ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 5, May 2015*

# Cloud Based Solutions to Safeguard Data Transfer

**Uma I Guranagoudr[1], Kumar H R[2], VivekBongale[3]**

M.Tech, Dept of Computer Science, SIET, Tumkur, India[1]

Assistant Professor, Dept of Computer Science, SIET, Tumkur, India[2]

M.Tech, Dept of Computer Science, SIET, Tumkur, India[3]

**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 data perturbation method to provide secure and efficient range query and kNN query services for protected data in the cloud. This 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.

**Keywords**: query services in the cloud, privacy, range query, kNN query

## I. INTRODUCTION

Hosting data-intensive query services in the cloud is increasingly popular because of the unique advantages in scalability & cost-saving. With 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 in efficient to serve such dynamic work loads with in-house infrastructures [2].

We summarize 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.

We propose the Random Space Perturbation(RASP) approach to constructing practical range query and k-nearest-neighbour (kNN) query services in the cloud. The proposed approach will address all the 2 four aspects of the CPEL criteria and aim to achieve a good balance on them. The basic idea is to randomly transform the multidimensional datasets 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 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 minimize the in-house processing workload because of the low perturbation cost and high precision query results.

## II. QUERY SERVICES IN THE CLOUD

### II.A. 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

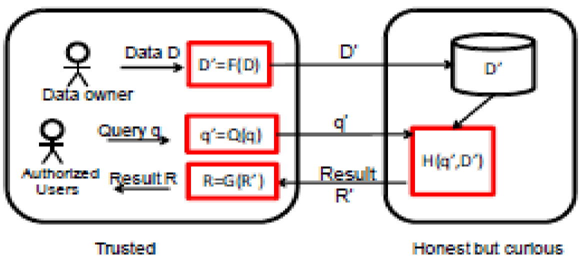

Fig.1.System architecture for RASP-based queryservices.

DOI 10.17148/IJARCCE.2015.45143

Record x in the out sourced database contains two parts: the RASP-processed attributes D′ =F(D,K) and the encrypted original records, Z =E(D,K′), where K and K′ are keys for perturbation and encryption, respectively. The RASP-perturbed data D′ are for indexing and query processing. Figure1 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 un trusted 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. Mean while, the authorized users can submit range queries or kNN queries to learn statistics or find some records. The un trusted 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.

There are a number of basic procedures in this framework: (1) F(D) is the RASP perturbation that transforms the original data D to the perturbed data D′; (2) Q(q) transforms the original query q to the protected form q′ that can be processed on the perturbed data; (3) H(q′,D′) is the query processing algorithm that returns the result R′.

II.B. 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 authorized users can query the proprietary database. Authorized 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 authorized users collude with
- Un trusted 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.

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.

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 categorize 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 ciper text-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.

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; (2) the attacker cannot effectively estimate the original data.

## III. RASP: RANDOMSPACEPERTURBATION

III.A. Definition of RASP

RASP is one type of multiplicative perturbation, witha novel combination of OPE, dimension expansion, random noise injection, and random projection. Let'sconsiderthe multidimensional data are numeric andin multidimensional vector space1.

The RASP perturbation involves three steps. Its security is based on the existence of random in vertiblereal-value matrix generator and random real value generator. For each k-dimensional input vector x.

1) An order preserving encryption (OPE) scheme[1], Eope with keys Kope, is applied to each dimension of x: Eope (x,Kope) ∈Rd to chang ethe dimensional distributions to normal distributions with each dimension's value order still preserved.
2) The vector is then extended to d+2 dimensions as G(x) = ((Eopt(x))T, 1, v)T , where the (d + 1)-th dimension is always a 1 and the (d + 2)-th dimension, v, is drawn from a random real number generator RNG that generates random values from a tailored normal distributions. We will discuss the design of RNG and OPE later.
3) The (d + 2)-dimensional vector is finally transformed to
4) F(x,K= {A,Kope,RG}) = A((Eope(x))T , 1, v)T, where A is a (d+2)×(d+2) randomly generated invertible matrix with aij∈R such that there are at least two non-zero values in each row of A and the last column of A is also non-zero2.

III.B. Properties of RASP

RASP has several important features. First, RASP doesnot preserve the order of dimensional values because of the

DOI 10.17148/IJARCCE.2015.45143

matrix multiplication component, which distinguishes itself from order preserving encryption(OPE) schemes, and thus does not suffer from the distribution-based attack. An OPE scheme maps a set of single-dimensional values to another, while keeping the value order unchanged.

Since the RASP perturbation can be treated as a combined transformation F(G(Eope(x))), it is sufficient to show that F(y) = Ay does not preserve the order of dimensional values, where y ∈Rd+2 and A ∈R(d+2)×(d+2). The proof is straightforward as shown in Appendix.

Second, RASP does not preserve the distances between records, which prevents the perturbed data from distance-based attacks [8].
2. Currently, we use a random invertible matrix generator that draws matrix elements uniformly at random from the standard normal distribution and check the matrix inevitability and the nonzero conditions.5

Third, the original range queries can be transformedto the RASP perturbed data space, which is the basisof our query processing strategy. A range query describes a hyper-cubic area (with possibly openbounds) in the multidimensional space.

III.C. Data Confidentiality Analysis
As the threat model describes, attackers might be interested in finding the exact original data records or estimating them based on the perturbed data. For estimation attack, if the estimation is sufficiently accurate (above certain accuracy threshold), we say the perturbation is not secure. Below, we define the measure for evaluating the effectiveness of estimation attacks.

III.C.a Evaluating Effectiveness of Estimation Attacks
Because attackers may not need to exactly recover theoriginal values, an accurate estimation will be sufficient.

A measure is needed to define the “accuracy” or “uncertainty” as we mentioned. We use the commonly used mean-squared-error (MSE) to evaluate the effectiveness of attack. To be semantically consistent, thej-th dimension can be treated as sample values drawn from a random variable Xj. Let xijbe the value of thei-th original record in j-th dimension and ˆxijbe the estimated value. The MSE for the j-th dimension can be defined as
MSE(Xj, ˆXj) =
1
n
n
X i
=
1
(xij− ˆxij)2,
which is equivalent to the variance: var(Xj− ˆXj). NR MSE(Xj) = 2qMSE(Xi,Xˆj)/domain length,(2)instead, which is intuitively the rate between theuncertain range and the whole domain.

To compare MSE for multiple columns, we also need to normalize these two series {xij} and {ˆxij}to eliminate the difference on domain scales. The normalization procedure [11] is described as follows. Assume the mean and variance of the series {xij} isμj and σ2j, correspondingly. The series is transformed by xij← (xij−μj)/σj. A similar procedure is also applied to the series {ˆxij}. For the normalized domains, the range [−2, 2] almost covers the whole population3[11]. Therefore, for normalized series, NR MSE is simply RMSE/2.

III. C.b. Prior-Knowledge Based Analysis
Below, we analyze the security under the two levelsof knowledge the attacker may have, according to thetwo levels of security definitions: exact match andstatistical estimation.

Naive Estimation.
The goal is to show the number of valid X datasetin terms of a known perturbed dataset P.

Proposition 1: For a known perturbed dataset P,there exists O(2(d+1)(d+2)n) candidate X datasets inthe original space. Proof: For a given perturbation P = AZ, whereZ is X with the two extended dimensions, we useBd+1 to represent the (d + 1)-th row of A−1. Thus,Bd+1P = [1, . . . ,1], i.e., the appended (d+1)-th row ofZ.

The total number of ˆB including non-invertibleones is 2(d+1)(d+2)n.

Thus, there are about (1 − exp−c(d+2))2(d+1)(d+2)n invertible ˆB. Correspondingly, there are a same number of candidate X. Thus, finding the exact X has a negligible probability in terms of the number of bits, n.

Distribution-based Estimation**.**
With the known distributional information, the attacker can do more on estimating the original data. The known most relevant method is called Independent Component Analysis(ICA) [17]. For a multiplicative perturbation P = AX,the basic idea is to find an optimal projection, wP, where w is a d + 2 dimension row vector, to result ina row vector with its value distribution close to thatof one original attribute. It can be extended to finda matrix W, so that WP gives independent and non-gaussian rows, i.e., a good estimate of X.

Proposition 2: There are O(2dn) candidate projection vectors, w, that lead to the same level of nongaussianity.
Proof: The OPE encrypted matrix ¯X (with the homogeneous dimension excluded, which can be possibly recovered) can be treated as a sample set drawn from a multivariate normal distribution N(μ,_). Anyinvertible transformation ¯ P = ¯ A ¯X will result in another multivariate normal distribution N(A¯μ,A¯_A¯T).

Thus, any projection w ¯ P will not change the gaussianity, and there are O(2dn) such candidates of w.Thus, the probability to identify the right projectionis

DOI 10.17148/IJARCCE.2015.45143

negligible in terms of the number of bits n. Thisshows that any ICA-style estimation that depends on non-guassianity is equally ineffective to the RASP perturbation.

## IV. RASP RANGE-QUERY PROCESSING

### IV.A. Transforming Range Queries

Let’s look at the general form of a range query condition. Let Xi be an attribute in the database. A simple condition in a range query involves only one attribute and is of the form “Xi <op>ai”, where aiis a constant in the normalized domain of Xi and op ∈ {<,>,=,≤,≥, 6=} is a comparison operator. For convenience we will only discuss how to process Xi <ai, 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., Snj=1(Tm i=1 Ci,j), where m, n are some integers depending on the original query conditions and Ci,jis a simple condition about Xi.Again, to simplify the presentation we restrict our discussion to a single conjunction condition ∩mi=1Ci, where Ciis in form of bi ≤ Xi ≤ ai.

Proposition 1: Order preserving encryption functions transform a hyper-cubic query range to another hyper-cubic query range.

Proof: The original range query condition consists of simple conditions like bi ≤ Xi ≤ aifor each dimension. Since the order is preserved, each simple condition is transformed as follows: Eope(bi) ≤ Eope(Xi) ≤ Eope (ai), which means the transformed range is still a hyper-cubic query range.

Let y = Eope(x) and ci = Eope(ai). A simple condition Yi ≤ ci defines a half-space. With the extended dimensions zT= (yT, 1, v), the half-space can be represented as wTz ≤ 0, where w is a d + 2 dimensional vector with wi= 1,wd+1 = −ci, and wj= 0 for j 6= i, d + 1. Finally, let u = Az, according to the RASP transformations

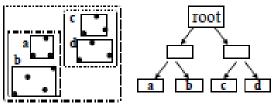

Fig. 2. R-tree index.

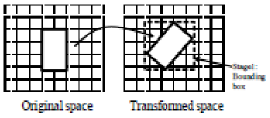

Fig. 3. Illustration of the two-stage processing algorithm.

### IV.B. Security Enhancement on Query Transformation

The attacker may also target on the transformed queries.

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 i-th dimension. If the query parameter wTA−1 is directly submitted to the cloud side, the server can apply wTA−1 to each record u in the server, and get wTA−1u = Eope(xi) − Eope(ai), where xi is the ithdimension of the corresponding original record x.

According to the design of noise, the extended (d+ 2)-th dimension v in the RASP perturbation: F(x) = A(Eope(x)T , 1, v)T is always greater than v0, which can be used to construct secure query conditions. Instead of processing a half space condition Eope(Xi) ≤ Eope(ai), we use (Eope(Xi) − Eope(ai))(v − v0) ≤ 0 instead. These two conditions are equivalent because v always satisfies v > v0. Using the similar transformations, we get Eope(Xi) − Eope(ai) = wTA−1u and v = qTA−1u, where qd+2 = −1, qd+1 = v0, and qj= 0, for j 6= d. Thus, we get the transformed quadratic query condition uT(A−1)TwqTA−1u ≤ 0. (4) randomly chosen for each record, the value Eope(Xi) − Eope(ai) is protected by the randomization. _i does not reveal the key parameters as well. Let ci = Eope(ai) and aibe the i-th row of A−1. _i is (ai−ciad+1)T(v0ad+1 −ad+2). As all the components: ai, ci, ad+1, and ad+2 are unknown and cannot be further reduced, _i provide no information to help drive information about A−1.

Other Potential Threats**.**

Because the query transformationmethod does not introduce randomness – thesame 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.
- We admit this is a weakness of the current design. 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.

### IV.C. A Two-Stage Query Processing Strategy with Multidimensional Index Tree

With the transformed queries, the next important taskis to process queries efficiently and return precise results to minimize the client-side post-processin effects. A commonly used method is to use multidimensional tree indices to improve the search performance.

Multidimensional Index Tree.

Most multidimensional indexing algorithms are derived from R-tree like algorithms [22], where the axis-aligned minimum8bounding region (MBR) is the construction

DOI 10.17148/IJARCCE.2015.45143

block for indexing the multidimensional data. For 2D data, an MBR is a rectangle. For higher dimensions, the shapeof MBR is extended to hyper-cube. Figure 2 shows the MBRs in the R-tree for a 2D dataset, 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 transformedquery describes a polyhedron in the perturbed space that cannot be directly processed by multidimensionaltree 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 {_1, . . . ,_m} to the server.
The server then uses the tree index to find the set of records enclosed by the MBR.

At the second stage, the server uses the transformed half space conditions to filter the initial result. In mostcases of tight ranges, the initial result set will be reasonably small so that it can be filtered in memoryby simply checking the transformed half-space conditions. However, in the worst case, the MBR ofthe polyhedron will possibly enclose the entire dataset and the second stage is reduced to a linear scan of the entire dataset. The result of second stage will return the exact range query result to the proxy server, which significantly reduces the post-processing cost that the proxy server needs to take. It is very important to the cloud-based service, because low post-processing cost requires low in-house investment.

## V. 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 akNN query processing algorithm based on range queries (the kNN-R algorithm).As a result, the use of index in range query processingalso 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 centered 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?

Definition 1: A square range is a hyper-cube that is centred at the query point and with equal-length edges.
Figure 5 illustrates the range-query-based kNN processing with two-dimensional data. The Inner Range is the square range that contains at least k points,and the Outer Range encloses the spherical rangethat encloses the inner range. The outer range surely contains the kNN results (Proposition 2) but it mayalso contain irrelevant points that need to be filteredout.

Proposition 2: The kNN-R algorithm returns results with 100% recall.
Proof: The sphere in Figure 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.

The kNN-R algorithm consists of two rounds ofinteractions between the client and the server. Figure 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 theinner range and sends it back to the server. The serverfinds the records in the outer range and sends themto the client. (3) The client decrypts the records andfind the top k candidates as the final result.

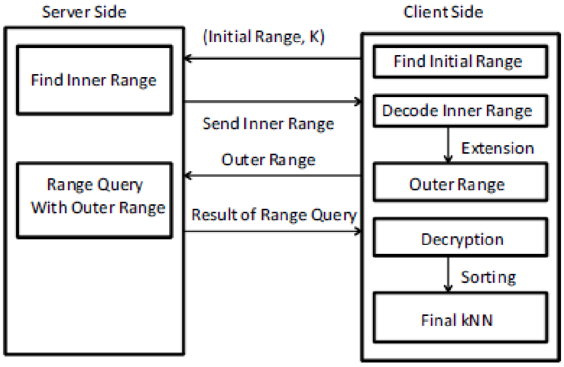

Fig. 4. Procedure of KNN-R algorithm

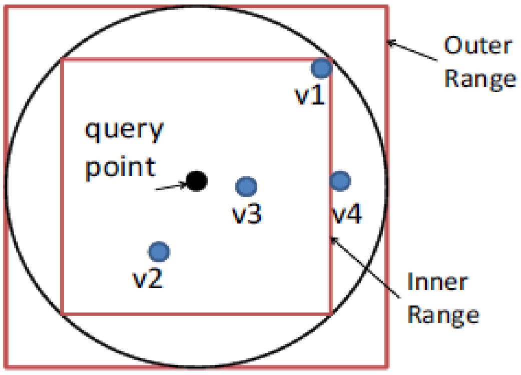

Fig. 5. Illustration for kNN-R Algorithm when k=3.

### B. Finding Compact Inner Square Range

An important step in the kNN-R algorithm is to find the compact inner square range to achieve highprecision. In the following, we give the (k, δ)-rangefor efficiently finding the compact inner range.

Definition 2: A (k, δ)-range is any square range centred at the query point, the number of points inwhich is in the range [k, k + δ], δ is a nonnegativeinteger.

We design an algorithm similar to binary searchto efficiently find the (k, δ)-range. Suppose a square range centred at the query point with length of Lin each dimension is represented as S(L). Let thenumber of points included by this range is N(L). Ifa square range S(in) is enclosed by another squarerange S(out), we say S(in) ⊂S(out). It directly followsthat N(in) ≤ N(out), and also
Corollary 1: If N(1) < N(2), S(1) ⊂S(2).

Using this definition and notation, we can always construct a series of enclosed square ranges centred on the query point: S(L1) ⊂S(L2) ⊂. . . ,⊂S(Lm).Correspondingly, the numbers of points enclosed by{S(Li)} have the ordering N(L1) ≤ N(L2) ≤ . . .N(Lm).

Selection of Initial Inner/Outer Bounds.
The selectionof initial inner bound can be the query point. If the query point is q(q1, . . . , qd), S(L1) is a hypercubedefined by {qi ≥ Xi ≥ qi, i = 1 . . . d}. The naïve selection of S(Lm) would be the whole domain.

### C. Finding Inner Range with RASP PerturbedData

Algorithm 4 gives the basic ideas of finding the compact inner range in iterations. There are two criticaloperations in this algorithm: (1) finding the numberof points in a square range and (2) updating the higherand lower bounds. Because range queries are securedin the RASP framework, the key is to update thebounds with the secured range queries, without thehelp of the client-side proxy server.
The problem of binary range search is to use thehigher bound range S(high) and the lower bound range S(low) to derive S(mid). When all of these ranges are secured, the problem is transformed to
(1) deriving _(mid)
ifrom _(high)
iand _(low)
i ; and (2)
deriving MBR(mid) from MBR(high) and MBR(low). The following discussion will be focused on the simplified RASP version without the OPE component, which will be extended with the OPE component.We show that
Proposition 3:
(_(high)
i + _(low)
i )/2 = _(mid)
i .
10

Proof: Remember that _i for Xi < ci can be represented as (ai− ciad+1)T(v0ad+1 − ad+2), where aiis the i-th row of the matrix A. Let the conditions be Xi < h, Xi < l, and Xi <(h+l)/2 for the high, low, and middle bounds, correspondingly. Thus, (_(high)i + _(low)i )/2 = (ai− ((h + l)/2)ad+1)T (v0ad+1 − ad+2), which is _(mid)i .

As we have mentioned, the MBR of an arbitrary polyhedron can be derived based on the vertices of the polyhedron.

Let the j-th dimension of MBR(L) represented as [s(L) j,min, s(L) j,max], where s(L) j,min= min{y(L) ij, i = 1 . . .m},and s(L) j,max= max{y(high) ij, i = 1 . . .m}. Now wechoose the MBR(MID) as follows: for j-th dimension we use [(s(low)j,min+ s(high)j,min)/2, (s(low)j,max+ s(high)j,max)/2].

Proposition 4: MBR(MID) encloses MBR(mid). The details of proof can be found in Appendix. Because the MBR is only used for the first stage of range query processing, a slightly larger MBR still encloses the polyhedron, which guarantees the correctness of the two-stage range query processing. Including the OPE component**.** The results on _(mid)i and MBR(MID) can be extended to the RASP scheme with the OPE component. Let the condition for the “between” bound be Xi < b that satisfies fi(b) = (fi(h) + fi(l))/2. According tothe OPE property, we have l < b < h, i.e., thecorresponding range is still between the lower rangeand higher range. Therefore, the same binary searchalgorithm can still be applied, according to Corollary1. The server can also derive (_(high)i + _(low) i )/2 =(ai− ((fi(h) + fi(l))/2)ad+1)T (v0ad+1 − ad+2) = _btwi,a result similar to Proposition 3.Similarly, we define MBR(BTW) withfi(s(BTW) i,max) = (fi(s(low)i,max) + fi(s(high)i,max))/2 and fi(s(BTW)i,min) = (fi(s(low)i,min) + fi(s(high) i,min))/2, whileMBR(btw) is defined based on the vertices to be Consistent with _(btw) i .

### D. Defining Initial Bounds

The complexity of the (k, δ)-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.Forq(q1, . . . , qd), the dimensional bounds are simply qj≤ Xj≤ qj. 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 γto the distribution center - as we always work on normalized distributions, the center is (0, . . . ,0). The upper bound is defined as qj− ϱγ≤ Xj≤ qj+ ϱγ,
whereepsilon ∈(0, 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 histgram) for quickly identifying the appropriate higher bound.

### E. 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

## VI. EXPERIMENTS

In this section, we present four sets of experimentalresults 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?

### VI.A. Datasets

Three datasets are used in experiments. (1) A syntheticdataset that draws samples from uniform distribution in the range [0, 1]. (2) The Adult dataset from UCI machine learning database5. We assign numeric values to the categorical values using a simple one-to- one mapping scheme, as described in Section 3.(3) The 2-dimensional NorthEast location data from treeportal.org.

5. http://archive.ics.uci.edu/ml/
11

### VI.B. Cost of RASP Perturbation

In this experiment, we study the costs of the componentsin the RASP perturbation. The major costs can be dividedinto two parts: the OPE and the restpart of RASP. We implement a simple OPE scheme [1] by mapping original column distributions to normal distributions. The OPE algorithm partitions the target distribution into buckets. Then, the sorted original values are proportionally partitioned 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 logD, where Dis the number of buckets). Figure 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.

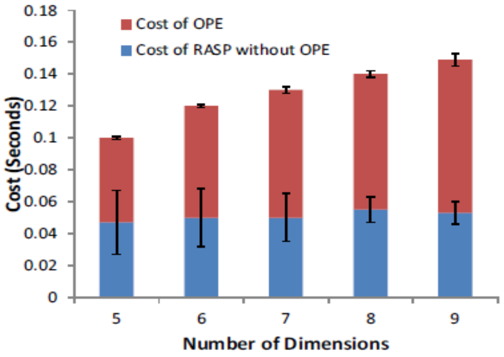

Fig.6:The cost distribution of the full RASP scheme Data:Adult(20K records,5-9dimensions)

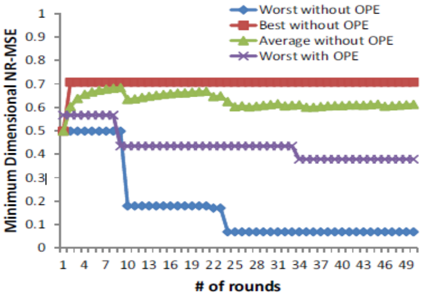

Fig.7:Randomly generated matrix Aand the progressive resilience to ICA attack Data Adult(10 dimensions,10Krecords)

### VI.C. Resilience to ICA Attack

We have discussed the methods for countering theICA distributional attack on the perturbed data. In this set of experiments, we evaluate how resilient the RASP perturbation is to the distributional attack.

**Results.**We simulate the ICA attack for randomlychosen matrices A. The data used in the experimentis the 10-dimensional Adult data with 10K records.

Figure 7 shows the progressive results in a numberof randomly chosen matrices A. The x-axis representsthe total number of rounds for randomly choosingthe matrix A; the y-axis represents the minimumdimensional NR MSE among all dimension. WithoutOPE, the label “Best-without-OPE” representsthe most resilient A at the round i, “Worst-without-OPE” represents the A of the weakest resilience, and“Average-without-OPE” is the average quality of thegenerated A matrices for i rounds.We see that the bestcase is already close to the upper bound 0.7 (Section III.C). With the OPE component, the worst case can alsobe significantly improved.

### VI.D. Performance of Two-stage Range Query Processing

In this set of experiments, we study the performanceaspects of polyhedron-based range query processing.We use the two-stage processing strategy described inSection 4, and explore the additional cost incurred bythis processing strategy.

**Results.**The first pair of figures (the left subfigures ofFigure 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 log10(# of block accesses) as the y-axis.

The cost of linear scan is simply the number of blocks for storing the whole dataset. The data dimensionality is fixed to 5 and the query range is set to 30% 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 transformed queries Wall clock cost distribution (milliseconds) and comparison.
We also studied the cost of the second stage. We use “PrepQ” to represent the client-side cost of transforming

DOI 10.17148/IJARCCE.2015.45143

queries, "purity" to represent the rate (final result count)/(1st 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% dimensional query range and 5 dimensions. Since the 2nd stage is done in memory, its cost is much lower than the 1st-stage cost. Overall, the two stage processing is much faster than linear scan and comparable to the original R*Tree processing.

VI.E. 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, δ)-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 [24] for privacy-preserving kNN search. **(k, δ)-**Range AlgorithmsIn this set of experiments, we want to understand how the setting of the δparameter affects the performance and the result precision. Figure 10 shows the effect of δ setting to the (k, δ)-range algorithm. Both datasets are twodimensional data. As δ becomes larger, both the precision and the number of rounds needs to reach the δ condition decreases. Note that each round corresponds to one server-side range query. The choice of δ represents a trade-off between the precision and the performance.

Comparing kNN-R with the Casper Approach**.**In this set of experiments, we compare our approach and the Casper approach with a focus on the trade off between the data confidentiality and the query result precision (which indicates the workload of the in-house proxy). Based on the description in the paper [24], we implement the 1NN query processing algorithm for the experiment.

## VII. CONCLUSION

We propose the RASP perturbation approach to hosting query services in the cloud, which satisfies theCPEL 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 features. 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-preserving features, we are able to develop efficient range query services to achieve sub lineartime 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; (2) formally analyze the leaked query and access patterns and the possible effect on both data and query confidentiality.

## REFERENCES

[1]. [1] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in Proceedings of ACM SIGMOD Conference, 2004.
[2]. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. K. andAndy Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," Technical Report, University 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, 2011.
[4]. S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.
[5]. N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacypreserving multi-keyword ranked search over encrypted cloud data," in INFOCOMM, 2011.
[6]. K. Chen, R. Kavuluru, and S. Guo, "Rasp: Efficient multidimensional range query on attack-resilient encrypted databases," in ACM Conference on Data and Application Security and Privacy, 2011, pp. 249–260.
[7]. K. Chen and L. Liu, "Geometric data perturbation for outsourced data mining," Knowledge and Information Systems, 2011.
[8]. K. Chen, L. Liu, and G. Sun, "Towards attack-resilient geometric data perturbation," in SIAM Data Mining Conference, 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,

**DOI 10.17148/IJARCCE.2015.45143**