Commun.Fac.Sci.Univ.Ank.Series A2-A3
Volume 61, Number 1, Pages 1-11 (2019)
DOI: 10.33769/aupse.533815
ISSN 1303-6009 E-ISSN 2618-6462
http://communications.science.ankara.edu.tr/index.php?series=A2A3

COMMUNICATIONS
SERIES A2-A3

PRIVACY-PRESERVING K-NEAREST NEIGHBOUR INTERPOLATION METHOD IN AN OUTSOURCED ENVIRONMENT

Muhammad Rifthy KALIDEEN, Murat OSMANOGLU and Bulent TUGRUL

Abstract. One of the most emerging computer technologies of this decade is cloud computing that allows data owners to outsource their storage and computing requirements. It enables data owners to avoid the costs of building and maintaining a private storage infrastructure. While outsourcing data to cloud promises significant benefits, it possesses substantial security and privacy concerns, especially when data stored in the cloud is sensitive and confidential, like a business plan. Encrypting the data before outsourcing can ensure privacy. However, it will be very difficult to process the cipher text created by the traditional encryption method. Considering this fact, we propose an efficient protocol that allows a query owner to retrieve the interpolation of the top k records from two different databases that are closest to a query point. Note that the databases are stored in two different cloud service providers in encrypted form. We also show that the proposed protocol ensures the privacy and the security of the data and the query point.

1. Introduction

Recent developments in networking technology and the increasing needs for computing resources have motivated many entities to outsource their storage requirements and computing needs. This new computing model is recognized as cloud computing that enables users to outsource their database and the processing functionalities to a cloud [1]. Moreover, they provide access mechanisms for the database, and allow users to make queries on the stored data. Outsourcing a database possesses substantial advantages such as avoidance of the costs of building and maintaining a private storage, and universal access to data without location dependency. On the other hand, clouds are untrustworthy in the context of preserving confidentiality and privacy of the data. Because, it is challenging for clouds to ensure the confidentiality of the stored data, i.e. a break in the cloud where

Received by the editors: November 21, 2018; Accepted: December 13, 2018.
Key word and phrases: Spatial interpolation, k-NN, privacy – preserving, homomorphic encryption.

the data is stored may cause all the data to be visible to the attackers [2-5]. Besides, an inside attacker from the cloud can reveal the data and some other related information such as the queries made for the data to a third party [6, 7]. Thus, public cloud can be considered as semi-honest (i.e. honest-but-curious) because of the presence of such attacker.

To overcome the privacy and security issues discussed above, the client should encrypt the data before outsourcing it to the cloud. The traditional encryption systems such as Advanced Encryption Standard (AES) [8], can be deployed within this aim. However, they do not allow processing the data directly. Particularly, if the client wants to do any operation on the data, he needs to download all the data from the cloud first, and to decrypt the data with its private key to be able to make the operations. Such limitation turns the cloud into a simple data storage service. On the other hand, a homomorphic encryption scheme that allows the data to be processed in the encrypted form is well suited for this type of scenarios. Within this scope, a large number of secure schemes [9-12] have been proposed that employ an homomorphic encryption to support processing on the encrypted data.

In this paper, we study the problem of k-Nearest Neighbour (k-NN) query processing over encrypted spatial data. We consider the following scenario: two data owners DO_1 and DO_2 holding two different spatial data sets D_1 and D_2 outsource their data to two different cloud service providers CSP_1 and CSP_2, respectively. Note that each data set consists of two-dimensional location vectors, and each data owner encrypts his data before outsourcing it. Assume there is a query owner that wants to extract the interpolation of the top k vectors of both data sets that are closest to a query point he provides. In the scenario, the query owner does not want to reveal the query point to the service providers, and the data owners do not want to reveal their data to the query owner or the service providers. Within this direction, we propose an efficient protocol that allows the query owner to retrieve the interpolation of the closest k records from both databases for his query point. The proposed protocol protects the confidentiality of data and the privacy of the query.

Paper overview. We discuss the related works at Section 2. In Section 3, we give the basic definitions. We propose the protocol in Section 4. In addition, we analyze the security and the complexity of our protocol in Section 5.

2. Related Work

k-NN problems have gained a lot of attention due to the wide usage of social media in recent years, and many solutions have been proposed within this direction [13]. These solutions are being used in most of the applications such as similarity check, data mining, and pattern recognition. The studies on k-NN problems can be divided into two categories: according to whether the data is encrypted or not; centralized and distributed k-NN problems.

In the former one, the database and its functionalities were outsourced to an untrusted cloud service provider (CSP) so that the cloud service provider manages the data on behalf of the data owner, and allows an authorized user to access the data. However, this may cause some security problems associated with the privacy and the confidentiality of the data. To overcome this obstacle, the data owner can use some cryptographic techniques to protect the data before outsourcing it to the cloud.

One of the techniques used to protect the confidentiality of the data is to employ a traditional encryption scheme such as AES. However, such encryption scheme does not allow the encrypted data to be processed. On the other hand, some researchers proposed several techniques to process range queries [14, 15] and aggregation queries [16, 17] on the encrypted data. However, the proposed techniques possess some limitations on the security for the k-NN queries.

In addition to these studies, different methods were proposed to solve k-NN problems for spatial data in recent years [18, 19]. Wong et al. [10] presented an encryption technique, named as Asymmetric Scalar – Product – Preserving Encryption (ASPE), in which the data and the queries are encrypted before the outsourcing. However, the encryption algorithms used to mask the data and the queries are slightly different, and the query owner knows the decryption key. On the other hand, Zhu et al. [20] proposed a better solution, named as novel SkNN, where the decryption key is not revealed to the query owner. Nevertheless, the data owner and the query owner jointly create the encryption of the queries. Other than this methods, Hu et al. [18] presented a solution for k-NN problem that utilizes a homomorphic encryption scheme. This scheme supports addition, subtraction and multiplication over encrypted data [21]. However, this solution is vulnerable to the chosen-plaintext attacks. Moreover, all the above techniques leak the access pattern of the data to the cloud servers.

In the distributed k-NN problems, data that is partitioned either vertically or horizontally is stored among the independent parties (servers). Most of the methods in this category rely on Secure Multi-Party Computation (SMC) that allows different parties to jointly compute a function over their inputs without revealing their inputs to the parties. Similar to the centralized k-NN, a number of different solutions have been proposed within this direction. Shaneck et al. [22] suggest a protocol for the horizontally distributed data, and Qi et al. [11] proposed a solution for the linear computation and communication. Besides, Vaidya et al. [23] presented an efficient method for the vertically partitioned data by. Also, a technique, called Private Information Retrieval (PIR), was proposed by Ghinita et al. [24] for the location-based services, that only protects the query privacy, but not the data confidentiality.

Note that no methods discussed above are suitable for our purpose since data are partitioned either vertically or horizontally, but stored as plaintext in the server. However, in our proposed solution, the data and the queries are in the encrypted format, and the data are stored in the server as cipher-text.

3. Background

In this section, we give some definitions that will be used as building blocks in our proposed solution.

3.1. k –NN Algorithm

k-Nearest Neighbor algorithm is one of the most common algorithms used in the similarity matching due to its easiness, efficiency, and applicability to the big amount of data. The algorithms can be implemented for the data in which the distance between the points and the query point can be calculated using distance metrics such as Euclidean distance, Manhattan distance. Briefly, the algorithm works as follows: (i) it first calculates the distance values between the data points and the query point using the corresponding distance metric, (ii) it then finds k indexes of the k nearest distances among the distance values, and (iii) it finally outputs the k points that correspond to the k indexes found in the previous step.

3.2. Homomorphic Encryption (HE)

Homomorphic encryption is an encryption system which allows performing mathematical operations on the cipher texts. This homomorphic feature makes the scheme a functional and strong tool for cloud computing. There are several homomorphic encryption schemes presented in the literature such as Goldwasser-Micali encryption scheme, Paillier system, ElGamal cryptosystem, Boneh-Goh-Nissim crypto scheme and so on [25].

In this paper, we utilize the Paillier encryption scheme as the homomorphic encryption scheme. Briefly, Pailler encryption is an additive homomorphic encryption scheme that allows processing addition and multiplication on the ciphertext as follows:

Addition function - $\mathrm{E_{pk}(a+b) = E_{pk}(a) * E_{pk}(b) \bmod N^2}$ (1)

Multiplication function - $\mathrm{E_{pk}(a*b) = E_{pk}(a)}^{b} \mathrm{\bmod N^2}$ (2)

where $\mathrm{a, b} \in Z_N$, pk is the public key, and N is the product of two large primes.

3.3. Basic Security Primitives

We here also present some basic security protocols that will be used in constructing our main protocol.

3.3.1 Secure Multiplication (SM) Protocol

In the protocol, a data owner (DO) with the secret key and a cloud service provider (CSP) with the inputs $\mathrm{E_{pk}(a)}$ and $\mathrm{E_{pk}(b)}$ securely compute the multiplication $\mathrm{E_{pk}(a*b)}$. During the computation, the plaintexts a and b are not revealed to any party, and the output is only revealed to CSP. Briefly, CSP first computes $\mathrm{E_{pk}}(a+r_1) = \mathrm{E_{pk}}(a) * \mathrm{E_{pk}}(r_1)$ and $\mathrm{E_{pk}}(b+r_2) = \mathrm{E_{pk}}(b) * \mathrm{E_{pk}}(r_2)$ for randomly chosen $r_1, r_2 \in Z_N$, then sends them to DO. Upon receiving, DO decrypts them as $a+r_1$ and $b+r_2$ using the corresponding secret key. DO then multiplies the decryptions as $h = (a+r_1)*(b+r_2) \; mod \; N$, encrpts h and sends it to CSP. After getting the encryption $\mathrm{E_{pk}}(h)$, CSP removes the randomness using the equation

$$a*b = (a+r_1)*(b+r_2) - a*r_1 - b*r_2 - r_1*r_2\,, \tag{3}$$

and gets $\mathrm{E_{pk}}(a*b)$.

3.3.2 Secure Squared Euclidean Distance (SSED) Protocol [26]

In this protocol, a data owner (DO) with the secret key and a cloud service provider (CSP) with the encrypted vectors $\mathrm{E_{pk}(a)}$ and $\mathrm{E_{pk}(b)}$ securely compute $\mathrm{E_{pk}}(|a-b|^2)$, the encryption of squared Euclidean distance between m dimensional vectors a and b. During the computation, the vectors a and b are not revealed to any party, and the output is only revealed to CSP. Briefly, CSP first computes $\mathrm{E_{pk}}(a_i-b_i)$ using the homomorphic properties of the underlying encryption scheme where $1 \leq i \leq m$. Then CSP securely computes $\mathrm{E_{pk}}((a_i-b_i)^2)$ with DO using the SM protocol for each i. Finally, CSP locally adds all $\mathrm{E_{pk}}((a_i-b_i)^2)$ using the homomorphic property in order to get the encryption of squared Euclidean distance $\mathrm{E_{pk}}(|a-b|^2)$.

4. Proposed Solution

In this section, we will give an efficient k-NN protocol that allows a query owner (QO) to extract the interpolation of the top k records, which are closest to a query point provided by QO. These k records are selected from two spatial databases stored in two different Cloud Service Providers. There are five parties involved in our proposed scheme:

- Data Owner #1 ($\mathrm{DO_1}$)
- Data Owner #2 ($\mathrm{DO_2}$)
- Cloud Service Provider #1 ($\mathrm{CSP_1}$)
- Cloud Service Provider #2 ($\mathrm{CSP_2}$)
- Query Owner (QO)

We assume that each data owner $\mathrm{DO_u}$ holds a spatial database $\mathrm{D_u}$ that consists of n records $t_1^{(u)}, \ldots, t_n^{(u)}$ such that each record is a two-dimensional vector as $t_i^{(u)} = (t_{i,1}^{(u)}, t_{i,2}^{(u)})$ obtained from a specific region. Our protocol is formalized as follows:

I. Each $\mathrm{DO_u}$ first encrypts its database $\mathrm{D_u}$ using his public key $\mathrm{pk_u}$, as $\{E_{pk_u}(t_1^{(u)}), \ldots, E_{pk_u}(t_n^{(u)})\}$ and sends the encryption of the database to the corresponding server $\mathrm{CSP_u}$.

II. When QO wants to extract the interpolation of the top k records, closest to a query point $Q = (q_1, q_2)$, she first creates two encryptions of her query point $E_{pk_1}(Q)$ and $E_{pk_2}(Q)$ under the public keys pk_1 and pk_2 of

DO$_1$ and DO$_2$, respectively. QO then sends each encryption $E_{pk_u}(Q)$ to the corresponding server CSP$_u$.

III. Upon receiving $E_{pk_u}(Q)$, each CSP$_u$ executes the SSED protocol with the corresponding data owner D$_u$ on the inputs $\left(E_{pk_u}(Q), E_{pk_u}\left(t_i^{(u)}\right)\right)$ where $1 \leq i \leq n$. As the output of the protocol, each CSP$_u$ receives the encryption of the squared Euclidean distance between the query point Q and each record $t_i^{(u)}$ as $E_{pk_u}\left(d_i^{(u)}\right)$, and sends these encryptions to the corresponding DO$_u$.

IV. Upon getting the encryptions, each DO$_u$ first decrypts them as $d_i^{(u)}$, then sorts these decryptions. At this moment, each DO$_u$ have n sorted distance values. DO$_2$ sends the first k values $\left\{d_1'^{(2)}, \ldots, d_k'^{(2)}\right\}$ of his sorted sequence to DO$_1$.

V. After that, DO$_1$ adds the first k values $\left\{d_1'^{(1)}, \ldots, d_k'^{(1)}\right\}$ of his sorted sequence to the k values he received from DO$_2$. DO$_1$ then sorts all 2k values. Based on the first k values in this sorting, DO$_1$ creates the index sets S$_1$ and S$_2$ such that S$_1$ includes the indexes of p values from the first k that are chosen among $\left\{d_1'^{(1)}, \ldots, d_k'^{(1)}\right\}$ and S$_2$ includes the indexes of q values from the first k that are chosen among $\left\{d_1'^{(2)}, \ldots, d_k'^{(2)}\right\}$ where $p + q = k$. DO$_1$ sends each S$_u$ to the corresponding server CSP$_u$.

VI. Upon receiving the index set $S_1 = \{i_1, \ldots, i_p\}$, CSP$_1$ computes the sum of the n records of the corresponding indexes as $E_{pk_1}\left(t_{i_1} + \cdots + t_{i_p}\right)$ using the homomorphic properties of the encryption scheme. In a similar way, CSP$_2$ computes the sum of the m records of the corresponding indexes in $S_2 = \{j_1, \ldots, j_q\}$ as $E_{pk_2}\left(t_{j_1} + \cdots + t_q\right)$. Then, each CSP$_u$ masks the sum with a random vector $r_u \in Z_N \times Z_N$ as $E_{pk_u}(\gamma_u) = E_{pk_u}(T_u + r_u)$ using the homomorphic property of the encryption scheme, and sends the result to the corresponding DO$_u$ and the random vector r_u to QO.

VII. After getting $E_{pk_u}(\gamma_u)$, each DO$_u$ decrypts it, and sends the decryption γ_u to QO.

VIII. Upon receiving γ_1 and γ_2, QO removes the randomness r_1 and r_2, and gets the sum of the closest records as $T_1 = \gamma_1 - r_1$ and $T_2 = \gamma_2 - r_2$. Finally, QO calculates the interpolation of the closest k records of both databases to his query point Q as $(T_1 + T_1)/k$.

5. Analysis of the Proposed Solution

In this section, we analyse the computational complexity of the proposed protocol, and discuss security issues.

5.1 Complexity Analysis

At the beginning of the protocol, each data owner creates the encryption of his data that costs only O(n) encryptions. At the second step, the query owner only makes O(1) encryptions to create the encryptions of his query point. Then, each CSP_u executes O(n) instantiations of SSED protocols with the corresponding DO_u. If we implement the SSED protocol proposed by [25] which involves O(m) encryptions and O(m) exponentiations for the m-dimensional records. Since we consider the data that contains two-dimensional vectors, O(n) instantiations of SSED cost only O(n) operations for each distance set. At the fourth step, the data owners sort the n distance vectors, that costs O(nlogn) operations with an efficient sorting algorithm. Later, DO_1 collects the first k values from each distance set, and sorts them. This part costs O(klogk) operations for each. At the sixth step, CSP_1 makes O(p) multiplications, and CSP_2 makes O(q) multiplications. Since $p + q = k$, O(p) and O(q) are bounded by O(k). At the final step, QO makes just a constant number of operations to calculate the interpolation of the top k records that are closest to the query point. Therefore, the computational complexity of our protocol is bounded by O(nlogn) operations.

5.2 Security Analysis

First of all, the query owner encrypts the query point using Paillier encryption scheme before giving it to the cloud service providers. Since the Paillier encryption scheme is semantically secure, the query point is not revealed to any DO or CSP. Thus, our protocol protects the privacy of the query point. Similarly, each data owner encrypts his database using Paillier cryptosystem before outsourcing it to the cloud. Since the Paillier encryption scheme is semantically secure, no data record is revealed to any CSP and QO. Thus, our protocol preserves the confidentiality of the data.

Besides, all communications between the parties are made in the encrypted form except the one in which the DO_2 sends k distance vectors to DO_1 and the other in which the data owners DO_1 and DO_2 send the decryptions γ_1 and γ_2 to QO, respectively. In the former one, DO_2 does not send the records directly. Instead, he sends the distance vectors that can be considered as the randomization of the records. In the latter one, each DO_u sends the decryption γ_u which is $\gamma_u = T_u + r_u$ for the random r_u. Thus, no information about the data are leaked to a third party during the protocol.

6. Conclusions

Interpolation plays a critical role in engineering problems. Obtaining measurement values across an entire region is both difficult and costly. When a measurement value is required for an unmeasured location, interpolation methods are employed to produce a prediction value. There is a variety of spatial interpolation methods. Due to its simplicity in geology and hydrology, k-NN is highly preferred as a spatial interpolation method. The accuracy and robustness of prediction models depend on the size of the data. Companies may collect data from the same region. However, no one may want to publicize its data. Therefore, they may prefer to use the methods that protects all parties' private data from each other. In this study, we proposed to execute the k-NN spatial interpolation method on outsourced data of two party without jeopardizing their private data. Our solution also protects the prediction coordinate from the data owners and cloud servers, which can be very important from the query owner's point of view. We also analysed our solution in terms of complexity and security.

References

[1] M. Armbrust, et al., A view of cloud computing, *Communications of the ACM*, 53/4 (2010) 50-58.

[2] T. Ermakova, B. Fabian, and R. Zarnekow, Security and privacy system requirements for adopting cloud computing in healthcare data sharing scenarios, *Proceedings of the Nineteenth Americas Conference on Information Systems* (2013).

[3] K. Hashizume, et al., An analysis of security issues for cloud computing, *Journal of Internet Services and Applications*, 4/1 (2013).

[4] J.J. Rodrigues, et al., Analysis of the security and privacy requirements of cloud-based electronic health records systems, *Journal of Medical Internet Research*, 15/8 (2013).

[5] M.D. Ryan, Cloud computing security: *The scientific challenge, and a survey of solutions, Journal of Systems and Software*, 86/9 (2013) 2263-2268.

[6] K. Ren, C. Wang, and Q. Wang, Security challenges for the public cloud, *IEEE Internet Computing,* 16/1 (2012) 69-73.

[7] C. Shahabi, et al., Privacy-preserving inference of social relationships from location data: a vision paper, *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 9 (2015) 1-4.

[8] J. Daemen, and V. Rijmen, The Design of Rijndael: AES - The Advanced Encryption Standard, Springer Berlin Heidelberg, (2013).

[9] D.X. Song, D. Wagner, and A. Perrig, Practical techniques for searches on encrypted data, *Proceeding 2000 IEEE Symposium on Security and Privacy,* (2000) 44-55.

[10] W.K. Wong, et al., Secure kNN computation on encrypted databases, *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data,* (2009) 139-152.

[11] Y. Qi and M.J. Atallah, Efficient privacy-preserving k-nearest neighbor search, *The 28th International Conference on Distributed Computing Systems.* (2008) 311-319.

[12] M.R. Kalideen, and B. Tugrul, Outsourcing of Secure k-Nearest Neighbours Interpolation Method, *International Journal of Advanced Computer Science and Applications,* 9/4 (2018) 319-323.

[13] J. Peng, K. K. R. Choo, and H. Ashman, Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles, *Journal of Network and Computer Applications,* 70 (2016) 171-182.

[14] B. Hore, et al., Secure multidimensional range queries over outsourced data, *The VLDB Journal,* 21/3 (2012) 333-358.

[15] B. Hore, S. Mehrotra, and G. Tsudik, A privacy-preserving index for range queries, *Proceedings of the Thirtieth International Conference on Very Large Data Bases* - Volume 30, (2004) 720-731.

[16] H. Hacıgümüş, B. Iyer, and S. Mehrotra, Efficient execution of aggregation queries over encrypted relational databases, *International Conference on Database Systems for Advanced Applications,* (2004) 125-136.

[17] E. Mykletun and G. Tsudik, Aggregation queries in the database-as-a-service model, *IFIP Annual Conference on Data and Applications Security and Privacy*, (2006) 89-103.

[18] H. Hu, et al., Processing private queries over untrusted data cloud through privacy homomorphism, *IEEE 27th International Conference on Data Engineering (ICDE)*, (2011) 601-612.

[19] B. Yao, F. Li, and X. Xiao, Secure nearest neighbor revisited, *IEEE 29th International Conference on Data Engineering (ICDE).* (2013) 733-744.

[20] Y. Zhu, R. Xu, and T. Takagi, Secure k-NN computation on encrypted cloud data without sharing key with query users, *Proceedings of the 2013 International Workshop on Security in Cloud Computing.* (2013) 55-60.

[21] J. Domingo-Ferrer, A provably secure additive and multiplicative privacy homomorphism, *International Conference on Information Security.* (2002) 471-483.

[22] M. Shaneck, Y. Kim, and V. Kumar, *Privacy preserving nearest neighbor search, Machine Learning in Cyber Trust.* (2009) 247-276.

[23] J. Vaidya and C. Clifton, Privacy-preserving top-k queries, *21st International Conference on Data Engineering (ICDE'05),* (2005) 545-546.

[24] G. Ghinita, et al., Private Queries in Location Based Services: Anonymizers Are Not Necessary, *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data,* (2008) 121-132.

[25] X. Yi, R. Paulet, and E. Bertino, Homomorphic Encryption and Applications, (2014) Springer.

[26] Y. Elmehdwi, B.K. Samanthula, and W. Jiang, Secure k-nearest neighbor query over encrypted data in outsourced environments, *2014 IEEE 30th International Conference on Data Engineering (ICDE)*, (2014) 664-675.

Current Address: Muhammad Rifthy KALIDEEN, Department of Islamic Studies, South Eastern University of Sri Lanka, Oluvil Park, Oluvil, Sri Lanka, 32360
E-mail: kmr@seu.ac.lk
ORCID: https://orcid.org/0000-0001-5790-1166

Current Address: Murat OSMANOGLU, Department of Computer Engineering, Ankara University, Turkey, 06100
E-mail: mosmanoglu@ankara.edu.tr
ORCID: https://orcid.org/0000-0001-8693-141X

Current Address: Bulent TUGRUL, Department of Computer Engineering, Ankara University, Turkey, 06100
E-mail: btugrul@eng.ankara.edu.tr
ORCID: https://orcid.org/0000-0003-4719-4298