cn (M. Wang), victortan@yeah.net (Y.-a. Tan).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.11.021

from Bob, she can search in the index and return the (encrypted) search result[1] to Bob. Then Bob is able to decrypt the search result and retrieve needed documents. Note that if the secret key is shared appropriately (such as the Multi-user SSE scheme in [11]), there can be multiple data users. For example, all employees of a company shared a secret key to perform keyword search on their data.

It has been widely accepted in the literature that both outsourced data and search query should leak as little information as possible to the third party. We note that searchable encryption can be achieved using the classical method called oblivious RAM [19,14], which attains the optimal security (i.e., nothing leaked to the third party). However, this kind of approaches is impractical due to the poly-logarithmic computation and communication overheads. Therefore a number of searchable symmetric encryption (SSE) schemes (e.g., [22,13,9,11,23,18,17]) have been proposed under weaker security guarantees for efficiency. However, the access pattern and the search pattern are leaked in their schemes. Informally, the access pattern is the information about which documents contain the queried keyword (i.e., the search result) for each of the user queries, while the search pattern is the information about whether any two queries are generated from the same keyword or not. In recent studies, [15] has discussed a concrete attack exploiting the access pattern to disclose the underlying keywords of user queries. However, the potential risks of the search pattern leakage are rarely studied in the literature. As it is noted in [17], a limitation of most known SSE schemes is the leakage of the search pattern. In fact, this limitation also exists in searchable asymmetric encryption schemes (e.g., [8,6]). For the reasons above, we are motivated to investigate on the search pattern leakage issue and develop a new searchable encryption construction under more rigorous security requirements.

1.1. Intuitions

Let's begin with analyzing why search pattern is leaked in SSE schemes. To the best of our knowledge, the query algorithms of existing SSE schemes in the literature are mostly deterministic, which means the same keyword will always generate the same query. In this sense, an adversary can easily judge whether any two queries are generated from the same keyword or not, so as to obtain the users' search pattern. One may apply probabilistic query algorithms to solve this problem. However, simply making use of probabilistic query algorithms [21] still cannot hide the search pattern, because the entry touched in each search process discloses the search pattern as well. In other words, for the same keyword, its corresponding entry in the index must be the same, so that the adversary can disclose users' search pattern just by observing the entries touched in the index during the search process. For this reason, the search pattern is also leaked in searchable asymmetric encryption (a.k.a. public key encryption with keyword search, PEKS), whose query algorithms are probabilistic [8,6,20,12]. Therefore, randomizing query algorithm only contributes to defend outer adversaries but not inner adversaries (e.g., cloud administrators). It is also worthwhile to point out that the access pattern might leak the search pattern in the same way (i.e., the same search result might mean the same queried keyword). We refer the reader to [15] for the details on how to hide the access pattern. The work of this paper focuses on the search pattern leakage caused by the queries and index. A delicate approach to hide the search pattern is launching several fake queries along with the real query, so that the adversary cannot identify the real query. Based on this idea, we develop a new searchable encryption construction where the number of subqueries (i.e., the real query and the fake queries) for each query is parametrized by a confusion parameter k. We will give detailed description in Section 5.

1.2. Our contributions

We outline the contributions of this paper as the following:

1. We address the search pattern leakage issue and demonstrate its potential risks in the practical applications by giving two concrete attack methods. In particular, with our attack methods, an adversary who has obtained users' search pattern can effectively uncover the underlying keywords of the user queries under the help of some public available knowledge.
2. We present a grouping-based construction (GBC) which transforms an existing index-based searchable encryption scheme to a new scheme hiding the search pattern. GBC is designed to be independent from the underlying searchable encryption scheme, so that most previous schemes [13,9,11,23,18,17] can be used in GBC.
3. We prove that the resultant scheme of GBC satisfies a stronger security guarantee than any existing searchable encryption scheme. GBC reduces the search pattern leakage to the group pattern leakage.
4. Based on the real-world dataset [5], we test the performance of the proposed attack methods and the proposed construction. The experiment results indicate the effectiveness and feasibility of proposed attack methods and construction.

1.3. Organization of the paper

The remainder of the paper is organized as follows: Section 2 briefly surveys the motivations of hiding the search pattern. Section 3 introduces some preliminaries. We formalize two attack methods in Section 4. In Section 5 we will describe the

[1] The search result is a collection of document identifiers whose corresponding documents contain the queried keyword.