RASP Data Perturbation – An Efficient Query Services in Cloud

Naveen Kumar M
Computer Science & Engineering,
PES College of Engineering,
Mandya, Karnataka.

***Abstract* – Hosting a data query services in cloud using public cloud computing infrastructures which has been deployed world-wide has become an appealing solution for advantages on scalability and cost-saving. Data owner does not want to move the sensitive data to cloud unless data confidentiality and query privacy 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. Attacks on data and queries have been carefully analyzed 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.**

Index Terms – Range query, kNN Query, Query Services in Cloud.

1 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]. 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.

1.1 Purpose:

The purpose of this project is to implement a **Random Space Perturbation** (RASP) data perturbation method to provide secure and efficient **Range Query** and **kNN** query services for protected data in the cloud.

1.2 Objective:

The objective of this project is to implement a secured and efficient query services in cloud by implementing a RASP framework which includes both Range Query and kNN query services for satisfying the following aspects:

- Data Confidentiality.
- Query Privacy.
- Quality of Service [Query Efficiency].
- Economics of using cloud [Low Cost].

2 LITERATURE SURVEY

2.1 *Existing System*

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. 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.

2.1.1 *Drawbacks of Existing System:*

- Cryptoindex[12], Order Preserving Encryption(OPE) [1] are vulnerable to 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.
- In OPE approach 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].
- In DRE approach drawback is the search algorithm is limited to linear scan and no indexing method can be applied.
- Private Information Retrieval(PIR) is very high cost and the techniques of PIR such as Pyramid hash Index[31], Space Twist[35] only provide efficient query privacy but these techniques doesn't provide data confidentiality.

2.2 *Proposed System:*

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.

2.2.1 *Advantages of Proposed System:*

- 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. This is an important feature enabling practical cloud-based solutions.
- RASP Perturbation technique is designed to meet all the four aspects which are defined as CPEL Criteria: i.e., Data Confidentiality, Query Privacy, Efficient Query Processing, and Low Cost Processing.

3 QUERY SERVICES IN CLOUD

This section presents the notations, the system architecture, and the threat model for the RASP approach. 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.

3.1 *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 € 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, 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 4. 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 € [ai, bi] and Xj € (aj, bj) and
Xk = ak;
```

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 € [ai, bi]can be described with two halfspace conditions Xi <= bi and –Xi <= ai. 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 = 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 neighbors, k.

3.2 *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 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 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.

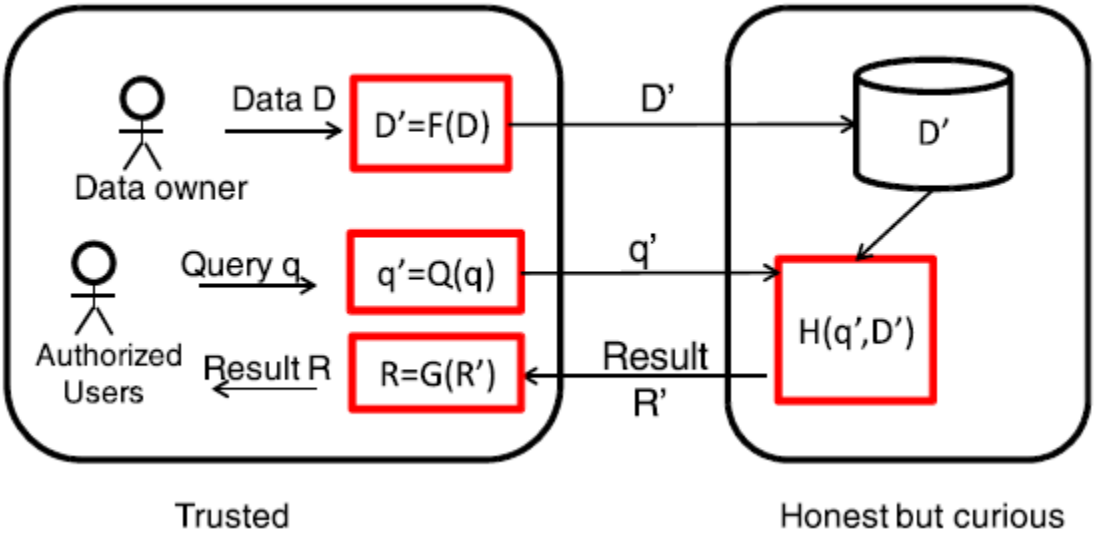

Fig.1. The system architecture for RASP-based query services

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.

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'.

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 G (R').

3.3 *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 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).

.

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 modeling.- 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 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.

2) The attacker cannot effectively estimate the original data. The effectiveness measure is defined with the NR_MSE measure

4 *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.

4.1 Overview of 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. ?

Procedure of kNN-R Algorithm

The kNN-R algorithm consists of two rounds of interactions between the client and the server. Following Fig. 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.

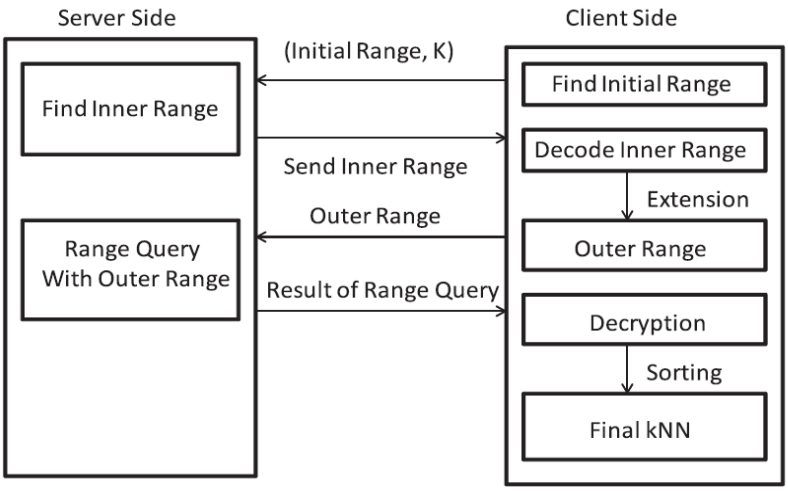

Fig 2: Procedure of kNN-R Algorithm.

5.2 Finding Compact Inner Square Range

An important step in the kNN-R algorithm is to find the compact inner square range to achieve high precision.

In the following, we give the (k, δ)-range for efficiently finding the compact inner range.

Definition . A (k ,δ) -range is any square range centered at the

query point, the number of points in which is in the range [k + k δ]is a nonnegative integer.

We design an algorithm similar to binary search to efficiently find the (k, δ) -range. Suppose a square range centered at the query point with length of L in each dimension is represented as S ^ (L). Let the number of points included by this range is N ^ (L). If a square range S ^ (in) is enclosed by another square range S ^ (out), we say S ^ (in) S ^ (out). It directly follows that NðinÞ _ NðoutÞ, and also

Corollary 1. If N ^(1) < N ^ (2), S ^(1) subset S ^ (2).

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

Assume that S ^ (L1) is the initial range containing less than k points and S ^ (Lm) is the initial upper bound range; both are sent by the client. The problem of finding the compact inner range S can be mapped to a binary search over the sequence {S ^ (Li) }.

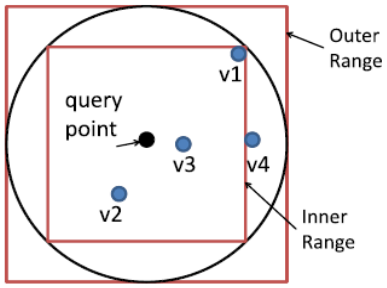

Fig 3: Illustration of kNN-R Algorithm when k=3

In each step of the binary search, we start with a lower bound range, denoted as S ^ (low) and a higher bound range, S ^ (high). We want the corresponding numbers of enclosed points to satisfy N ^ (low) < k <= N ^ (high) in each step, which is achieved with the following procedure. First, we find the middle square range S ^ (mid), where mid=(low=+high)/2. If S ^(mid) covers no less than k points, the higher bound: S ^(high) is updated to S ^ (mid); otherwise, the lower bound: S ^ (low) is updated to S ^ (mid). At the beginning step S ^ (low) is set to S ^ (L1) and S ^ (high) is S ^ (Lm). This process repeats until N ^ (mid) < k + δ or high - low < E , where E is some small positive number.

Selection of initial inner/outer bounds. The selection of initial inner bound can be the query point. If the query point is q(q1; . . . ; qd), S ^ (L1) is a hypercube defined by {qi >= Xi >=
qi , i = 1 . . . d}. The naive selection of S ^ (Lm) would be the whole domain. However, we can effectively reduce the range with a coarse density map organized in a tiny flat multidimensional tree, which can be included in the preprocessing step in the client side.

6 EXPERIMENTS:

In this section we present the result for the

1) How expensive is the RASP perturbation?

2) How efficient is the kNN-R query processing and what are the advantages?

6.1 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 original

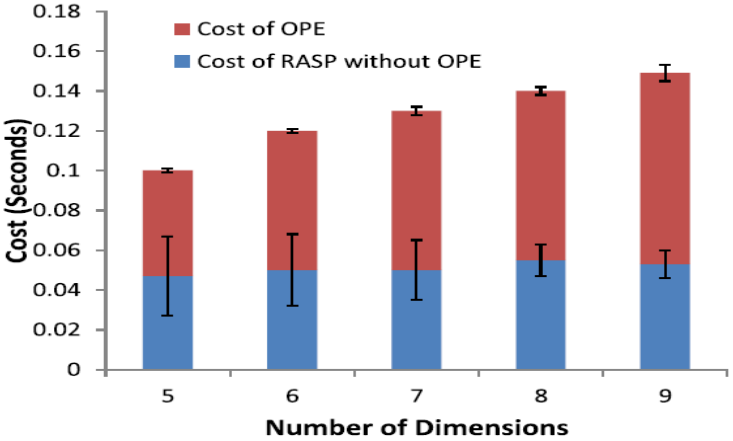

Fig 4: Cost distribution of the full RASP scheme.

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 D is the number of buckets). Fig 4: 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.

6.2 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 serverside 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 approrch [23] for privacy-preserving kNN search.

(k, δ)-range algorithms. In this set of experiments, we want to understand how the setting of the δ parameter affects the performance and the result precision. Fig 5 : shows the effect of δ setting to the (k , δ) -range algorithm. Both data sets are 2D 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 serverside range query. The choice of δ 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 increased dimensionality. When the dimensionality increases, the precision can significantly drop, which will increase the cost of postprocessing in the client side. Fig. 6 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 dotproducts 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. One concern with the kNN-R approach is the workload on the proxy server. Different from range query, the proxy

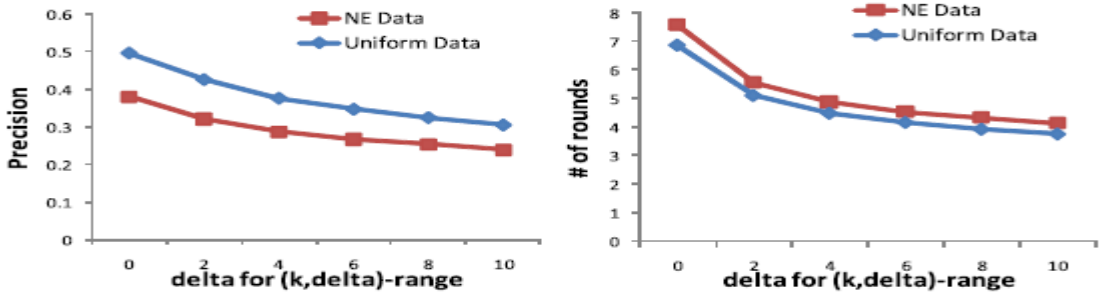

Fig 5: Performance and result precision for different δ setting of the (k, δ)-range algorithm for 2D data.

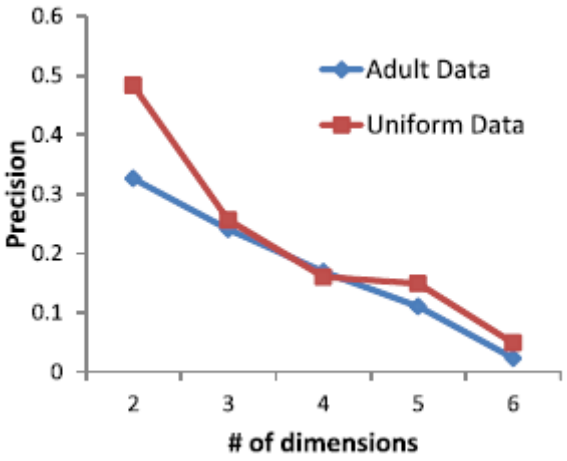

Fig 6: Precision reduction with more dimension

server will need to filter out the points returned by the server to find the final kNN. A reduced precision due to the increased dimensionality will imply an increased burden for the proxy server. We need to show how significant this proxy cost is.

Per-Query Performance Comparison (Milliseconds) between Linear Scan on the Original Nonperturbed Data and Index-Aided kNN-R Processing on Perturbed Data

Data& setting	LinerScan	Pre-proc.	Server Cost	Post-proc.
Uniform2D	27.37	0.01	13.54	0.04
Adult2D	26.09	0.01	14.48	0.06
Uniform5D	33.03	0.01	13.79	0.34
Adult5D	31.96	0.01	2.56	0.05

We use the database of 100 thousands of data points and 1,000 randomly selected queries for the 1NN experiment. The wall clock time (milliseconds) is used to show the average cost per query in Table 2. We also list the cost of the secure dot-product method [32] for comparison. Table 2 shows that the proxy server takes a negligible preprocessing cost and a very small postprocessing cost, even for reduced precision in the 5D data sets. We use 5 percent domain
length to extend the query point to form the initial higher bound. Compared to the dot-product method, the userspecified higher bound setting can cut off uninteresting regions, giving significant performance gain for sparse or skewed data sets, such as Adult5D. This cut-off effect cannot be implemented with the dot-product method. Furthermore, even for dense cases like the 2D data sets, the overall cost is only about half of the dot-product method.

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 tradeoff 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 [23], we implement the 1NN query processing algorithm for the experiment.

The Casper approach uses cloaking boxes to hide both the original data points in the database and the query points. It can also use the index to process kNN queries. The confidentiality of data in Casper is solely defined by the size of cloaking box. Roughly speaking, the actual point has the same probability to be anywhere in the cloaking box. However, the size of cloaking box also directly affects the precision of query results. Thus, the decision on the box size represents a tradeoff between the precision of query results and the data confidentiality.

For clear presentation, we assume each dimension has the same length of domain, h and each cloaking box is square with an edge-length e. Assume the whole domain also has a uniform distribution. According to the variance of uniform distribution, the NR_MSE measure is $\sqrt{6}e/(3h)$. To achieve the protection of 10 percent domain length, we have $e \approx 0{:}12h$.

In Fig. 7, the x-axis represents NR_MSE, i.e., the Casper's relative cloaking-edge length. It shows that when the edge length is increased from 2 to 10 percent, the precision dramatically drops from 62 to 13 percent for the 2D uniform data and 43 to 10 percent for the 2D NE data,

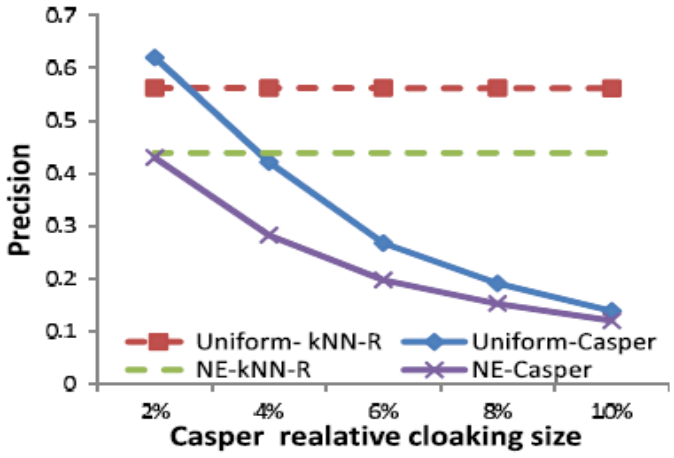

Fig 7: The impact of cloaking-box size on precision for Casper for the NE data

which shows the severe conflict between precision and confidentiality. The kNN-R's results are also shown for comparison

7 RELATED WORK

7.1 *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 = F(x) , for all xi, xj, xi < (>, =) xj ⇔yi < (>, =)yj. 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 bucketbased 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, Xi < ai might be replaced with X'i € [ID1, 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 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.

7.2 *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. [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 confidentiality, as well. The Casper approach [23] considers both data confidentiality and query privacy, the detail of which has been discussed in our experiments.

7.3 *Other 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], [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.

8 CONCLUSION

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 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 sublinear time complexity of processing queries. We then develop the kNN query service based on the range query service. The security of both the perturbeddata and the protected queries is carefully analyzed under aprecisely 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 Computing," technical report, Univ. of Berkerley, 2009.

[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.

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

[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.

[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 Constructions," Proc. 13th ACM Conf. Computer and Comm. Security, pp. 79-88, 2006.

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

[6] K. Chen, R. Kavuluru, and S. Guo, "RASP: Efficient Multidimensional Range Query on Attack-Resilient Encrypted Databases," Proc. ACM Conf. Data and Application Security and Privacy, pp. 249-260, 2011.

[31] P. Williams, R. Sion, and B. Carbunar, "Building Castles Out of Mud: Practical Access Pattern Privacy and Correctness on Untrusted Storage," Proc. ACM Conf. Computer and Comm. Security, 2008.