Adaptive Data Protection methodology Using K-Nearest Neighbour Query Services over the Cloud Environment

#1 P.Suresh, #2 M.Sireesha #3 Sk. Abdul Rasheed

[1]*M.tech student.Computer ScienceEngineering,Narasaraopeta Engineering College,Narasaraopeta,Guntur Dist,Andrapradesh,INDIA*

[2,3]*Assistant Professor,Computer Science&Engineering,Narasaraopeta Engineering College,Narasaraopeta,Guntur Dist,Andrapradesh,INDIA*

[1] spathuri24@gmail.com
[2] moturisireesha@gmail.com
[2] rasheed4321@gmail.com

Abstract**—**For the past decade, query processing on relational data has been studied extensively, and many theoretical and practical solutions to query processing have been proposed undervarious scenarios. With the recent popularity of cloud computing, users now have the opportunity to outsource their data as well as the data management tasks to the cloud. However, due to the rise of various privacy issues, sensitive data (e.g., medical records) need to be encrypted before outsourcing to the cloud. In addition, query processing tasks should be handled by thecloud; otherwise, there would be no point to outsource the dataat the first place. To process queries over encrypted data without the cloud ever decrypting the data is a very challenging task. In this paper, we focus on solving the k-nearest neighbour (kNN) query problem over encrypted database outsourced to a cloud:a user issues an encrypted query record to the cloud, and the cloud returns the k closest records to the user. We first presenta basic scheme and demonstrate that such a naive solution is not secure. To provide better security, we propose a secure Knn protocol that protects the confidentiality of the data, user's input query, and data access patterns. Also, we empirically analyse the efficiency of our protocols through various experiments. The seresults indicate that our secure protocol is very efficient on the user end, and this lightweight scheme allows a user to use any mobile device to perform the kNN query.

I. INTRODUCTION

As an emerging computing paradigm, cloud computing attracts many organisations to consider utilising the benefits of a cloud in terms of cost-efficiency, flexibility, and off load of administrative overhead. In cloud computing model [1], [2], a data owner outsources his/her database T and the DBMS functionalities to the cloud that has the infrastructure to host out sourced databases and provides access mechanisms for querying and managing the hosted database. On one hand, by out sourcing, the data owner gets the benefit of reducing thedata management costs and improves the quality of service.On the other hand, hosting and query processing of data out of the data owner control raises security challenges such as preserving data confidentiality and query privacy.One straightforward way to protect the confidentiality of the outsourced data from the cloud as well as from the unauthorized users is to encrypt data by the data owner before outsourcing [3]. By this way, the data owner can protect the privacy of his/her own data. In addition, to preserve query privacy, authorized users require encrypting their queries before sending them to the cloud for evaluation. Furthermore, during query processing, the cloud can also derive useful and sensitive information about the actual data items by observing the data access patterns even if the data and query are encrypted [4],[5]. Therefore, following from the above discussions, securequery processing needs to guarantee

(1) confidentiality of the encrypted data
(2) confidentiality of a user's query record and
(3) Hiding data access patterns.

Using encryption as a way to achieve data confidentiality may cause another issue during the query processing step in the cloud. In general, it is very difficult to process encrypted data without ever having to decrypt it. The question here is how the cloud can execute the queries over encrypted data while the data stored at the cloud are encrypted at all times.In the literature, various techniques related to query processing over encrypted data have been proposed, including range queries [6]–[8] and other aggregate queries [9], [10]. However, these techniques are either not applicable or inefficient to solve advanced queries such as the k-nearest neighbour (kNN) query.

In this paper, we address the problem of secure processing of k-nearest neighbour query over encrypted data (SkNN) in the cloud. Given a user's input query Q, the objective ofthe SkNN problem is to securely identify the k-nearest data tuples to Q using the encrypted database of T in the cloud, without allowing the cloud to learn anything regarding the actual contents of the database T and the query record Q.More specifically, when encrypted data are outsourced to thecloud, we observe that an effective SkNN protocol needs to satisfy the following properties:

- Preserve the confidentiality of T and Q at all times
- Hiding data access patterns from the cloud
- Accurately compute the k-nearest neighbours of

Query Q

Incur low computation overhead on the end-user Hosting data-intensive query services in the cloud is increasingly popular because of the unique advantages in scalability and cost-saving. With the cloud infrastructures, the service owners can conveniently scale up or down the service and only pay for the hours of using the servers. This is an attractive feature because the workloads of query services are highly dynamic, and it will be expensive and inefficient to serve such dynamic workloads with in-house infrastructures [2]. However, because the service providers lose the control over the data in the cloud, data confidentiality and query privacy have become the major concerns. Adversaries, such as curious service providers, can possibly make a copy of the database or eavesdrop users' queries, which will be difficult to detect and prevent in the cloud infrastructures.

While new approaches are needed to preserve data confidentiality and query privacy, the efficiency of query services and the benefits of using the clouds should also be preserved. It will not be meaningful to provide slow query services as a result of security and privacy assurance. It is also not practical for the data owner to use a significant amount of in-house resources, because the purpose of using cloud resources is to reduce the need of maintaining scalable in-house infrastructures. Therefore, there is an intricate relationship among the data confidentiality, query privacy, the quality of service, and the economics of using the cloud.

We summarise these requirements for constructing a practical query service in the cloud as the CPEL criteria: data confidentiality, query privacy, efficient query processing, and low in-house processing cost. Satisfying these requirements will dramatically increase the complexity of constructing query services in the cloud. Some related approaches have been developed to address some aspects of the problem. However, they do not satisfactorily address all of these aspects. For example, the cryptoindex [12] and order preserving encryption (OPE) [1] are vulnerable to the attacks. The enhanced cryptoindex approach [14] puts heavy burden on the in-house infrastructure to improve the security and privacy. The New Casper approach [23] uses cloaking boxes to protect data objects and queries, which affects the efficiency of query processing and the inhouse workload. We have summarized the weaknesses of the existing approaches in Section 7.

We propose the random space perturbation (RASP) approach to constructing practical range query and knearest- neighbor (kNN) query services in the cloud. The proposed approach will address all the four aspects of the CPEL criteria and aim to achieve a good balance on them. The basic idea is to randomly transform the multidimensional data sets with a combination of order preserving encryption, dimensionality expansion, random noise injection, and random project, so that the utility for processing range queries is preserved. The RASP perturbation is designed in such a way that the queried ranges are securely transformed into polyhedra in the RASP-perturbed data space, which can be efficiently processed with the support of indexing structures in the perturbed space. The RASP kNN query service (kNN-R) uses the RASP range query service to process kNN queries. The key components in the RASP framework include

1. the definition and properties of RASP perturbation;
2. the construction of the privacy-preserving range query services;
3. the construction of privacy-preserving kNN query services; and
4. an analysis of the attacks on the RASP-protected data and queries.

In summary, the proposed approach has a number of unique contributions:

- The RASP perturbation is a unique combination of OPE, dimensionality expansion, random noise injection, and random projection, which provides strong confidentiality guarantee.
- The RASP approach preserves the topology of multidimensional range in secure transformation, which allows indexing and efficiently query processing.
- The proposed service constructions are able to minimise the in-house processing workload because of the low perturbation cost and high precision query results. This is an important feature enabling practical cloud-based solutions.

We have carefully evaluated our approach with synthetic and real data sets. The results show its unique advantages on all aspects of the CPEL criteria.

This paper is organized as follows: In Section 3, we define the RASP perturbation method, describe its major properties, and analyse the attacks to the RASP perturbed data. We also introduce the framework for constructing the query services with the RASP perturbation. In Section 4, we describe the algorithm for transforming queries and processing range queries. In Section 5, the range query service is extended to handle kNN queries. When describing these two services, we also analyze the attacks on the query privacy. Finally, we present some related approaches in Section 7 and analyze their weaknesses in terms of the CPEL criteria.

II. QUERY SERVICES IN CLOUD

This section presents the notations, the system architecture, and the threat model for the RASP approach, and prepares for the security analysis [3] in later sections. The design of the system architecture keeps the cloud economics in mind so that most data storage and computing tasks will be done in the cloud. The threat model makes realistic security assumptions and clearly defines the practical threats that the RASP approach will address.

A. Definitions and Notations

First, we establish the notations. For simplicity, we consider only single database tables, which can be the result of denormalization from multiple relations. A database table consists of n records and d searchable attributes. We also frequently refer to an attribute as a dimension or a column, which are exchangeable in the paper. Each record can be represented as a vector in the multidimensional space, denoted by low case letters. If a record x is d-dimensional, we say x 2 IRd, where IRd means the d-dimensional vector space. A table is also treated as a d _ n matrix, with records represented as column vectors. We use capital letters to represent a table, and indexed capital letters, for example, Xi, to represent columns. Each column is defined on a numerical domain. Categorical data columns are allows in range query, which are converted to numerical domains as we will describe in Section 3.

Range query is an important type of query for many data analytic tasks from simple aggregation to more sophisticated machine learning tasks. Let T be a table and Xi, Xj, and Xk be the real valued attributes in T, and a and b be some constants. Take the counting query for example. A typical range query looks like

select count(*),from T
where Xi 2 ½; bi_ and Xj

which calculates the number of records in the range defined by conditions on Xi, Xj, and Xk. Range queries may be applied to arbitrary number of attributes and conditions on these attributes combined with conditional operators "and"/"or." We call each part of the query condition that involves only one attribute as a simple condition. A simple condition like Xi 2 ½; bi_ can be described with two halfspace conditions Xi _ bi and _Xi _. Without loss of generality, we will discuss how to process half-space conditions like Xi _ bi in this paper. A slight modification will extend the discussed algorithms to handle other conditions like Xi < bi and Xi ¼ bi.

kNN query is to find the closest k records to the query point, where the euclidean distance is often used to measure the proximity. It is frequently used in locationbased services for searching the objects close to a query point, and also in machine learning algorithms such as hierarchical clustering and kNN classifier. A kNN query consists of the query point and the number of nearest neighbours, k.

B. System Architecture

We assume that a cloud computing infrastructure, such as Amazon EC2, is used to host the query services and large data sets. The purpose of this architecture is to extend the proprietary database servers to the public cloud, or use a hybrid private-public cloud to achieve scalability and reduce costs while maintaining confidentiality.

Each record x in the outsourced database contains two parts: the RASP-processed attributes D0 ¼ FðD;KÞ and the encrypted original records, Z ¼ EðD;K0Þ, where K and K0 are keys for perturbation and encryption, respectively. The RASP-perturbed data D0 are for indexing and query processing. Fig. 1 shows the system architecture for both RASP-based range query service and kNN service.

There are two clearly separated groups: the trusted parties and the untrusted parties. The trusted parties include the data/service owner, the in-house proxy server, and the authorized users who can only submit queries. The data owner exports the perturbed data to the cloud. Meanwhile, the authorized users can submit range queries or kNN queries to learn statistics or find some records. The untrusted parties include the curious cloud provider who hosts the query services and the protected database. The RASP-perturbed data will be used to build indices to support query processing.

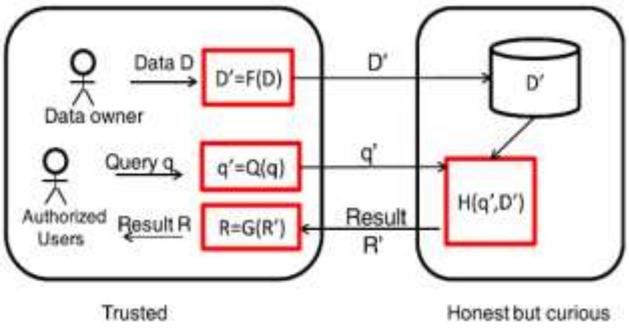

Fig. 1. The system architecture for RASP-based query services.

There are a number of basic procedures in this framework: 1) FðDÞ is the RASP perturbation that transforms the original data D to the perturbed data D0; 2) transforms the original query q to the protected form q0 that can be processed on the perturbed data; and 3) Hadoop is the query processing algorithm that returns the result R0. When the statistics such as SUM or AVG of a specific dimension are needed, RASP can work with partial homomorphic encryption such as Paillier encryption [24] to compute these statistics on the encrypted data, which are then recovered with the procedure GðR0Þ.

C. Threat Model

Assumptions. Our security analysis is built on the important features of the architecture. Under this setting, we believe the following assumptions are appropriate:

- Only the authorised users can query the proprietary database. Authorised users are not malicious and will not intentionally breach the confidentiality. We consider insider attacks are orthogonal to our research; thus, we can exclude the situation that the authorised users collude with the untrusted cloud providers to leak additional information.

- The client-side system and the communication channels are properly secured and no protected data records and queries can be leaked.

- Adversaries can see the perturbed database, the transformed queries, the whole query processing procedure, the access patterns, and understand the same query returns the same set of results, but nothing else.

- Adversaries can possibly have the global information of the database, such as the applications of the database, the attribute domains, and possibly the attribute distributions, via other published sources (e.g., the distribution of sales, or patient diseases, in public reports).

These assumptions can be maintained and reinforced by applying appropriate security policies. Note that this model is equivalent to the eavesdropping model equipped with the plaintext distributional knowledge in the cryptographic setting.

Protected assets. Data confidentiality and query privacy should be protected in the RASP approach. While the integrity of query services is also an important issue, it is orthogonal to our study. Existing integrity checking and preventing techniques [33], [29], [18] can be integrated into our framework. Thus, the integrity problem will be excluded from the paper, and we can assume the curious cloud provider is interested in the data and queries, but it will honestly follow the protocol to provide the infrastructure service.

Attacker modelling. The goal of attack is to recover (or estimate) the original data from the perturbed data, or identify the exact queries (i.e., location queries) to breach users' privacy. According to the level of prior knowledge the attacker may have, we categorise the attacks into two categories:

- Level 1: The attacker knows only the perturbed data and transformed queries, without any other prior knowledge. This corresponds to the cipertext-only attack in the cryptographic setting.

- Level 2: The attacker also knows the original data distributions, including individual attribute distributions and the joint distribution (e.g., the covariance matrix) between attributes. In practice, for some applications, whose statistics are interesting to the public domain, the dimensional distributions might have been published via other sources.

These levels of knowledge are appropriate according to the assumptions we hold. We will analyze the security based on this threat model.

Security definition. Different from the traditional encryption schemes, attackers can also be satisfied with good estimation. Therefore, we will investigate two levels of security definitions: 1) it is computationally intractable for the attacker to recover the exact original data based on the perturbed data; and 2) the attacker cannot effectively estimate the original data. The effectiveness measure is defined with the NR_MSE measure in Section 3.3.

IV. KNN QUERY PROCESSING

Because the RASP perturbation does not preserve distances (and distance orders), kNN query cannot be directly processed with the RASP perturbed data. In this section, we design a kNN query processing algorithm based on range queries (the kNN-R algorithm). As a result, the use of index in range query processing also enables fast processing of kNN queries.

A. Overview of the kNN-R Algorithm

The original distance-based kNN query processing finds the nearest k points in the spherical range that is centred at the query point. The basic idea of our algorithm is to use square ranges, instead of spherical ranges, to find the approximate kNN results, so that the RASP range query service can be used. There are a number of key problems to make this work securely and efficiently. 1) How to efficiently find the minimum square range that surely contains the k results, without many interactions between the cloud and the client? 2) Will this solution preserve data confidentiality and query privacy? 3) Will the proxy server's workload increase? to what extent?

The algorithm is based on square ranges to approximately find the kNN candidates for a query point, which are defined as follows.

Definition 1. A square range is a hypercube that is centred at the query point and with equal-length edges.

Fig. 5 illustrates the range-query-based kNN processing with 2D data. The Inner Range is the square range that contains at least k points, and the Outer Range encloses the spherical range that encloses the inner range. The outer range surely contains thekNNresults (see Proposition 2) but it may also contain irrelevant points that need to be filtered out.

Proposition 2. The kNN-R algorithm returns results with 100 percent recall.

Proof. The sphere in Fig. 5 between the outer range and the inner range covers all points with distances less than the radius r. Because the inner range contains at least k points, there are at least k nearest neighbours to the query points with distances less than the radius r. Therefore, the k nearest neighbours must be in the outer range.

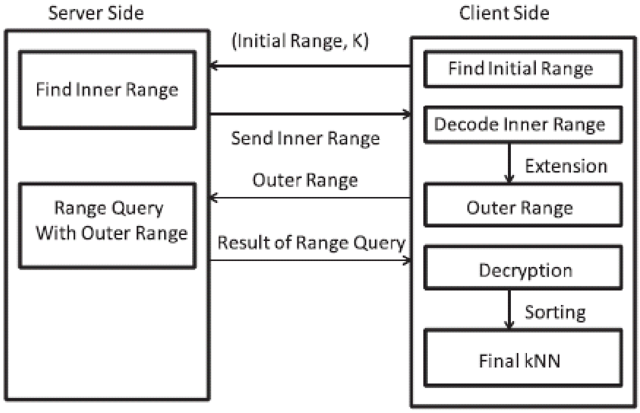

Fig. 2. Procedure of the KNN-R algorithm

The kNN-R algorithm consists of two rounds of interactions between the client and the server. Fig. 4 demonstrates the procedure. 1) The client will send the initial upper bound range, which contains more than k points, and the initial lower bound range, which contains less than k points, to the server. The server finds the inner range and returns to the client. 2) The client calculates the outer range based on the inner range and sends it back to the server. The server finds the records in the outer range and sends them to the client. 3) The client decrypts the records and find the top k candidates as the final result.

If the points are approximately uniformly distributed, we can estimate the precision of the returned result. With the uniform assumption, the number of points in an area is proportional to the size of the area. If the inner range contains m points, m >¼ k, the outer range contains q points, and the dimensionality is d, we can derive q ¼ 2d=2m. Thus, the precision is k=q ¼ k=ð2d=2mÞ. If m _ k and d ¼ 2, the precision is around 0.5. When d increases, the precision decreases exponentially due to the curse of dimensionality [22], which suggests kNN-R should not work effectively on high-dimensional data. We will show this weakness in experiments.

V. RELATED WORK

A. Protecting Outsourced Data

Order preserving encryption. Order preserving encryption [1] preserves the dimensional value order after encryption. It can be described as a function y, A well-known attack is based on attacker’s prior knowledge on the original distributions of the attributes. If the attacker knows the original distributions and manages to identify the mapping between the original attribute and its encrypted counterpart, a bucket based distribution alignment can be performed to break the encryption for the attribute [6]. There are some applications of OPE in outsourced data processing. For example, Yiu et al. [20] use a hierarchical space division method to encode spatial data points, which preserves the order of dimensional values and thus is one kind of OPE.

Cryptoindex. Cryptoindex is also based on column-wise bucketization. It assigns a random ID to each bucket; the values in the bucket are replaced with the bucket ID to generate the auxiliary data for indexing. To utilise the index for query processing, a normal range query condition has to be transformed to a set-based query on the bucket IDs. For example, Xi < ai might be replaced with X0i 2 ½ID1; ID2; ID3_. A bucket-diffusion scheme [14] was proposed to protect the access pattern, which, however, has to sacrifice the precision of query results, and thus increase the client’s cost of filtering the query result.

Distance-recoverable encryption. DRE is the most intuitive method for preserving the nearest neighbour relationship. Because of the exactly preserved distances, many attacks can be applied [19], [8]. Wong et al. suggest preserving dot products instead of distances to find kNN, which is more resilient to distance-targeted attacks. One drawback is the search algorithm is limited to linear scan and no indexing method can be applied.

B. Preserving Query Privacy

Private information retrieval (PIR) [9] tries to fully preserve the privacy of access pattern, while the data may not be encrypted. PIR schemes are normally very costly. Focusing on the efficiency side of PIR, Williams et al. use a pyramid hash index to implement efficient privacy preserving data-block operations based on the idea of Oblivious RAM. It is different from our setting of high throughput range query processing.

Papadopoulos et al. [25] use private information retrieval methods [9] to enhance location privacy. However, their approach does not consider protecting the confidentiality of data. Space Twist proposes a method to query kNN by providing a fake user’s location for preserving location privacy. But the method does not consider data confidentiality, as well. The Casper approach [23] considers both data confidentiality and query privacy, the detail of which has been discussed in our experiments.

C.Other Related Work

Another line of research facilitates authorized users to access only the authorized portion of data, for example, a certain range, with a public key scheme. However, the underlying encryption schemes do not produce indexable encrypted data. The setting of multidimensional range query in [28] is different from ours. Their approach requires that the data owner provides the indices and keys for the server, and authorized users use the data in the server. While in the cloud database scenario, the cloud server takes more responsibilities of indexing and query processing. Secure keyword search on encrypted documents [10], [5] scans each encrypted document in the database and finds the documents containing the keyword, which is more like point search in database. The research on privacy preserving data mining has discussed multiplicative perturbation methods [7], which are similar to the RASP encryption, but with more emphasis on preserving the utility for data mining.

V. CONCLUSION

The k-nearest neighbours is one of the commonly used query in many data mining applications. Under an out sourced database environment, where encrypted data are stored in the cloud, secure query processing over encrypted data becomes challenging. The existing SkNN techniques over encrypted data are not secure. In this paper, we proposed two novel SkNN protocols over encrypted data in the cloud. The first protocol, which acts as a basic solution, leaks some information to the cloud. On the other hand, our second protocol isfully secure, that is, it protects the confidentiality of the data,user's input query, and also hides the data access patterns.However, the second protocol is more expensive compared to the basic protocol. Also, we evaluated the performance of our protocols under different parameter settings. As a future work,we will investigate and extend our research to other complex conjunctive queries over encrypted data.

REFERENCES

[1] H. Hu, J. Xu, C. Ren, and B. Choi, "Processing private queries overuntrusted data cloud through privacy homomorphism," in ICDE. IEEE,2011, pp. 601–612.

[2] P. Mell and T. Grance, "The nist definition of cloud computing (draft),"NIST special publication, vol. 800, p. 145, 2011.

[3] M. Li, S. Yu, W. Lou, and Y. T. Hou, "Toward privacy-assured clouddata services with flexible search functionalities," in ICDCSW. IEEE,2012, pp. 466–470.

[4] P. Williams, R. Sion, and B. Carbunar, "Building castles out of mud:practical access pattern privacy and correctness on untrusted storage,"in CCS. ACM, 2008, pp. 139–148.

[5] M. S. Islam, M. Kuzu, and M. Kantarcioglu, "Access pattern disclosureon searchable encryption: Ramification, attack and mitigation," in NDSS,2012.

[6] B. Hore, S. Mehrotra, and G. Tsudik, "A privacy-preserving index forrange queries," in VLDB, 2004, pp. 720–731.

[7] E. Shi, J. Bethencourt, T.-H. Chan, D. Song, and A. Perrig, "Multidimensionalrange query over encrypted data," in IEEE Symposium onSecurity and Privacy (SP'07). IEEE, 2007, pp. 350–364.

[8] B. Hore, S. Mehrotra, M. Canim, and M. Kantarcioglu, "Secure multidimensionalrange queries over outsourced data," The VLDB Journal,vol. 21, no. 3, pp. 333–358, 2012.

[9] H. Hacıgum¨ us¸, B. Iyer, and S. Mehrotra, "Efficient execution of aggregation queries over encrypted relational databases," in Database Systemsfor Advanced Applications. Springer, 2004, pp. 125–136.

[10] E. Mykletun and G. Tsudik, "Aggregation queries in the database-as-aservicemodel," in Data and Applications Security XX. Springer, 2006,pp. 89–103.

[11] W. K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "Secure knncomputation on encrypted databases," in SIGMOD, 2009, pp. 139–152.

[12] Y. Zhu, R. Xu, and T. Takagi, "Secure k-nn computation on encryptedcloud data without sharing key with query users," in Cloud Computing.ACM, 2013, pp. 55–60.

[13] B. Yao, F. Li, and X. Xiao, "Secure nearest neighbor revisited," in IEEEICDE, Brisbane, Australia, April 2013.

[14] O. Goldreich, The Foundations of Cryptography. Cambridge, UniversityPress, 2004, vol. 2, ch. General Cryptographic Protocols, pp. 599–746.

[15] P. Paillier, "Public-key cryptosystems based on composite degree residuosityclasses," in EUROCRYPT. Springer-Verlag, 1999.

[16] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano,
"Heart disease data set," The UCI KDD Archive, 1988,
http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[17] J. Domingo-Ferrer, "A provably secure additive and multiplicative privacyhomomorphism," Information Security, pp. 471–483, 2002.

[18] M. Shaneck, Y. Kim, and V. Kumar, "Privacy preserving nearest neighbour search," Machine Learning in Cyber Trust, pp. 247–276, 2009.

[19] Y. Qi and M. J. Atallah, "Efficient privacy-preserving k-nearest neighbour search," in ICDCS. IEEE, 2008, pp. 311–319.

[20] J. Vaidya and C. Clifton, "Privacy-preserving top-k queries," in ICDE.IEEE, 2005, pp. 545–546.

[21] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, "Privatequeries in location based services: anonymizers are not necessary,"in SIGMOD. ACM, 2008, pp. 121–132.

[22] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexityof interactive proof systems," SIAM Journal of Computing, vol. 18, pp.186–208, February 1989.

[23] B. K. Samanthula and W. Jiang, "An efficient and probabilistic securebit-decomposition," in ACM ASIACCS, 2013, pp. 541–546.

[24] S. Bugiel, S. Nurnberger, A.-R. Sadeghi, and T. Schneider, "Twin clouds: ¨An architecture for secure cloud computing (extended abstract)," inWorkshop on Cryptography and Security in Clouds, March 2011.

[25] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, "Secure k-nearestneighbor query over encrypted data in outsourced environments," eprintarXiv:1307.4824, 2013.

Selected Paper from International Conference on Computing (NECICC-2k15)