International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 13
October 2017

Query Privacy And Data Confidentiality Using Random Space Perturbation In The Cloud

Sk.Rizwana & V Sriharsha
[1]PG Scholar, Dept of CSE, PACE Institute of Tech and Sciences, Vallur, Ongole, AP, India.
[2] Assistant Professor, Dept of CSE , D PACE Institute of Tech and Sciences, Vallur, Ongole, AP, India.

Abstract—*With the wide deployment of public cloud computing infrastructures, using clouds to host data query services has become an appealing solution for the advantages on scalability and cost-saving. However, some data might be sensitive that the data owner does not want to move to the cloud unless the data confidentiality and query privacy are guaranteed. On the other hand, a secured query service should still provide efficient query processing and significantly reduce the in-house workload to fully realize the benefits of cloud computing. We propose the random space perturbation (RASP) data perturbation method to provide secure and efficient range query and kNN query services for protected data in the cloud. The RASP data perturbation method combines order preserving encryption, dimensionality expansion, random noise injection, and random projection, to provide strong resilience to attacks on the perturbed data and queries. It also preserves multidimensional ranges, which allows existing indexing techniques to be applied to speedup range query processing. The kNN-R algorithm is designed to work with the RASP range query algorithm to process the kNN queries. We have carefully analyzed the attacks on data and queries under a precisely defined threat model and realistic security assumptions. Extensive experiments have been conducted to show the advantages of this approach on efficiency and security.*

Keywords —Query services in the cloud, privacy, range query, kNN query.

INTRODUCTION

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.We summarize these requirements for constructing a practical query service in the cloud as the CPEL criteria: data confidentiality, query privacy, efficient query proces-sing, 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 in-house 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 k-nearest-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 definition and properties of RASP perturbation;

1. the construction of the privacy-preserving range query services;
2. the construction of privacy-preserving kNN query services; and
3. 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 multi-dimensional range in secure transformation, which allows indexing and efficiently query processing.

. The proposed service constructions are able to minimize 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 their unique advantage 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 analyze 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 describ-ing 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.

2 QUERY SERVICES IN THE 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.

2.1 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 IR^d, where IR^d 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, X_i, to represent columns. Each column is defined on a numerical domain.

.

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

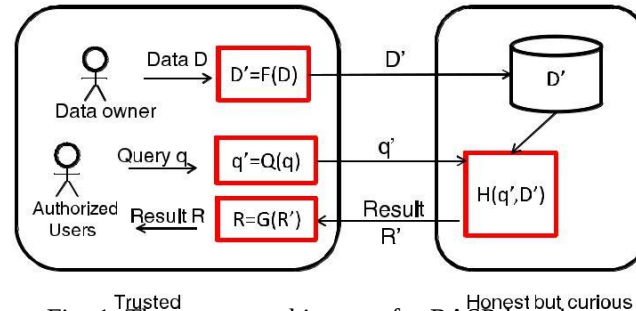

Fig. 1. The system architecture for RASP-based query services.

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.For an attack that can only result in low-accuracy estimation (e.g., NR_MSE 20%, the uncertainty is more than 20

percent of the domain length.), we call the RASP-perturbed data set is resilient to that attack. Intuitively, NR_MSE higher than 100 percent will not be very mean-ingful. Thus, we set the absolute upper bound to be 100 percent. We will discuss the specific upper bounds according to the level of prior knowledge.

3.3.2 Prior-Knowledge-Based Analysis

Below, we analyze the security under the two levels of knowledge the attacker may have, according to the two levels of security definitions: exact match and statistical estimation.Naive estimation. We assume each value in the vector or matrix is encoded with n bits. Let the perturbed vector p be drawn from a random variable P, and the original vector x be drawn from a random variable X. We show that naive estimation is computationally intractable to identify the exact original data with the perturbed data, if we use a random invertible real matrix generator and a random real-value generator. The goal is to show the number of valid X data set in terms of a known perturbed data set P . Below, we discuss a simplified version that contains no OPE component—the OPE version has at least the same level of security.

4 RASP RANGE-QUERY PROCESSING

Based on the RASP perturbation method, we design the services for two types of queries: range query and kNN query. This section will dedicate to range query processing. We will first show that a range query in the original space can be transformed to a polyhedron query in the perturbed space, and then we develop a secure way to do the query transformation. Then, we will develop a two-stage query processing strategy for efficient range query processing.

4.1Transforming Range Queries

Let's look at the general form of a range query condition. Let X_i be an attribute in the database. A simple condition in a range query involves only one attribute and is of the form "X_i<op>a_i," where a_i is a constant in the normalized domain of X_i and op 2 f<; >; ¼; ; ; ¼g6 is a comparison operator. For convenience, we will only discuss how to process X_i < a_i, while the proposed method can be slightly changed for other conditions.

Any complicated range query can be transformed into the disjunction of a set of conjunctions, i.e., $^{Sn}_{j¼1}ð^{Tm}_{i¼1}C_{i;j}Þ$, where m and n are some integers depending on the original query conditions and $C_{i;j}$ is a simple condition about X_i. Again, to simplify the presentation, we restrict our discussion to a single conjunc-tion condition $\^{m}_{i¼1}C_i$, where C_i is in form of b_i X_i a_i. Such a conjunction conditions describes a hypercubic area in the multidimensional space.According to the three nested transformations in RASP F ðGðE$_{ope}$ ðxÞÞÞ, we will first show that an OPE will trans-form the original hypercubic area to another hypercubic area in the OPE space. Proposition 1. Order preserving encryption functions transform a hypercubic query range to another hypercubic query range.Proof. The original range query condition consists of simple conditions like b_i X_i a_i for each dimension.

4.2 Security Enhancement on Query Transformation

The attacker may also target on the transformed queries. In this section, we discuss such attacks and describe the methods countering the attacks. Note that the attack on small ranges will be described in kNN query processing.Countering dimensional selection attack. We show that the dimensional selection attack can reveal partial information of the selected data dimensions, if the attacker knows the distribution of the dimension. Assume the query condition is applied to the ith dimension. If the query parameter w^T A ^{1} is directly submitted to the cloud side, the server can apply w^T A ^{1} to each record u in the server, and get w^T A ^{1}u ¼ E$_{ope}$ðx$_i$Þ E$_{ope}$ða$_i$Þ, where x_i is the ith dimension of the corresponding original record x. After getting all such values for the dimension i, with the known original data distributions, the attacker can apply the bucket-based distributional attack on the OPE encrypted data (see Section 7) to get an accurate estimate.According to the design of noise, the extended ðd þ 2Þth dimension v in the RASP perturbation: F ðxÞ ¼ AðE$_{ope}$ðx ÞT ; 1; vÞ^T is always greater than v_0, which can be used to construct secure query conditions. Instead of processing a half-space condition E$_{ope}$ðX$_i$Þ E$_{ope}$ða$_i$Þ, we use ðE$_{ope}$ðX$_i$Þ E$_{ope}$ða$_i$ÞÞðv v_0Þ 0 instead. These two conditions are equivalent because v always satisfies v > v_0. Using the similar transformations, we get E$_{ope}$ðX$_i$Þ E$_{ope}$ða$_i$Þ ¼ w^T A ^{1}u and v ¼ q^T A ^{1}u, where $q_{d\ þ2}$ ¼ 1, $q_{dþ1}$ ¼ v_0, and q_j ¼ 0, for j ¼6 d. Thus, we get the transformed

quadratic query condition. Let ¼ ðA ^{1}Þ^{T} wq^{T} A ^{1}. Now, is submitted to the server and the server will use u^{T} u 0 to filter out the results.Other potential threats. Because the query transformation method does not introduce randomness—the same query will always get the same transformation, and thus the confidentiality of access pattern is not preserved. We summarize the leaked information related to access patterns as follows:

. Attackers know the exact frequency of each transformed query.

. The set relationships (set intersection, union, difference, etc.) between the query results are revealed as a result of exact range query processing.

. Some query matrices on the same dimension may have special relationship preserved as shown in Proposition 3, which we will discuss later.We admit this is a weakness of the current design. However, according to the threat model, the adversary will not know any of the original data and queries. Thus, by simply observing the query frequency or relationships between queries, one cannot derive useful information. An important future work is to formally define the specific information leakage caused by the leaked query and access patterns, and then precisely analyze the data and query confidentiality affected by this information leakage under different security assumptions.

4.3 A Two-Stage Query Processing Strategy with Multidimensional Index Tree

With the transformed queries, the next important task is to process queries efficiently and return precise results to minimize the client-side postprocessing effects. A com-monly used method is to use multidimensional tree indices to improve the search performance. However, multidimen-sional tree indices are normally used to process axis-aligned "bounding boxes"; whereas the transformed queries are in arbitrary polyhedra, not necessarily aligned to axes. In this section, we propose a two-stage query processing strategy to handle such irregular-shape queries in the perturbed space.Multidimensional index tree. Most multidimensional in-dexing algorithms are derived from R-tree like algorithms [21], where the axis-aligned minimum bounding region (MBR) is the construction block for indexing the multi-dimensional data. For 2D data, an MBR is a rectangle. For higher dimensions, the shape of MBR is extended to hypercube. Fig. 2 shows the MBRs in the R-tree for a 2D data set, where each node is bounded by a node MBR. The R-tree range query algorithm compares the MBR and the queried range to find the answers.The two-stage processing algorithm. The transformed query describes a polyhedron in the perturbed space that cannot be directly processed by multidimensional tree algorithms. New tree search algorithms could be designed to use arbitrary polyhedron conditions directly for search. However, we use a simpler two-stage solution that keeps the existing tree search algorithms unchanged.At the first stage, the proxy in the client side finds the MBR of the polyhedron (as a part of the submitted transformed query) and submit the MBR and a set of secured query conditions f $_{1}$; . . . ; $_{m}$g to the server. The server then uses the tree index to find the set of records enclosed by the MBR.The MBR of the polyhedron can be efficiently founded based on the original range. The original query condition constructs a hypercube shape. With the described query transformation, the vertices of the hyper cube are also transformed to vertices of the polyhedron. Therefore, the MBR of the vertices is also the MBR of the polyhedron [26]. Fig. 3 illustrates the relationship between the vertices and the MBR and the two-stage processing strategy.At the second stage, the server uses the transformed half-space conditions to filter the initial result. In most cases of tight ranges, the initial result set will be reasonably small so that it can be filtered in memory by simply checking the transformed half-space conditions. However, in the worst case, the MBR of the polyhedron will possibly enclose the entire data set and the second stage is reduced to a linear scan of the entire data set. The result of second stage will return the exact range query result to the proxy server, which significantly reduces the postprocessing cost that the proxy server needs to take. It is very important to the cloud-based service, because low postprocessing cost requires low in-house investment.

5 KNN QUERY PROCESSING WITH RASP

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.

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 centered at the query point and with equal-length edges.

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 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 $2^{d=2}$m. Thus, the precision is k=q ¼ k=ð$2^{d=2}$mÞ. 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.

Proposition 3.

Proof. Remember that $_i$ for $X_i < c_i$ can be represented as ða_i $c_i a_{dþ1}$Þ^T ð$v_0 a_d$ þ1 $a_{dþ2}$Þ, where a_i is the ith row of the matrix A. Let the conditions be $X_i < h$, $X_i < l$, and

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 the kNN results (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 points, there are at least k nearest neighbors to the query points with distances less than the radius r. Therefore, the k nearest neighbors must be in the outer range.

As we have mentioned, the MBR of an arbitrary polyhedron can be derived based on the vertices of the polyhedron. A polyhedron is mapped to another polyhedron after the RASP perturbation. Concretely, let a polyhedron P has m vertices fx_1; x_mg, which are mapped to the vertices in the perturbed space: fy_1; . . . ; y_mg. Then, the upper bound and lower bound of dimension j of the MBR of the polyhedron in the perturbed space are determined by maxf_{yij}; i ¼ 1 . . . mg and minf_{yij}; i ¼ 1 . . . mg, respectively. Let the jth dimension of MBR^{ðLÞ} represented as ½$s^{ð}_{j;min}$^{LÞ};

5.4Defining Initial Bounds

The complexity of the ðk; Þ-range algorithm is determined by the initial bounds provided by the client. Thus, it is important to provide compact ones to help the server process queries more efficiently. The initial lower bound is defined as the query point. For qðq$_1$; . . . ; q_dÞ, the dimensional bounds are simply q_j X_j q_j.The higher bounds can be defined in multiple ways.

1) Applications often have a user-specified interest bound, for example, returning the nearest gas station in 5 miles, which can be used to define the higher bound. 2) We can also use center-distance-based bound setting. Let the query point has a distance to the distribution center—as we

always work on normalized distributions, the center is ð0; . . . ; 0Þ. The upper bound is defined as q_j X_j q_j þ

, where epsilon 2 ð0; 1& defines the level of conservativity. 3) If it is really expected to include all candidate kNN regardless how distant they are, we can include a rough density-map (a multidimensional histogram) for quickly identifying the appropriate higher bound. However, this method works best for low-dimensional data as the number of bins exponentially increases with the number of dimen-sions. In experiments, we simply use the method (1) and 5 percent of the domain length for the extension.

5.5 Security of kNN Queries

As all kNN queries are completely transformed to range queries, the security of kNN queries are equivalent to the security of range queries. According to the previous discussion in Section 4.2, the transformed range queries are secure under the assumptions. Therefore, the kNN queries are also secure. Detailed proofs have to be skipped for space limitation.

6 EXPERIMENTS

In this section, we present four sets of experimental results to investigate the following questions, correspondingly.

1. How expensive is the RASP perturbation?
2. How resilient the OPE enhanced RASP is to the ICA-based attack?
3. How efficient is the two-stage range query processing?
4. How efficient is the kNN-R query processing and what are the advantages?

6.1 Data Sets

Three data sets are used in experiments. 1) A synthetic data set that draws samples from uniform distribution in the range [0, 1]. 2) The Adult data set from UCI machine learning database.[5] We assign numeric values to the categorical values using a simple one-to-one mapping scheme, as described in Section 3. 3) The 2D NorthEast location data from rtreeportal.org.

6.2 Cost of RASP Perturbation

In this experiment, we study the costs of the components in the RASP perturbation. The major costs can be divided into two parts: the OPE and the rest part of RASP. We implement a simple OPE scheme [1] by mapping originalcolumn distributions to normal distributions. The OPE algorithm partitions the target distribution into buckets. Then, the sorted original values are proportionally parti-tioned according to the target bucket distribution to create the buckets for the original distribution. With the aligned original and target buckets, an original value can be mapped to the target bucket and appropriately scaled. Therefore, the encryption cost mainly comes from the bucket search procedure (proportional to log D, where D is the number of buckets). Fig. 6 shows the cost distributions for 20K records at different number of dimensions. The dimensionality has slight effects on the cost of RASP perturbation. Overall, the cost of processing 20K records is only around 0.1 second.

6.4 Performance of Two-Stage Range Query Processing

In this set of experiments, we study the performance aspects of polyhedron-based range query processing. We use the two-stage processing strategy described in Section 4, and explore the additional cost incurred by this processing strategy. We implement the two-stage query processing based on an R tree implementation provided by Dr. Hadjieleftheriou at AT&T Lab.[6] The block size is 4 KB and we allow each block to contain only 20 entries to mimic a large database with many disk blocks. Samples from the original databases in different size (10,000-50,000 records, i.e., 500-2,500 data blocks) are perturbed and indexed for query processing. Another set of indices is also built on the original data for the performance comparison with non-perturbed query processing. We will use the number of disk block accesses, including index blocks and data blocks, to assess the performance to avoid the possible variation caused by other parts of the computer system. In addition, we will also show the wall-clock time for some results.Recall the two-stage

processing strategy: using the MBR to search the indexing tree, and filtering the returned result with the secured query in quadratic form. We will study the performance of the first stage by comparing it to two additional methods: 1) the original queries with the index built on the original data, which is used to identify how much additional cost is paid for querying the MBR of the transformed query; 2) the linear scan approach, which is the worst case cost. Range queries are generated randomly within the domain of the data sets, and then transformed with the method described in Section 4. We also control the range of the queries to be [10, 20, 30, 40, and 50 percent] of the total range of the domain, to observe the effect of the scale of the range to the performance of query processing.

Results. The first pair of figures (the left subfigures of Figs. 8 and 9) shows the number of block accesses for 10,000 queries on different sizes of data with different query processing methods. For clear presentation, we use $\log_{10}$ (# of block accesses) as the y-axis. The cost of linear scan is simply the number of blocks for storing the whole data set. The data dimensionality is fixed to 5 and the query range is set to 30 percent of the whole domain. Obviously, the first stage with MBR for polyhedron has a cost much cheaper than the linear scan method and only moderately higher than R tree processing on the original data. Interestingly, different distributions of data result in slightly different patterns. The costs of R tree on trans-formed queries are very close to those of original queries for Adult data, while the gap is larger on uniform data. The costs over different dimensions and different query ranges show similar patterns.We also studied the cost of the second stage. We use "PrepQ" to represent the client-side cost of transforming queries, "purity" to represent the rate (final result count)/ (first stage result count), and records per query ("RPQ") to represent the average number of records per query for the first stage results. The quadratic filtering conditions are used in experiments. Table 1 compares the average wall-clock time (milliseconds) per query for the two stages, the RPQ values for stage 1, and the purity of the stage-1 result. The tests are run with the setting of 10K queries, 20K records, 30 percent dimensional query range and 5 dimensions. Since the second stage is done in memory, its cost is much lower than the first-stage cost. Overall, the two stage processing is much faster than linear scan and comparable to the original R Tree processing.

6.5 Performance of kNN-R Query Processing

In this set of experiments, we investigate several aspects of kNN query processing. 1) We will study the cost of (k,)-Range algorithm, which mainly contributes to the server-side cost. 2) We will show the overall cost distribution over the cloud side and the proxy server. 3) We will show the advantages of kNN-R over another popular approach: the Casper approach [23] for privacy-preserving kNN search.(k,)-range algorithms. In this set of experiments, we want to understand how the setting of the parameter affects the performance and the result precision. Fig. 10 shows the effect of setting to the ðk; Þ-range algorithm. Both data sets are 2D data. As becomes larger, both the precision and the number of rounds needs to reach the condition decreases. Note that each round corresponds to one server-side range query. The choice of represents a tradeoff between the precision and the performance. As we have discussed, the major weakness with the kNN-R algorithm is the precision reduction with increaseddimensionality. When the dimensionality increases, the precision can significantly drop, which will increase the cost of postprocessing in the client side. Fig. 11 shows this phenomenon with the real Adult data and the simulated uniform data. However, compared to the overall cost, the client-side cost increase is still acceptable. We will show the comparison next.Overall costs. Many secure approaches cannot use indices for query processing, which results in poor performance. For example, the secure dot-product approach [32] encodes the points with random projections and recovers dot-products in query processing for distance comparison. The way of encoding data disallows the index-based query processing. Without the aid of indices, processing a kNN query will have to scan the entire database, leaving many optimization impossible to implement.

7 RELATED WORK

7.1Protecting Outsourced Data

Order preserving encryption. Order preserving encryption [1] preserves the dimensional value order after encryption. It can be described as a function y ¼ F ðxÞ; 8x_i; x_j; x_i < ð>; ¼Þx_j , y_i < ð>; ¼Þy_j. A well-known attack is based on attacker's prior knowledge on the original distributions of the attributes. If the attacker knows the original distribu-tions 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 utilize 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, X_i < a_i might be replaced with X_i^0 2 ½ID_1; ID_2; 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 neighbor relationship. Because of the exactly preserved distances, many attacks can be applied [32], [19], [8]. Wong et al. [32] 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.

7.2Preserving 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, Williamst al. [31] 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. SpaceTwist [35] proposes a method to query kNN by providing a fake user's location for preserving location privacy. But the method does not consider data confidenti-ality, as well. The Casper approach [23] considers both data confidentiality and query privacy, the detail of which has been discussed in our experiments.

7.3Other Related Work

Another line of research [28] 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], [30],

8CONCLUSION

We propose the RASP perturbation approach to hosting query services in the cloud, which satisfies the CPEL criteria: data confidentiality, query privacy, efficient query processing, and low in-house workload. The requirement on low in-house workload is a critical feature to fully realize the benefits of cloud computing, and efficient query processing is a key measure of the quality of query services.RASP perturbation is a unique composition of OPE, dimensionality expansion, random noise injection, and random projection, which provides unique security fea-tures. It aims to preserve the topology of the queried range in the perturbed space, and allows to use indices for efficient range query processing. With the topology-preser-ving features, we are able to develop efficient range query services to achieve sublinear time complexity of processing queries. We then develop the kNN query service based on the range query service. The security of both the perturbed data and the protected queries is carefully analyzed under a

precisely defined threat model. We also conduct several sets of experiments to show the efficiency of query processing and the low cost of in-house processing.We will continue our studies on two aspects: 1) further improve the performance of query processing for both range queries and kNN queries; and 2) formally analyze the leaked query and access patterns and the possible effect on both data and query confidentiality.

REFERENCES

[1] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order Preserving Encryption for Numeric Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2004.

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.K. Andy Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Comput-ing," technical report, Univ. of Berkerley, 2009.

[3] J. Bau and J.C. Mitchell, "Security Modeling and Analysis," IEEE Security and Privacy, vol. 9, no. 3, pp. 18-25, May/June 2011.

[4] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge Univ. Press, 2004.

[5] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOMM, 2011.

[6] K. Chen, R. Kavuluru, and S. Guo, "RASP: Efficient Multi-dimensional Range Query on Attack-Resilient Encrypted Data-bases," Proc. ACM Conf. Data and Application Security and Privacy,

pp. 249-260, 2011.

[7] K. Chen and L. Liu, "Geometric Data Perturbation for Outsourced Data Mining," Knowledge and Information Systems, vol. 29, pp. 657-695, 2011.

[8] K. Chen, L. Liu, and G. Sun, "Towards Attack-Resilient Geometric Data Perturbation," Proc. SIAM Int'l Conf. Data Mining, 2007.

[9] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private Information Retrieval," ACM Computer Survey, vol. 45, no. 6,
pp. 965-981, 1998.

[10] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Con-structions," Proc. 13th ACM Conf. Computer and Comm. Security,

pp. 79-88, 2006.

[11] N.R. Draper and H. Smith, Applied Regression Analysis. Wiley, 1998.

[12] H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra, "Executing SQL over Encrypted Data in the Database-Service-Provider Model,"
Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD),
2002.

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer-Verlag, 2001.

[14] B. Hore, S. Mehrotra, and G. Tsudik, "A Privacy-Preserving Index for Range Queries," Proc. Very Large Databases Conf. (VLDB), 2004.

[15] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD Int'l Conf. Manage-ment of Data (SIGMOD), 2005.

[16] A. Hyvarinen, J. Karhunen, and E. Oja, Independent Component Analysis. Wiley, 2001.

[17] I.T. Jolliffe, Principal Component Analysis. Springer, 1986.

[18] F. Li, M. Hadjieleftheriou, G. Kollios, and L. Reyzin, "Dynamic Authenticated Index Structures for Outsourced Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.

[19] K. Liu, C. Giannella, and H. Kargupta, "An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining,"

Proc. 10th European Conf. Principle and Practice of Knowledge Discovery in Databases (PKDD), 2006.

[20] M.L. Liu, G. Ghinita, C.S. Jensen, and P. Kalnis, "Enabling Search Services on Outsourced Private Spatial Data," The Int'l J. Very Large Data Base, vol. 19, no. 3, pp. 363-384, 2010.

[21] Y. Manolopoulos, A. Nanopoulos, A. Papadopoulos, and Y. Theodoridis, R-Trees: Theory and Applications. Springer-Verlag, 2005.

[22] R. Marimont and M. Shapiro, "Nearest Neighbour Searches and the Curse of Dimensionality," J. Inst. of Math. and Its Applications, vol. 24, pp. 59-70, 1979.

[23] M.F. Mokbel, C. yin Chow, and W.G. Aref, "The New Casper: Query Processing for Location Services without Compromising Privacy," Proc. 32nd Int'l Conf. Very Large Databases Conf. (VLDB)
pp. 763-774, 2006.

[24] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes," Proc. 17th Int'l Conf. Theory and Application of Cryptographic Techniques (EUROCRYPT), pp. 223-238, 1999.

[25] S. Papadopoulos, S. Bakiras, and D. Papadias, "Nearest Neighbor Search with Strong Location Privacy," Proc. Very Large Databases Conf. (VLDB), 2010.

[26] F.P. Preparata and M.I. I, Computational Geometry: An Introduction.
Springer-Verlag, 1985.

Author Details:

Inkollu Uma Manikanta,
PG Scholar, Department of Computer Science and Engineering, PACE Institute of Tech and Sciences, Vallur, Ongole, AP, India..

DavarapalliAnandam, Department of Computer Science and Engineering, PACE Institute of Tech and Sciences, Vallur, Ongole, AP, India.He Completed B.Tech and M.Tech.He have 8 Years Teaching Experiance.His Interested area is Cloud Computing