Analysis of One Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data

Zhengjun Cao[1], Chong Mao[1], Lihua Liu[2], Wenping Kong[2], Jinbo Wang[3]
(Corresponding author: Zhengjun Cao)

Department of Mathematics, Shanghai University[1]
No.99, Shangda Road, Shanghai, China
(Email: caozhj@shu.edu.cn)
Department of Mathematics, Shanghai Maritime University[2]
Science and Technology on Communication Security Laboratory, China[3].

(Received Mar. 20, 2017; revised and accepted June 26 & July 2, 2017)

Abstract

In 2016, Xia et al. have proposed a scheme for privacy-preserving multi-keyword ranked search over encrypted cloud data [IEEE TPDS, 2016, 340-352]. In this note we show that Xia et al.'s scheme is flawed because the introduced relevance scores do not indicate the true Euclidean distances between the index vectors and the query vector. The scheme has not developed a proper procedure for distance comparison which should be compatible with the technique of Scalar-Product-Preserving Encryption. In the scheme the returned documents are not indeed related to the queried keywords. We also present an improvement using the technique developed by Wong *et al.*'s work [ACM SIGMOD 2009].

Keywords: Cloud Computing; Multi-Keyword Ranked Search; Privacy-Preserving Search; Scalar-Product-Preserving Encryption

1 Introduction

Cloud computing benefits scientific and engineering applications, such as data mining, computational financing, and many other data-intensive activities by supporting a paradigm shift from local to network-centric computing and network-centric content [23]. It enables customers with limited computational resources to outsource large-scale computational tasks to the cloud.

In 2010, Kamara and Lauter [16] discussed the security problem of cloud storage. In 2013, Liu *et al.* [21] explored the problem of multiowner data sharing for dynamic groups in the cloud. Chen *et al.* [12,29] investigated on achieving secure role-based access control on encrypted data in cloud storage. Nabeel *et al.* [24] designed a scheme with privacy preserving policy based content sharing in public clouds.

In 2014, Chen *et al.* proposed two computation outsourcing schemes for linear equations and for linear programming [9, 10]. But the schemes are insecure because the technique of masking a vector with a diagonal matrix is vulnerable to statistical analysis attacks [6]. The Wang *et al.*'s scheme for outsourcing linear equations is flawed [5], too.

In 2016, Khaleel *et al.* [17, 25] discussed the possibility of using caching search engine for files retrieval system, and using cloud based technique for blog search optimization. Hsien *et al.* [8, 11, 15, 19] have presented some surveys on public auditing for secure data storage in cloud computing.

Searchable encryption [1–3, 7, 13, 18, 26] is a very appreciated tool that allows a user to securely search over encrypted data through keywords and retrieve documents of interest. Lu *et al.* [22] have discussed how existing additive homomorphic encryption can be potentially used for image search, and proposed two confidentiality-preserving image search schemes based on Paillier's encryption.

In the proposed model, a client has many images and wants to store the image data online for convenient data access anywhere anytime. The client has to encrypt each image and its features and upload the encrypted data to a cloud server. In 2016, Liu and Cao [20] pointed out that Lu *et al.*'s schemes did not make use of the additive homomorphic property at all and the additive homomorphic encryption in one scheme was unnecessary and can be replaced by a more efficient symmetric key encryption.

Recently, Xia *et al.* [28] proposed a scheme for privacy-preserving multi-keyword ranked search over encrypted cloud data. In this note we show that in Xia *et al.*'s scheme the cloud server cannot determine which encrypted index vector I_u is more similar to the encrypted query vector TD. Actually, the relevance score $s_u := I_u \cdot TD = D_u \cdot Q$ does not represent the true similarity between the unencrypted index vector D_u and the unencrypted query vector Q.

The remainder of this paper is organized as follows.

Table 1: Scalar-product-preserving encryption

Key	A $(d+1)\times(d+1)$ invertible matrix M.
DataEnc	For a d-dimensional vector p, set $\hat{p} = (p^T, -0.5\|p\|^2)^T$ and encrypt it as $p' = M^T\hat{p}$.
QueryEnc	For a querying vector q, pick a random number $r > 0$, set $\hat{q} = r(q^T, 1)^T$ and encrypt it as $q' = M^{-1}\hat{q}$.
DistanceComp	Let p'_1 and p'_2 be the encrypted p_1 and p_2 respectively. To determine whether p_1 is nearer to a query q than p_2 is, check whether $(p'_1 - p'_2)\cdot q' > 0$.
DataDecry	Given p', compute $p = (I_d, 0)(M^T)^{-1}p'$ where I_d is the $d\times d$ identity matrix.

In Section 2, we describe the technique of scalar-product-preserving encryption (SPPE) and explain in detail that the technique is compatible with the formal routine of distance-comparison. In Section 3, we provide an explicit description of Xia *et al.*'s scheme (see Table 2). We then point out that Xia *et al.*'s scheme is flawed because the introduced variation of SPPE is not compatible with the routine of distance-comparison (Euclidean distance). In Section 5, we present an improvement of Xia *et al.*'s scheme by extending an index vector to a higher dimension one in order to keep the compatibility between SPPE and distance-comparison. At last, we stress that SPPE must be integrated with the common mechanism for distance comparison in order to represent the similarity scores of vectors.

2 Scalar Product Preserving Encryption

Given two n-dimension vectors X_1, X_2 and another n-dimension vector Y, to determine which $X_i, i = 1, 2$, is more similar to Y, it is usual to compute the distances

$$d(X_i, Y) = \|X_i - Y\| = \sqrt{\|X_i\|^2 - 2X_i\cdot Y + \|Y\|^2},$$

where $i = 1, 2$ and $\|X\|$ represents the Euclidean norm of X, and compare the distances. If $d(X_1, Y) < d(X_2, Y)$, then we assert X_1 is more similar to Y.

In 2009, Wong *et al.* [27] introduced the technique of scalar-product-preserving encryption which can be explained as follows (see Table 1).

The encryption is distance-recoverable because

$$\begin{aligned}
& (p'_1 - p'_2)\cdot q' = (p'_1 - p'_2)^T q' \\
= \ & (M^T\hat{p_1} - M^T\hat{p_2})^T M^{-1}\hat{q} \\
= \ & (\hat{p_1} - \hat{p_2})^T\hat{q} = \big((p_1^T, -0.5\|p_1\|^2)^T \\
& - (p_2^T, -0.5\|p_2\|^2)^T\big)^T r(q^T, 1)^T \\
= \ & r(p_1^T - p_2^T, -0.5\|p_1\|^2 + 0.5\|p_2\|^2)(q^T, 1)^T \\
= \ & \frac{1}{2}r(2p_1^Tq - 2p_2^Tq - \|p_1\|^2 + \|p_2\|^2) \\
= \ & \frac{1}{2}r\left((\|p_2\|^2 - 2p_2^Tq + \|q\|^2)\right. \\
& \left. - (\|q\|^2 - 2p_1^Tq + \|p_1\|^2)\right) \\
= \ & \frac{1}{2}r\left(\|p_2 - q\|^2 - \|p_1 - q\|\right)^2 \\
= \ & \frac{1}{2}r\left(\|p_2 - q\| + \|p_1 - q\|\right) \\
& \cdot\left(\|p_2 - q\| - \|p_1 - q\|\right) \\
= \ & \frac{1}{2}r\left(d(p_2, q) + d(p_1, q)\right)\left(d(p_2, q) - d(p_1, q)\right)
\end{aligned}$$

Set the similarity score as $s_i = p'_i\cdot q', i = 1, 2$. Since $r\left(d(p_2, q) + d(p_1, q)\right) > 0$, we have

$$(p'_1 - p'_2)\cdot q' > 0 \Leftrightarrow d(p_2, q) - d(p_1, q) > 0,$$

$$s_1 > s_2 \Leftrightarrow d(p_2, q) > d(p_1, q).$$

Thus, the similarity score can be used to indicate the Euclidean distance between the original vector p and the query vector q.

3 Review of Xia *et al.*'s Scheme

The scheme [28] involves three entities: data owner, data user and cloud server.

Data owner has a collection of documents $\mathcal{F} = \{f_1, f_2, \cdots, f_n\}$ that he wants to outsource to the cloud server in encrypted form while still keeping the capability to search on them for effective utilization. Data users are authorized ones to access the documents of data owner. Cloud server stores the encrypted document collection $\mathcal{C}$ and the encrypted searchable tree index $\mathcal{I}$ for data owner.

Upon receiving the trapdoor TD from the data user, the cloud server executes search over the index tree $\mathcal{I}$, and finally returns the corresponding collection of top-k ranked encrypted documents.

The scheme consists of the following phases (see Table 2). We refer to the original for the full description of the scheme [28].

Table 2: Xia *et al.*'s scheme

Date Owner		Server
Setup. Pick a m-bit S and two		
$m \times m$ invertible matrices M_1, M_2.		
Set (S, M_1, M_2) as the secret key.		
Pick a symmetric key encryption $(\mathcal{E}, \mathcal{D})$.		
GenIndex. For files $\mathcal{F} = \{f_1, f_2, \cdots, f_n\}$		
and keywords $\mathcal{W} = \{w_1, w_2, \cdots, w_m\}$, set		
the index $\mathcal{T}$ for $\mathcal{F}$. For the vector D_u in		
node u, split it into $(D'_u, D^{''}_u)$:		
if $S[j] = 0$, then $D'_u[j] = D^{''}_u[j] = D_u[j]$;		
if $S[j] = 1$, then $D'_u[j] + D^{''}_u[j] = D_u[j]$.		
Set the encrypted index tree as $\mathcal{I}$, where		
the node u stores $I_u = \{M_1^T D'_u, M_2^T D^{''}_u\}$.	$\xrightarrow{\mathcal{I}, c_i=\mathcal{E}(f_i), i=1,\cdots,n}$	Store $\mathcal{I}$ and all c_i.
Date user		Server
Query. Given $\mathcal{W}_q \subset \mathcal{W}$, generate Q		
for $\mathcal{W}_q$ and split it into $Q', Q^{''}$:		
if $S[j] = 1$, then $Q'[j] = Q^{''}[j] = Q[j]$;		
if $S[j] = 0$, then $Q'[j] + Q^{''}[j] = Q[j]$.	$\xrightarrow{TD=\{M_1^{-1}Q', M_2^{-1}Q^{''}\}}$	**Response**. Compute all
		scores $s_u = I_u \cdot TD$, return
Output. Decrypt all files in $\mathcal{C}_{\mathcal{W}_q}$.	$\xleftarrow{\mathcal{C}_{\mathcal{W}_q}}$	the top ranked id list $\mathcal{C}_{\mathcal{W}_q}$.

4 Xia *et al.*'s Scheme is Flawed

In Xia *et al.*'s scheme, the cloud server has to compute the relevance score

$$\begin{aligned} s_u &= I_u \cdot TD \\ &= \{M_1^T D'_u, M_2^T D^{''}_u\} \cdot \{M_1^{-1}Q', M_2^{-1}Q^{''}\} \\ &= D'_u \cdot Q' + D^{''}_u \cdot Q^{''} = D_u \cdot Q \end{aligned}$$

for all nodes. The server then sorts them and returns the top ranked id list $\mathcal{C}_{\mathcal{W}_q}$. We would like to point out that the proposed mechanism fails because the score s_u cannot work well when one considers a true Euclidean distance between the index vector D_u and the query vector Q.

In fact, given two scores $s_i, s_j, i \neq j$, we have

$$s_i - s_j = (D_i - D_j) \cdot Q.$$

If $s_i < s_j$, one cannot determine whether the Euclidean distance $\mathrm{d}(D_i, Q)$ is less than $\mathrm{d}(D_j, Q)$.

Xia *et al.*'s scheme is inspired by Wong *et al.*'s work [27]. The technique of Scalar-Product-Preserving Encryption (SPPE) introduced in [27], *i.e.*, $I_u \cdot TD = D_u \cdot Q$, must be integrated with the routine of Distance-Comparison in order to help the cloud server to sort the final scores according to all $\mathrm{d}(D_u, Q)$. But Xia *et al.* have forgotten to check the compatibility of the variant of SPPE in their scheme with the routine of Distance-Comparison.

5 An Improvement

We now describe an improvement of Xia *et al.*'s scheme by using the technique developed by Wong *et al.* [27]. First, the data owner has to replace S with a $(m+1)$-bit vector rather than the original m-bit vector. Second, the owner sets both M_1, M_2 be of order $m+1$. Third, for the vector D_u in node u, the owner extends it as $\hat{D}_u = (D_u^T, -0.5\|D_u\|^2)^T$. See Table 3 for the details.

The correctness of the improvement is easy to check. In fact, we have

$$\begin{aligned} & s_1 - s_2 = (I_1 - I_2) \cdot TD \\ = \ & \{M_1^T(\hat{D}'_1 - \hat{D}'_2), M_2^T(\hat{D}^{''}_1 - \hat{D}^{''}_2)\} \\ & \cdot\{M_1^{-1}\hat{Q}', M_2^{-1}\hat{Q}^{''}\} \\ = \ & (\hat{D}'_1 - \hat{D}'_2) \cdot \hat{Q}' + (\hat{D}^{''}_1 - \hat{D}^{''}_2) \cdot \hat{Q}^{''} \\ = \ & (\hat{D}_1 - \hat{D}_2) \cdot \hat{Q} \\ = \ & (D_1^T - D_2^T, -0.5\|D_1\|^2 + 0.5\|D_2\|^2)^T \cdot (Q^T, 1)^T \\ = \ & (D_1 - D_2) \cdot Q - 0.5\|D_1\|^2 + 0.5\|D_2\|^2 \\ = \ & 0.5\left(\|D_2\|^2 - 2D_2 \cdot Q + \|Q\|^2\right) \\ & -0.5\left(\|Q\|^2 - 2D_1 \cdot Q + \|D_1\|^2\right) \\ = \ & 0.5\left(\|D_2 - Q\|^2 - \|D_1 - Q\|^2\right) \\ = \ & 0.5\left(\|D_2 - Q\| + \|D_1 - Q\|\right) \\ & \cdot\left(\|D_2 - Q\| - \|D_1 - Q\|\right) \end{aligned}$$

Table 3: An improvement of Xia *et al.*'s scheme

Date owner		Server
Setup. See the original except that		
S is replaced by a $(m+1)$-bit vector,		
and *both M_1, M_2 are of order $m+1$.*		
GenIndex. For the vector D_u in node u,		
extend it as $\hat{D}_u = (D_u^T, -0.5\|D_u\|^2)^T$		
split it into $(\hat{D}'_u, \hat{D}''_u)$:		
if $S[j]=0$, then $\hat{D}'_u[j] = \hat{D}''_u[j] = \hat{D}_u[j]$;		
if $S[j]=1$, then $\hat{D}'_u[j] + \hat{D}''_u[j] = \hat{D}_u[j]$.		
Set the tree as $\mathcal{I}$, where		
the node u stores $I_u = \{M_1^T\hat{D}'_u, M_2^T\hat{D}''_u\}$.	$\xrightarrow{\mathcal{I}, c_i=\mathcal{E}(f_i), i=1,\cdots,n}$	Store $\mathcal{I}$ and all c_i.
Date user		Server
Query. Given Q, extend it as $\hat{Q} = (Q^T, 1)^T$		
and split it into into $\hat{Q}', \hat{Q}''$:		
if $S[j]=1$, then $\hat{Q}'[j] = \hat{Q}''[j] = \hat{Q}[j]$;		
if $S[j]=0$, then $\hat{Q}'[j] + \hat{Q}''[j] = \hat{Q}[j]$.	$\xrightarrow{TD=\{M_1^{-1}\hat{Q}', M_2^{-1}\hat{Q}''\}}$	**Response**. Compute all
		scores $s_u = I_u \cdot TD$, return
Output. Decrypt all files in $\mathcal{C}_{\mathcal{W}_q}$.	$\xleftarrow{\mathcal{C}_{\mathcal{W}_q}}$	the top ranked id list $\mathcal{C}_{\mathcal{W}_q}$.

Thus,

$$s_1 > s_2 \Leftrightarrow \|D_2 - Q\| > \|D_1 - Q\|.$$

In such case the server can determine that D_1 is nearer to Q than D_2, although D_1, D_2, Q are still unknown to the server.

Xia *et al.*'s scheme [28] is similar to Cao *et al.*'s scheme [4]. Both two schemes are the variations of Wong *et al.*'s scheme [27] except the method to build the unencrypted index vector for each file. But the two schemes failed to develop the technique to integrate the scalar-product-preserving encryption with the routine of distance-comparison (Euclidean distance). The improvement adopts the method developed in [27] and split a vector into two parts. It then encrypts these two parts using two invertible matrixes. The mechanism is useful to resist statistical attacks [14]. This strengthens the security at the expense of a little computational cost.

6 Conclusion

We show that Xia *et al.*'s scheme is flawed and present a possible improvement. We also point out that it is conventional to compare the Euclidean distances between a set of encrypted vectors and a given encrypted vector so as to determine their similarities. We would like to stress that the technique of Scalar-Product-Preserving Encryption must be integrated with the common mechanism for distance comparison in order to represent the similarity scores of these vectors.

Acknowledgements

We thank the National Natural Science Foundation of China (Project 61303200, 61411146001). The authors gratefully acknowledge the reviewers for their valuable suggestions.

References

[1] M. Abdalla *et al.*, "Searchable encryption revisited: Consistency properties, relation to anonymous ibe, and extensions," *Journal of Cryptology*, vol. 21, no. 3, pp. 350–391, 2008.

[2] D. Boneh *et al.*, "Public key encryption with keyword search," in *Proceedings of International Conference on Theory and Applications of Cryptographic Techniques (EUROCRYPT'04)*, pp. 506–522, May 2004.

[3] D. Boneh *et al.*, "Public key encryption that allows pir queries," in *Advances in Cryptology (CRYPTO'07)*, pp. 50–67, 2007.

[4] N. Cao *et al.*, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 222–233, 2014.

[5] Z. J. Cao and L. H. Liu, "Comment on 'harnessing the cloud for securely outsourcing large-scale systems of linear equations'," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 5, pp. 1551–1552, 2016.

[6] Z. J. Cao, L. H. Liu, and O. Markowitch, "Comment on 'highly efficient linear regression outsourcing to a cloud'," *IEEE Transactions on Cloud Computing*, 10.1109/TCC.2017.2709299, 2017.

[7] Y. C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in *Proceedings of Third International Conference on Applied Cryptography and Network Security (ACNS'05)*, pp. 442–455, June 2005.

[8] W. Y. Chao, C. Y. Tsai, and M. S. Hwang, "An improved key-management scheme for hierarchical access control," *International Journal of Network Security*, vol. 19, no. 4, pp. 639–643, 2017.

[9] F. Chen, T. Xiang, X. Lei, and J. Chen, "Highly efficient linear regression outsourcing to a cloud," *IEEE Transactions on Cloud Computing*, vol. 2, no. 4, pp. 499–508, 2014.

[10] F. Chen, T. Xiang, and Y. Y. Yang, "Privacy-preserving and verifiable protocols for scientific computation outsourcing to the cloud," *Journal of Parallel and Distributed Computing*, vol. 74, pp. 2141–2151, 2014.

[11] J. S. Chen, C. Y. Yang, and M. S. Hwang, "The capacity analysis in the secure cooperative communication system," *International Journal of Network Security*, vol. 19, no. 6, pp. 863–869, 2017.

[12] T. Y. Chen, C. C. Lee, M. S. Hwang, and J. K. Jan, "Towards secure and efficient user authentication scheme using smart card for multi-server environments," *Journal of Supercomputing*, vol. 66, no. 2, pp. 1008–1032, 2013.

[13] R. Curtmola *et al.*, "Searchable symmetric encryption: Improved definitions and efficient constructions," in *Proceedings of 13th ACM Conf. Computer and Communication Security (CCS'06)*, pp. 79–88, Nov. 2006.

[14] D. Hankerson., A. Menezes, and S. Vanstone, *Guide to Elliptic Curve Cryptography*, Springer, 2004.

[15] W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of public auditing for secure data storage in cloud computing," *International Journal of Network Security*, vol. 18, no. 1, pp. 133–142, 2016.

[16] S. Kamara and K. Lauter, "Cryptographic cloud storage," in *Proceedings of 14th International Conference on Financial Cryptography Data Security (FC'10)*, pp. 136–149, Jan. 2010.

[17] M. Khaleel, H. El-Bakry, and A. Saleh, "A new efficient files retrieval system using caching search engine," *International Journal of Electronics and Information Engineering*, vol. 4, no. 1, pp. 22–31, 2016.

[18] J. Li *et al.*, "Fuzzy keyword search over encrypted data in cloud computing," in *Proceedings of 29th IEEE International Conference on Computer Communications (INFOCOM'10)*, pp. 441–445, Mar. 2010.

[19] C. W. Liu, W. F. Hsien, C. C. Yang, and M. S. Hwang, "A survey of attribute-based access control with user revocation in cloud data storage," *International Journal of Network Security*, vol. 18, no. 5, pp. 900–916, 2016.

[20] L. H. Liu and Z. J. Cao, "Analysis of two confidentiality-preserving image search schemes based on additive homomorphic encryption," *International Journal of Electronics and Information Engineering*, vol. 5, no. 1, pp. 1–5, 2016.

[21] X. Liu, Y. Zhang, B. Wang, and J. Yang, "Mona: Secure multiowner data sharing for dynamic groups in the cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 6, pp. 1182–1191, 2013.

[22] W. J. Lu, A. L. Varna, and M. Wu, "Confidentiality-preserving image search: A comparative study between homomorphic encryption and distance-preserving randomization," *IEEE Access*, no. 2, pp. 125–141, 2014.

[23] D. Marinescu, *Cloud Computing Theory and Practice*, Elsevier, 2013.

[24] M. Nabeel, N. Shang, and E. Bertino, "Privacy preserving policy based content sharing in public clouds," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2602–2614, 2013.

[25] J. Singh, "Cloud based technique for blog search optimization," *International Journal of Electronics and Information Engineering*, vol. 4, no. 1, pp. 32–39, 2016.

[26] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *Proceedings of IEEE Symp. Security and Privacy (IEEE S&P'00)*, pp. 44–55, May 2000.

[27] W.K. Wong, D.W. Cheung, B. Kao, and N. Mamoulis, "Secure knn computation on encrypted databases," in *Proceedings of 35th ACM SIGMOD Int'l Conference on Management of Data*, pp. 139–152, June 2009.

[28] Z. H. Xia, X. H. Wang, X. M. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 2, pp. 340–352, 2016.

[29] L. Zhou, V. Varadharajan, and M. Hitchens, "Achieving secure role-based access control on encrypted data in cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 1947–1960, 2013.

Biography

Zhengjun Cao is an associate professor with the Department of Mathematics at Shanghai University. He received his Ph.D. degree in applied mathematics from

Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He served as a post-doctor in Computer Sciences Department, Université Libre de Bruxelles, from 2008 to 2010. His research interests include cryptography, discrete logarithms and quantum computation.

Chong Mao is currently pursuing his M.S. degree from Department of Mathematics, Shanghai university. His research interests include information security and cryptography.

Lihua Liu is an associate professor with the Department of Mathematics at Shanghai Maritime University. She received her Ph.D. degree in applied mathematics from Shanghai Jiao Tong University. Her research interests include combinatorics, cryptography and information security.

Wenping Kong is currently pursuing her M.S. degree from Department of Mathematics, Shanghai Maritime university. Her research interests include combinatorics and cryptography.

Jinbo Wang received his Ph.D. degree in applied mathematics from Shanghai University. His research interests include applied cryptography and network security.