Privacy-Preserving Keyword Search Schemes over Encrypted Cloud Data: An Extensive Analysis

Prasanthi Sreekumari
Department of Computer Science,
Grambling State University, Grambling, LA, USA
Email: sreekumarip@gram.edu

Abstract—Big Data has rapidly developed into a hot research topic in many areas that attracts attention from academia and industry around the world. Many organization demands efficient solution to store, process, analyze and search huge amount of information. With the rapid development of cloud computing, organization prefers cloud storage services to reduce the overhead of storing data locally. However, the security and privacy of big data in cloud computing is a major source of concern. One of the positive ways of protecting data is encrypting it before outsourcing to remote servers, but the encrypted significant amounts of cloud data brings difficulties for the remote servers to perform any keyword search functions without leaking information. Various privacy-preserving keyword search (PPKS) schemes have been proposed to mitigate the privacy issue of big data encrypted on cloud storage. This paper presents an extensive analysis of the existing PPKS techniques in terms of verifiability, efficiency and data privacy. Through this analysis, we present some valuable directions for future work.

Keywords—Big Data; Cloud Storage; Encryption; Verifiability; Efficiency; Security

I. INTRODUCTION

Cloud computing has become an unavoidable component for data storage due to its unlimited data storage capabilities, easy access and decreased costs. With the rapid development of cloud computing, individuals and enterprise users prefer cloud storage services to outsource large volume of documents to reduce the overhead of storing data locally. However, the privacy of these data on cloud storage is a major concern. This is because once data are outsourced, the owners lose direct control of these data and the cloud service providers cannot be fully trusted [1] [2]. There are several reports that confirm data breaches related to cloud servers, due to malicious attack, theft or internal errors [3]. As a result, to protect the privacy of data, users/organizations encrypt the data before outsourcing to remote servers. However, the encrypted significant amounts of cloud data brings difficulties for the remote servers to perform any keyword search on encrypted cloud data without leaking information.

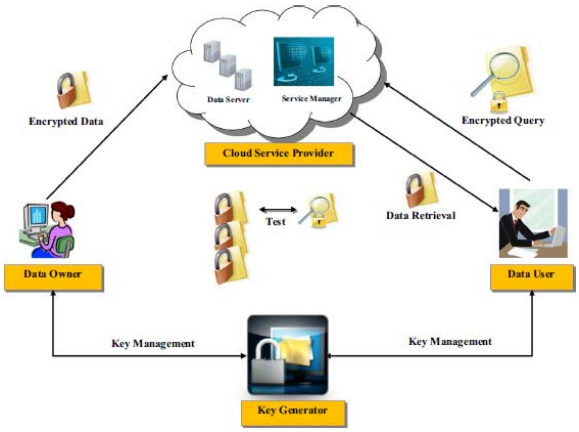

Fig. 1. Searchable encryption scheme

Keyword search is one of the fundamental and most frequent data operations [2]. It enables the user to search for a certain keyword on the cloud data [3][4]. To provide a secure and efficient retrieval of data, one needs to ensure that the user can perform a search over the encrypted data without revealing the information to the server. The cryptographic primitive that provides this feature is widely known as searchable encryption [3][5][6]. Searchable encryption enables the users to generate a search token from the searched keyword in such a way that the cloud server can retrieve the encrypted contents containing the searched keyword [3].

As shown in Fig.1 [3], a data owner has a collection of documents to outsource and generates searchable encrypted indexes for outsourced data to protect the data on the cloud storage. The user can generate some search encrypted keywords based on the

978-1-5386-4399-0/18/$31.00 ©2018 IEEE
DOI 10.1109/BDS/HPSC/IDS18.2018.00035

granted key from the data owner and keywords based on the interest of data owner, meanwhile, it is required that nothing should be leaked from the encrypted keywords, indexes or files or keywords of search queries. Although traditional searchable encryption schemes could achieve the guarantee of security and efficiency, they may fail to protect the privacy of cloud data [4][5]. To mitigate the issue of data privacy, privacy-preserving keyword search techniques have been proposed. In this paper, we presents an extensive analysis of the existing PPKS techniques in terms of privacy, verifiability and efficiency. Through analysis, we present some valuable directions for future work.

The remainder of this paper is organized as follows: Section II discusses about big data in cloud computing. Section III discusses about PPKS. We introduce the existing research related to PPKS in section IV. Section V compares and analyze the existing solutions. Finally, section VI draws conclusions of our work and present some directions for future work.

II. BIG DATA IN CLOUD COMPUTING

Storing and processing big volumes of data requires scalability, fault tolerance and availability [9]. The traditional infrastructure of storing and managing data is now proving to be slower and not easy to manage. Cloud computing delivers all the essential requirements for storing big data through hardware virtualization. Thus, big data and cloud computing are two compatible concepts as cloud enables big data to be available, scalable and fault tolerant. Several new companies such as Cloudera, Hortonworks, Teradata and many others, have started to focus on delivering big data as a Service or Database as a Service [6]. Companies such as Google, IBM, Amazon and Microsoft also provide ways for consumers to consume big data on demand.

However, in the high speed connectivity era, moving large datasets on cloud and providing the details needed to access it, is a current issue. Because, these large sets of data often carry sensitive information like credit or debit card numbers, addresses, medical records and other details, raising data privacy concerns [10]. In order to ensure data privacy, the data can be encrypted at the client side before outsourcing the data to the cloud server. However, once the data is encrypted, the server cannot make a plaintext keyword search

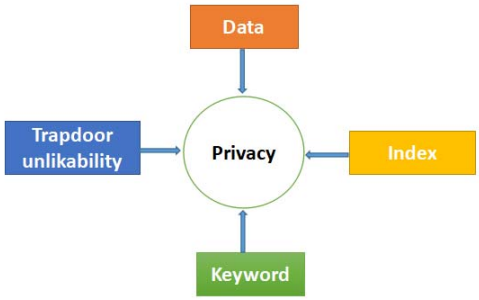

Fig.2. Requirements for Protecting Privacy

and it affects the efficient retrieval of data using keyword search from the cloud [10][11]. For mitigating this issue, researchers proposed efficient privacy-preserving keyword search solutions. We present the recently proposed solutions in section V.

III. PRIVACY-PRESERVING KEYWORD SEARCH

To maintain data privacy, the keyword search functionality need to be performed over encrypted cloud data without leaking any information about the search keyword or the retrieved document. This is known as privacy-preserving keyword search [12]. As shown in Fig.2., following are the list of some requirements for protecting the privacy of users from cloud server.

- **Data privacy-** We protect the outsourced data without leaking any confidential information to unauthorized party from the cloud server.
- **Index privacy**- We must guarantee that the third party cannot steal the information from index stored on cloud.
- **Keyword privacy**- The cloud server should not get any details about data collection, index and encrypted keywords.
- **Trapdoor unlinkability**- We should randomize the trapdoors (encrypted keywords) for protecting the privacy of data. And we also should make each trapdoor generated for query is different. The cloud cannot infer the relationship between these trapdoors.

IV. PRIVACY-PRESERVING KEYWORD SEARCH SCHEMES

This section surveys the most recently proposed privacy-preserving keyword search schemes.

Jiang et al. [13] proposed a novel privacy preserving keyword search scheme over encrypted cloud data. The authors firstly use a structure named as Inverted Matrix (IM) to create search index. The IM is made up of a number of index vectors which are sub-indices for distinct words in the data set. Specifically, during the index construction, each keyword is associated with an index vector denoted by binary bits, where each bit represents whether the keyword appears in the corresponding document. Then, map a keyword to a value as an address used to locate the corresponding index vector. As a result, users can avoid the overhead of storing a dictionary locally.

In addition, to preserve users' privacy, the authors blind the index vectors using pseudo-random bits to obtain an Encrypted Enlarged Inverted Matrix (EEIM) which can prevent the server from learning information from the index. Therefore, when a search query is submitted, the server just needs to find the matching index vector via the inputted valid request. Users only need a single interaction with the server to get the data files they are interested in.

Zhang et al.[14] proposed PRMSM, a privacy preserving ranked multi-keyword search protocol in a multi-owner cloud model as shown in Fig.3 [4]. To enable cloud servers to perform secure search without knowing the actual value of both keywords and trapdoors, the authors systematically construct a novel secure search protocol. As a result, different data owners use different keys to encrypt their files and keywords. Authenticated data users can issue a query without knowing secret keys of these different data owners. To rank the search results and preserve the privacy of relevance scores between keywords and files, a new additive order and privacy preserving function family (AOPPF) was proposed, which helps the cloud server return the most relevant search results to data users without revealing any sensitive information.

To prevent the attackers from eavesdropping secret keys and pretending to be legal data users submitting searches, a novel dynamic secret key generation protocol and a new data user authentication protocol was proposed. As a result, attackers who steal the secret key and perform illegal searches would be easily detected. Extensive experiments on real-world datasets confirm the efficacy and efficiency of the proposed schemes.

Gurjar et al. [15] proposed a new approach using MIR-tree, it provides privacy of outsourced data,

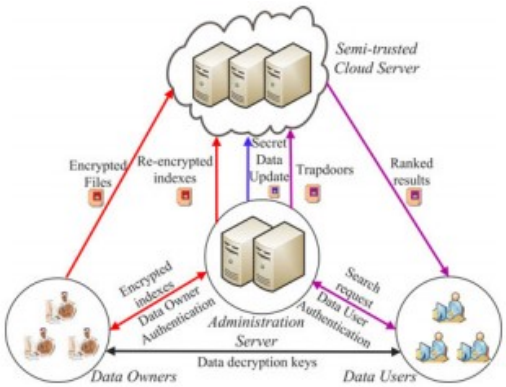

Fig. 3. Architecture of PRMSM

and also provide the authentication of the search results, thus ensures correctness and the completeness of top-*k* results. An authentication set is designed to verify the top-k result and provide efficient searching and ranking techniques to get top-k result. The MIR-tree based index is developed to make a secure and flexible index structure which supports efficient search or traversal with the help of the query trapdoor. In addition, the authors used the ranking function to get top-*k* results and an authentication set is designed to verify the top-k results. The security analysis proved that the scheme is fully able to perform privacy-preserving multi-keyword search along with the verification of the top-k results.

Wan et al. [16] design an efficient, verifiable and privacy-preserving multi-keyword ranked searchable encryption scheme called VPSearch for outsourced cloud data under the partially honest cloud server model. It is realized by integrating an adapted homomorphic MAC technique with a privacy preserving multi-keyword search scheme. The proposed scheme is very efficient as it relies on only one-way function for security. The authors also provide the random challenge technique to verify top-*k* search results for a given query.

With this solution, the client can be sure that the top-*k* results are authentic for probability close to 1. Also, the authors provide detailed analysis on security, privacy, verifiability and efficiency of VPSearch. Specifically, the underlying homomorphic MAC scheme used in VPSearch can be proved to be secure. The authors implement VPSearch using Matlab and evaluate its performance over three UCI data sets. Experiment results on a laptop showed that VPSearch is very efficient on authentication tag generation and keyword search operations.

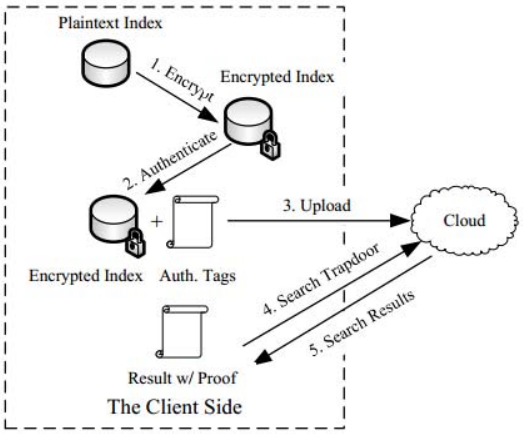

Fig. 4. Overview of VPSearch

As shown in Fig.4. [16] the client first encrypts the plaintext index, then the encrypted index is authenticated with homomorphic MAC technique, which produces authentication tags for the encrypted index. Next, the index and authentication tags are uploaded to the cloud. Then the client can generate a search trapdoor, and uses our homomorphic MAC technique to authenticate the trapdoor. With the authenticated trapdoor, the cloud server can homomorphically execute the search function over the authentication tags to derive the result with a proof, which can certify the search result.

Li et al. [17] focused on enabling efficient and privacy preserving similarity keyword search scheme termed PSS in a multi-cloud scenario. The key contributions of this paper can be summarized as follows: First, the authors formalized the problem of efficient and privacy preserving similarity search in a multi-cloud scenario as shown in Fig 5. [17], and establish corresponding system and threat models. Second, the authors proposed algorithms for keyword-order computation, Chord-ring construction and similarity search processing, over encrypted data stored in multiple cloud servers with high efficiency, and without the privacy leakage of keywords and files. Third, the authors justified the privacy-preserving property of proposed scheme. Also, performance evaluations are done on a real-world dataset, and the results show the efficiency of proposed scheme.

Fu et al. [18], propose two practical and processing schemes to solve the challenging problem: conceptual graph (CG) match in the encrypted form. The authors used conceptual graphs as a knowledge representation to substitute traditional keywords and solve the problem of privacy-preserving smart

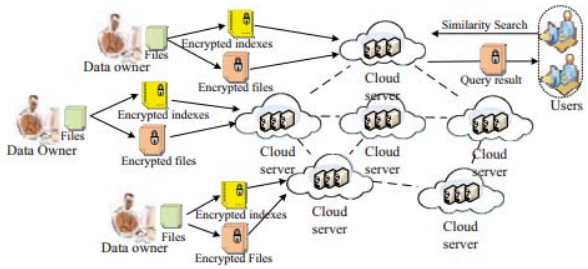

Fig. 5. Architecture of PSS over multiple cloud

semantic search based on conceptual graphs over encrypted outsourced data. In addition, the authors creatively propose a modified linear form of conceptual graphs which makes quantitative calculation on conceptual graphs possible. In a sense, the authors facilitates fuzzy retrieval on conceptual graphs in semantic level. Moreover, CG's scheme present two practical schemes from different aspects to solve the problem of privacy-preserving smart semantic search based on conceptual graphs over encrypted outsourced data. They are both secure and efficient, but have their own focus on different aspects.

Xu et al. [19] introduced a complete framework of privacy preserving ranked fuzzy keyword search over encrypted cloud data. For the efficiency and security consideration, the scheme adopted the dictionary-based fuzzy keyword search and the One-to-many Order-Preserving Mapping scheme to build the inverted index and search. Through performance and security analysis, the scheme showed that the proposed scheme is privacy preserving and efficient.

V. COMPARISON AND ANALYSIS OF PRIVACY-PRESERVING KEYWORD SEARCH SCHEMES

We presented the most recently proposed keyword search schemes for mitigating the privacy issue over encrypted data in cloud computing. Table 1 summarizes the comparative study of each scheme. Apart from the novelty of the proposed search schemes, we evaluate each solution using the following seven main criteria: Search functionality, system model, threat model, entities, privacy, efficiency and verifiability. From our analysis, we observed that evaluation of verifiability was missing except the scheme VPSearch.

Furthermore, in addition to the existing requirements, privacy preserving search schemes should meet the requirements such as dynamic and verifiability.

Table 1. Comparison and Analysis of Existing Schemes

Scheme	Search Functionality	System Model	Threat Model	Entities	Privacy	Verifiability	Efficiency
Jiang et al. [13]	Single Keyword	Single-owner and single user	Honest-but-curious	Data owner, Data user The Cloud server	Data, Index and Trapdoor	No	No
Zhang et al. [14]	Ranked Multi-Keyword	Multi-owner and Multi-user	Curious-but-Honest	Data owners The cloud server, Administration server, Data users	Data and Keyword	No	No
Gurjar et al. [15]	Ranked Multi-Keyword	Single-owner and Single-user	Known Cipher text model and Known Background model	Data owner, Data user, The Cloud server	Data, Index, Keyword and Trapdoor	No	Yes
Wan et al. [16]	Multi-Keyword	Single-owner and Single User	Honest-but-Curious	Data User The Cloud Server	Keyword and Data	Yes	Yes
Li et al. [17]	Similarity-Keyword	Multi-owner and Multi-users	Honest-but-Curious	Data owners, Data Users, The Cloud servers	Keywords and files	No	Yes
Fu et al. [18]	Smart Semantic Search	Single-owner and Single-user	Known Cipher text model and Known Background model	Data Owner, Data user The Cloud Server	Data, Index, Keyword and Trapdoor	No	No
Xu et al. [19]	Ranked Fuzzy keyword search	Single-owner and single -user	Honest-but-curious	Data owners, Data Users The Cloud servers	Data and index	No	Yes

Moreover, we observed that the search functionality of each scheme varied from single keyword, multi-keyword, ranked multi-keyword, similarity search, smart semantic search and ranked fuzzy keyword search. We noted that these schemes greatly reduces the system usability and efficiency, except the scheme ranked fuzzy keyword search. In terms of system model, cloud servers mostly support multiple data owners to share the benefits from cloud computing.

Among these compared schemes only two schemes [17] and [14] used multiple data owner system model. Most importantly, the malicious cloud service provider may delete the unused encrypted files for saving space. As a result the scheme should have the ability to verify the correctness of the search results. From our analysis, only one scheme [16] has the verifiability technique.

VI. Conclusion and Future Work

In this paper, we presented the survey of recently proposed privacy-preserving keyword search schemes specifically to mitigate the privacy issues in cloud data by highlighting the advantages and limitations of the prominent search schemes. In addition, we have provided a comparative study of the currently existing keyword search schemes in detail to further assist readers to understand the specific design goal required for the keyword search techniques especially in terms of privacy.

For future work, the readers can design an efficient privacy-preserving ranked fuzzy keyword search scheme based on the following requirements such as security, privacy, efficiency and verifiability.

REFERENCES

[1] C. Wang, S. S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," IEEE Trans. Comput., vol. 62, no. 2, pp. 362–375, Feb. 2013.

[2] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. IEEE Int. Symp. Security Privacy, Nagoya, Japan, Jan. 2000, pp. 44–55.

[3] E. Goh. (2003). Secure indexes [Online]. Available: http://eprint.iacr.org/

[4] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: Improved definitions and efficient constructions,"
in Proc. 13th ACM Conf. Comput. Commun. Security, Oct. 2006, pp. 79–88.

[5] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in Advances in Cryptology Eurocrypt
2004, Springer, 2004, pp. 506–522.

[6] P. Golle, J. Staddon, and B. Waters, "Secure conjunctive keyword search over encrypted data," in Proc. Appl. Cryptography Netw. Security, Yellow Mountain, China, Jun. 2004, pp. 31–45.

[7] L. Ballard, S. Kamara, and F. Monrose, "Achieving efficient conjunctive keyword searches over encrypted data," in Proc. Inf. Commun. Security, Beijing, China, Dec. 2005, pp. 414–426.

[8] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data," in Proc. IEEE Distrib. Comput. Syst., Genoa, Italy, Jun. 2010, pp. 253–262.

[9] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," in Proc. IEEE INFOCOM, Shanghai, China, Apr. 2011, pp. 829–837.

[10] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 1, pp. 222–233, Jan. 2014.

[11] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li,"Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 11, pp. 3025–3035, Nov. 2014.

[12] Z. Xu, W. Kang, R. Li, K. Yow, and C. Xu, "Efficient multikeyword ranked query on encrypted data in the cloud," in Proc. IEEE 19th Int. Conf. Parallel Distrib. Syst., Singapore, Dec. 2012, pp. 244–251.

[13] X. Jiang, J. Yu, F. Kong, X. Cheng and R. Hao, "A Novel Privacy Preserving Keyword Search Scheme over Encrypted Cloud Data," *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, Krakow, 2015, pp. 836-839.
doi: 10.1109/3PGCIC.2015.48

[14] W. Zhang, Y. Lin, S. Xiao, J. Wu and S. Zhou, "Privacy Preserving Ranked Multi-Keyword Search for Multiple Data Owners in Cloud Computing," in *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1566-1577, May 1 2016. doi: 10.1109/TC.2015.2448099

[15] S. P. S. Gurjar and S. K. Pasupuleti, "A privacy-preserving multi-keyword ranked search scheme over encrypted cloud data using MIR-tree," *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, Pune, 2016, pp. 533-538. doi: 10.1109/CAST.2016.7915026

[16] Z. Wan and R. H. Deng, "VPSearch: Achieving Verifiability for Privacy-Preserving Multi-Keyword Search over Encrypted Cloud Data," in *IEEE Transactions on Dependable and Secure Computing*, vol. PP, no. 99, pp. 1-1. doi: 10.1109/TDSC.2016.2635128

[17] J. Li, M. Wen, C. Gu and H. Li, "PSS: Achieving high-efficiency and privacy-preserving similarity search in multiple clouds," *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, 2016, pp. 1-6. doi: 10.1109/ICC.2016.7511324

[18] Z. Fu, F. Huang, K. Ren, J. Weng and C. Wang, "Privacy-Preserving Smart Semantic Search Based on Conceptual Graphs Over Encrypted Outsourced Data," in *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1874-1884, Aug. 2017.

[19] Q. Xu, H. Shen, Y. Sang and H. Tian, "Privacy-Preserving Ranked Fuzzy Keyword Search over Encrypted Cloud Data," *2013 International Conference on Parallel and Distributed Computing, Applications and Technologies*, Taipei, 2013, pp. 239-245. doi: 10.1109/PDCAT.2013.44