On Security of the Utility Preserving RASP Encryption

Keke Chen

Data Intensive Analysis and Computing Lab
Ohio Center of Excellence in Knowledge Enabled Computing
Department of Computer Science and Engineering
Wright State University, Dayton, OH 45435, USA
keke.chen@wright.edu

Abstract. Many potential users hesitate to use cloud computing because of the data confidentiality issue. Can we compute on untrusted public cloud platform with both data confidentiality and data utility preserved? Recent study has revealed that a convexity preserving encryption RASP can be used to construct confidentiality preserving and efficient range query service, which is one of the most frequently used query types for online data analytics. Convexity preserving encryption schemes, such as the RASP encryption, preserve the topology of the queried range in the encrypted space. It allows the encrypted data to be indexed and queried with transformed secure range queries. The initial study shows the range query service built on the RASP encrypted data can efficiently handle queries. However, there is no in-depth security analysis on the RASP encryption. In this paper, we focus on the security of the RASP encryption method. Concretely, we show that RASP is resilient to distributional attack, but it is not indistinguishable to chosen plaintext attack. We propose a relaxed security definition based on the statistical learning theory. We develop the Amount of Preserved Confidentiality (APC) measure to evaluate the security in terms of estimation attacks. We also show that the RASP encryption is resilient to estimation attacks and its encryption parameters can be appropriately tuned to meet different levels of confidentiality requirements.

1 Introduction

With the increasing popularity of web-based applications and the cloud infrastructures, service-based computing has become a major computing paradigm. Service providers take advantage of low cost cloud infrastructures, while service users enjoy convenient services without worrying about the cost of maintaining hardware and software. On the other hand, large datasets have been collected, stored, and analyzed in business intelligence and scientific computing for several years, which are expensive to maintain. An appealing solution is to outsource data-intensive services to the cloud platform or a service provider. However, data confidentiality on the untrusted platform is a major concern impeding the adoption of outsourced data services.

Encryption schemes for outsourced data services share a common characteristics - they need to preserve the utility of the provided data services, and thus often involve an intrinsic tradeoff between security and utility. Fully homomorphic encryption [11] represents an extreme case that preserves the lowest level of utility - the addition and multiplication operations, which, however, provides very poor performance for constructing

the upper level data-intensive services. As the author of [11] mentioned, this is still too expensive to be practical even for a simple application like encrypted keyword search. On the other side, many approaches developed in the database community focuses on performance, while providing weak security. For example, Crypto-index [15, 17] and order-preserving encryption (OPE) [1, 2] assume the attacker does not have sufficient prior knowledge about the data; thus powerful attacks are excluded from the consideration when the outsourced services are designed and deployed, which is unrealistic and dangerous.

Recently, the RASP encryption [6] is proposed for constructing secure range query service, one of the major database services. The RASP encryption approach introduces an interesting idea of preserving the convexity of datasets. More generally speaking, it tries to preserve the topology of convex sets with a secure transformation that is resilient to attacks. An important feature is to include random noise into the secure transformation, which, however, does not damage the preserved utility. The random noise component may force the attackers turn to *estimation attacks*. Experimental evaluation shows the RASP-based secure range query service can provide satisfactory performance.

Scope of the Paper. Although the original RASP paper includes a thorough discussion on possible attacks and the methods to enhance the attack resilience, it does not include a formal treatment. In this paper, we aim to formally analyze its security based on a precisely defined threat model. Concretely, there are three contributions.

1. The distributional attack is identified and analyzed. We show the shared theory behind the distributional attacks and show that the combination of the RASP encryption and dimensional order preserving encryption can render this type of attacks ineffective.
2. We argue that the traditional security definition of computational indistinguishability might be too strong for the situation of outsourced data services. A relaxed definition based on learning and statistical estimation theory is proposed, targeting on the estimation attacks. We also propose a measure *the Amount of Preserved Confidentiality* (APC) for evaluating security under the new definition.
3. Our analysis shows that the RASP encryption is not computationally indistinguishable under the chosen plaintext attack (CPA). Therefore, we use the relaxed definition to analyze the regression-based estimation attack that uses the chosen plaintext-ciphertext pairs. A theoretical lower bound of APC for this attack is derived based on the learning and estimation theory.

The rest of the paper is organized as follows. Section 2 describes the RASP encryption method and its application. Section 3 presents the threat model, the distributional attack, and the analysis on chosen plaintext attacks. Section 4 gives some of the related utility preserving encryption approaches.

2 Preliminaries

In this section, we give the formal definition of the RASP encryption method and briefly describe its benefits in constructing secure range query services.

2.1 Definition of RASP Encryption

Chen et al. proposed a construction of convexity preserving encryption, called the RASP scheme. We formally define the RASP construction as follows. Message m_i is a d-dimension vector, i.e., $m_i = (m_{i1}, \ldots, m_{id})$, where m_{ij} is a floating point number of length n bits. The basic RASP private-key encryption scheme is defined as follows.

- **Gen:** assume we have a pseudorandom invertible matrix generator $\mathcal{K}_m$. Choose a $(d+2) \times (d+2)$ random matrix A uniformly at random, where each element is of length n bits, and output it as the key.
- **Enc:** on the key A, and a message m_i, choose r, $r \in \mathbb{R}$, with a pseudorandom positive floating number generator $\mathcal{K}_r$, and output the ciphertext

$$c_i = A * (m_i^T, 1, r)^T, \tag{1}$$

  where the operation '*' is matrix multiplication and m_i^T means vector transpose. Thus, c_i is a $d+2$ dimension vector.
- **Dec:** on input a key A and a ciphertext c_i, output the plaintext message

$$m_i = P(A^{-1} * c_i, d), \tag{2}$$

  where the function $P(x, d)$ is a projection function that projects the vector x to its first d elements, and A^{-1} is the inverse of A.

The introduction of the noise component r is the key to increase the attack resilience of the encryption. In practice, the encryption construction [6] uses a combined scheme, RASP', which is a combination of two encryption schemes: order preserving encryption and the basic RASP scheme. Concretely, the OPE scheme is first applied to each dimension to change the dimensional distribution to normal distribution, which transforms the plain message matrix M to $\tilde{M}$. Then, the basic RASP scheme is applied to the records in $\tilde{M}$. It was shown that the RASP' scheme still preserves the utility for range query processing.

2.2 Preserved Utility for Secure Range Query

Range query is an important query in databases. The range is often a multidimensional range that describes a bounding box in the original space. Range query service will return records enclosed by the bounding box. Let X_i represent the i-th dimension. A dimensional range $[s_{i,min}, s_{i,max}]$ is represented by the condition "$X_i \leq s_{i,max}$ and $X_i \geq s_{i,min}$", which are linear half-space functions. Because the additional $d+2$ random dimension is always positive, the upper bound range condition, $X_i \leq s_{i,max}$, is equivalent to $(X_i - s_{i,max})X_{d+2} \leq 0$, while the lower bound range condition $X_i \geq s_{i,min}$ is equivalent to $(-X_i + s_{i,min})X_{d+2} \leq 0$. Thus, a d-dimensional range query that consists of a set of dimensional ranges $\{[s_{i,min}, s_{i,max}]\}$ is transformed to a series of quadratic conditions by using the positiveness of X_{d+2}. With the transforms defined in RASP, a condition $(X_i - s_{i,max})X_{d+2} \leq 0$ can be further transformed to a quadratic condition in the encrypted space: $y^T \Theta_i y \leq 0$. Paper [6] gives the detail of the matrix Θ_i, which is resilient to query-based attacks.

A two-stage query processing strategy is used to process the encrypted query. As Figure 1 shows, the original range is transformed to a polyhedron in the encrypted space. In the first stage, the server will find the points in the bounding box, MBR, of the polyhedron, which was calculated and sent by the proxy server based on the original range query. In the second stage, the server uses the conjunction of the quadratic conditions $y^T \Theta_i y \leq 0$ to filter out the points retrieved in the first stage. Experimental results have shown that this strategy works very efficiently [6].

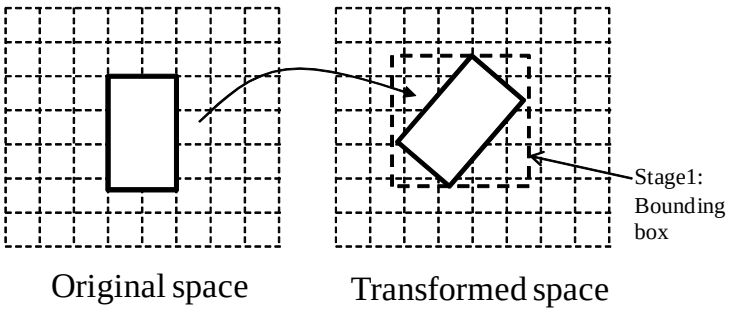

Fig. 1. Illustration of the two-stage range query processing algorithm.

3 Security Analysis of RASP

In this section, we will define a precise threat model for the outsourced services. Then, a unique attack - distributional attack is analyzed with a focus on the independent component analysis (ICA) attack. Third, we move on to the chosen plaintext attack and show that RASP is not IND-CPA. Finally, we develop a relaxed security definition to address the estimation attacks, and the RASP encryption is analyzed with the measure of the Amount of Preserved Confidentiality in terms of the regression-based estimation attack.

3.1 Threat Model

We consider several aspects of a threat model. First, we consider only the confidentiality of the outsourced data and range queries. Data in our research refers to multidimensional table-like data (columns by rows), which is most commonly used in scientific research and business analysis. Second, data disclosure attack is our major concern in this research - attackers might be interested in data distributions and/or the original data. Data tampering or dishonest service providers can be addressed by integrity preserving techniques [25, 23, 20], which will not be covered by the proposed research. In this context, we can assume the service provider is *honest-but-curious* [13].

Third, we model the attackers according to their prior knowledge. Active attackers will try to obtain as much knowledge as possible to help break the encryption or estimate the original data. We categorize the knowledge to three levels.

Level 1: the attacker observes only the encrypted data (and possibly queries), without any other additional knowledge, and might be interested in data distributions or the original data. This corresponds to the ciphertext-only attack (COA) in cryptanalysis.

Level 2: apart from the ciphertext, the attacker also knows distributions of plaintext, including the attribute domain (the maximum and minimum values), attribute distributions (e.g., the probability density function (PDF) or histogram), and the covariance between attributes. In practice, some applications may already expose distributional information through statistical database interface [8], but do not want to expose the exact data to the public, which may breach privacy.

Level 3: the attacker obtains a small set of plaintext tuples and their cipher tuples in the outsourced data. This corresponds to the chosen-plaintext attack (CPA). The CPA attack might be possible in some situations. For instance, the attacker may break in an authorized user's account in a query-based service, submit designed queries, and eavesdrop the communication channel to get plaintext-ciphertext pairs.

3.2 Distributional Attacks

Distributional Attack on OPE. The known distribution attack is fatal to some existing utility preserving encryption schemes, such as order preserving encryption. Let's take a look how this attack damages an OPE scheme for one dimensional data. The attacker can collect a sufficiently large number of cipher records and sort them in ascending order. Meanwhile, the original distribution is partitioned into bins. Each bin occupies certain percentage of the population. Let's denote the bins with $B = \{b_1, b_2, \ldots, b_m\}$ and the percentage of first i bins to the population as p_i and $p_0 = 0$. Because the order of the cipher records is preserved, we can find a mapping between the sorted cipher records and the bins B, bin by bin. In the i-th bin, we take the first p_i percentage of the sorted cipher records, denoted as the set S_i, $S_0 = \emptyset$, and assign the difference $S_i - S_{i-1}$ to the bin b_i. If b_i is bounded by the range $[a_i, a_{i+1}]$, we say the set $S_i - S_{i-1}$ has value approximately in $[a_i, a_{i+1}]$. We can choose arbitrary small bins to increase the precision of estimation. It was shown that RASP encryption does not preserve dimensional value orders [6]. Thus, the bin-based attack does not work on RASP.

ICA Distributional Attack. Independent Component Analysis (ICA) [18] is a fundamental problem in signal processing that has many applications such as blind source separation of mixed electro-encephalographic(EEG) signals, audio signals and the analysis of functional magnetic resonance imaging (fMRI) data. Let matrix X composed by source signals, where row vectors represent source signals. Suppose we can observe the mixed signals Y, which is generated by linear transformation $Y = AX$. The ICA model is designed to separate the independent components (the row vectors) of the original signals X from the mixed signals Y, if the following conditions are satisfied:

1. The source signals are independent, i.e., the row vectors of X are independent;
2. All source signals must be non-Gaussian with possible exception of one signal;

3. The number of observed signals, i.e. the number of row vectors of Y, must be at least as large as the independent source signals.
4. The transformation matrix A must be of full column rank.

All ICA algorithms [18, 16] share the same idea that tries to find a matrix $\tilde{A}$ that $\tilde{A}Y$ contains non-Gaussian components to approximate the signals in X.

RASP's Resilience to ICA Attack. RASP's protection from the ICA attack is to use the combination of OPE and RASP, i.e., the RASP' scheme. Because each of the d original dimensions is transformed to normal distribution via the OPE scheme and the additional noise $d+2$-th dimension has an approximate normal distribution, the condition 2 of effective ICA is not satisfied. The ICA algorithms can possibly identify the constant $d+1$-th dimension '1', but they cannot distinguish the other $d+1$ dimensions.

3.3 RASP Encryption is not IND-CPA

We present a chosen plaintext attack, *Plane Attack*, on the RASP encryption, which demonstrates that the RASP encryption is not indistinguishable to chosen plaintext attack.

Proposition 1 *RASP is not computationally indistinguishable to chosen plaintext attack.*

Sketch of Proof. Let $c = E(m)$ represent a RASP encryption. The distinguisher experiment is described as follows.

1. m_0 and m_1 are two points randomly sampled from the original data space $\mathbb{R}^d$.
2. The point m_b, where $b \in \{1, 0\}$ is randomly selected, is encrypted to c_b with the function E and given to the adversary.
3. The adversary can request a polynomial number of plaintext records $\{m_i, i > 1, m_i \in \mathbb{R}^d\}$ to be encrypted, where $m_i \neq m_0$ and $m_i \neq m_1$. With some attacking algorithm, the adversary finally outputs a bit b'.

If $|Pr(b' = b) - Pr(b' \neq b)| < 1/p(n)$, $p(n)$ is some polynomial function in terms of the key length n, we say the encryption scheme is indistinguishable under chosen plaintext attack. We show that the following "plane attack" allows the adversary to accurately predict b, i.e., $Pr(b' = b) = 1$. Thus, RASP is not IND-CPA.

If $m_0 \neq m_1$, there is always a plane separating these two points. Let's use $w^T m + a = 0$ to represent the plane, where m is a variable in the original data space $\mathbb{R}^d$, while w and a are parameters determining the plane. Therefore, we have $w^T m_0 + a < 0$ and $w^T m_1 + a > 0$. Let $c_i = E(m_i)$ be the ciphertext. By definition of RASP encryption, the separation plane $w^T m + a = 0$ is transformed to another plane in the encrypted space, $u^T A^{-1} y = 0$, where $u^T = (w^T, -a, 0)$. By definition, we still have $u^T A^{-1} c_0 < 0$, and $u^T A^{-1} c_1 > 0$. Therefore, once we can determine the plane parameter $u^T A^{-1}$, we can test c_b to distinguish whether $b = 0$ or $b = 1$.

We show that although it is impossible to find the exact value of $v^T = u^T A^{-1}$, we can possibly find $v^T = \alpha t^T$ that t is known but the scalar α is unknown. Plugging αt^T to the separation plane, the decision problem is transformed to determining the sign of

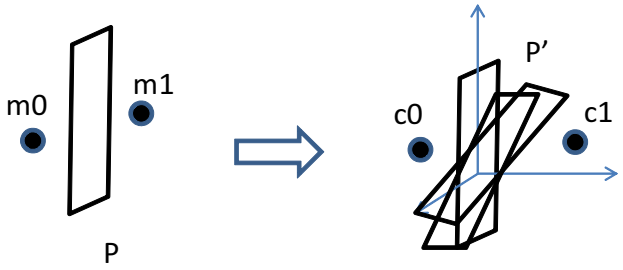

Fig. 2. Many planes passing the original can be the candidate of the original plane P's image. However, all of them can correctly separate the ciphertext c_0 and c_1.

$\alpha t^T c_b$. Thus, knowing the sign of α will solve the decision problem completely. Next, we describe the methods of (1) deriving the vector t and (2) determining the sign of α.

We sample a sufficient number of points $\{m_i\}$, where $i > 1$ and $m_i \neq m_0$ and $m_i \neq m_1$, from the original separation plane $w^T m + a = 0$. With extended $d+2$-dimensional vectors $\tilde{m}_i$ as the column vectors, we get a matrix $\tilde{M}$. They are then encrypted with the RASP encryption to $\{c_i\}$. c_i should be in the plane $v^T c = 0$ according to the definition of v^T. Using the vectors $\{c_i\}$ as the column vectors we get a matrix $\tilde{C}$ so that $v^T \tilde{C} = 0$. It is clear that there are an unlimited number of solutions for v^T. To find the representation of v^T, we need to find the basis of the null space of $\tilde{C}$. Because $\tilde{C} = A\tilde{M}$ and Rank($\tilde{C}$) = Rank($A\tilde{M}$), it follows Rank($\tilde{C} \leq$ min {Rank (A), Rank ($\tilde{M}$) }, according to the property of matrix rank [21]. Because Rank(A)=d+2 and $\tilde{M}$ has one row of 1, Rank($\tilde{C}$) $\leq$ Rank($\tilde{M}$) $\leq d + 1$. We can easily find sample points to make Rank($\tilde{C}$) $= d + 1$. According to the rank-nullity theorem [21], the basis of the null space of $\tilde{C}$ consists of one vector t and the solution to $v^T \tilde{C} = 0$ can be represented as $v^T = \alpha t^T$, where α can be any value. However, the sign of α can be determined with an additional known pair of (m_k, c_k) that is not on the separation plane. Without loss of generality, we assume $w^T m_k + a < 0$, which leads to $\alpha t^T c_k < 0$. With known t^T and c_k we derive the sign of α.

Now with the known vector t^T and the sign of α the adversary can easily determine the sign of $\alpha t^T c_b$. If $\alpha t^T c_b < 0$, $b = 0$ is returned, otherwise $b = 1$. This gives the exact answer to the distinguisher experiment. □

.

Figure 2 illustrates the Plane attack. One may wonder whether the adversary can use the plane attack to figure out the key matrix A - i.e., using different planes and samples on the planes to construct a sufficient number of equations to infer A. We show that it is impossible. Based on the previous discussion, the best we can get for a plane i is $u_i^T A^{-1} = \alpha_i t_i^T$. Putting $d + 2$ planes (with linearly independent u_i) together we can reduce the number of unknowns of A to $d + 2$, i.e., $A^{-1} = U^{-1}(\alpha_1 t_1, \ldots, \alpha_{d+2} t_2)$, where $U = (u_1, \ldots, u_{d+2})^T$. However, the unknowns α_i cannot be further eliminated.

3.4 Refined Security Definition

We think the indistinguishability definition might be too strong for the situation of encrypted data for outsourced services. Distinguishing which record is encrypted is not very meaningful to the attacker. The attacker is more interested in estimating the original values based on his/her prior knowledge. In the following, we investigate a relaxed definition based on the effectiveness of estimation attacks.

Amount of Preserved Confidentiality. The existing security definition for symmetric encryption is based on the *computationally indistinguishability* model [19]. This definition says that the adversary can only do *negligibly* better than random guess to find the plaintext for a ciphertext. Because of the preserved properties, the strong indistinguishability definition may not be achieved - the recent study on order-preserving encryption [2, 3] has shown this. In the context of outsourced data, the attacker's goal may not be finding the exact plaintext for a ciphertext (which becomes impossible when random noise is available) or distinguishing ciphertexts. We believe more interesting is the precise estimation of each record based on the known information.

We propose a new security definition based on estimation accuracy. According to different applications, the tolerance of the level of adversarial estimation error could vary. Thus, *the accuracy of estimation* needs to be tunable according to users's preferences. Based on this understanding, attacks can be modeled as a learning problem - with the known information the attacker wants to *learn* an approximate decryption function that outputs *approximate plaintext* for the corresponding ciphertext. Correspondingly, the theoretical lower error bound [22] of the learning problem defines the security for a specific dataset and a specific encryption method.

Specifically, we define the *amount of preserved confidentiality* (APC) based on the distribution of estimation error. Let X represent the random variable that generates the values for certain attribute in the dataset. Let $\hat{X}$ be the estimate of X. $\{x_i\}$ are the plaintexts and $\{\hat{x}_i\}$ are the corresponding estimated plaintexts. In general, if the estimation error $X - \hat{X}$ approximately follows some distribution with mean μ and standard deviation σ, we can derive the normalized error $(X - \hat{X} - \mu)/\sigma$ in the range $[-\delta, \delta]$ with certain confidence 1-α (e.g., 0.95).

$$Pr(-\delta \leq \frac{X - \hat{X} - \mu}{\sigma} \leq \delta) \geq 1 - \alpha, \tag{3}$$

which is justified by the interval estimation theory [5]. It follows that the error $X - \hat{X}$ is bounded by $[\mu - \delta\sigma, \mu + \delta\sigma]$. μ and σ together determine the interval. With the known distributional information, however, μ can be easily estimated and removed from $\hat{X}$. Thus, only σ is meaningful to confidentiality. To precisely evaluate the confidentiality crossing different domains, we refine the definition according to the length of domain. The length of the domain is defined as the length of the portion of the domain that contains the majority of elements. For example, if the domain has a normal distribution, the majority (95.4%) of points is in the range $[\mu' - 2\sigma', \mu' + 2\sigma']$, where μ' is the mean and σ' is the standard deviation. In this case, we can derive the length of domain is $4\sigma'$. Let D denote the length of domain. We define the amount of preserved confidentiality as the relative length of the interval: $APC = \frac{2\delta\sigma}{D}$ under confidence level $1 - \alpha$.

We can derive the APCs under the indistinguishability definition, which are instructional for users setting a specific APC threshold. If X is uniformly distributed in the domain $[0, D]$, $D > 0$, indistinguishability tells $\hat{X}$ has the same distribution. It follows that $X - \hat{X}$ is a triangle distribution [10] in $[-D, D]$ with $\mu = 0$ and $\sigma = D/\sqrt{6}$. For normalized $X - \hat{X}$ distribution (triangle distribution in $[-\sqrt{6}, \sqrt{6}]$), we choose $\delta = \sqrt{6}$ to include the complete domain (thus with confidence level 1). It follows that $APC = \frac{2\delta\sigma}{D} = 2$ with confidence level 1. Similarly, we can derive $APC = 2$ with confidence level 0.954 if we assume the encryption is indistinguishable under normal plaintext distribution.

More interesting is the theoretical lower bound of APC for a specific dataset and an encryption construction. Since the APC measure is determined by the variance of the error (σ^2), the key question to security analysis is whether we can determine the minimum variance, regardless of any learner, for a specific plain dataset and a specific encryption method. This problem is nicely linked to the existing results in statistics and machine learning. We will analyze the lower bound of APC for the RASP method in terms of CPA attacks.

3.5 Security Analysis on Estimation Attacks

We apply the APC measure to evaluate the effectiveness of estimation attack with the chosen plaintext records. First, we will define the estimation attack; then we derive the lower bound of APC for the estimation attack.

Assume the attacker knows a number of plaintext/ciphertext record pairs, which are chosen by the attacker. Concretely, let $P_{d\times m}$ be the known m d-dimensional original records $(x_1, \ldots, x_m), m > d + 2$ and $x_i \in \mathbb{R}^d$, that include $d + 2$ linearly independent records, and $Q_{d+2\times m}$ be the corresponding $d + 2$-dimensional ciphertext records $(y_1, \ldots, y_m)$, $y_i \in \mathbb{R}^{d+2}$.

We first transform the RASP equation for easier manipulation. Let the key matrix A decomposed into blocks $A = (A_1, A_2, A_3)$, where A_1, A_2 and A_3 have block sizes $(k + 2) \times k$, $(k + 2) \times 1$ and $(k + 2) \times 1$, respectively. Let C and M are the ciphertext and plaintext datasets, respectively. The extended data is $\begin{pmatrix} M \\ \mathbf{1} \\ v \end{pmatrix}$ where $\mathbf{1}$ is the row vector with '1' and v is a row vector with random positive values. According to the RASP definition, the relationship between C and M is

$$C = (A_1, A_2, A_3) \begin{pmatrix} M \\ \mathbf{1} \\ \mathbf{v} \end{pmatrix} = A_1 M + A_2 \mathbf{1} + A_3 \mathbf{v}. \tag{4}$$

$A_2\mathbf{1}$ can be treated a translation matrix that adds the constant vector A_2 to each of the column vectors in $A_1 X$; $A_3\mathbf{v}$ is a random noise matrix.

At the first look, Eq. 4 is a standard affine transformation with a noise component. Assume the noise vector v consists of independent and identical random variables with mean value μ_v and variance σ_v^2. We have $v = \mu_v + \tilde{v}$, where $\tilde{v}$ has mean value zero and the same variance σ_v^2. Thus, the noise component is decomposed to $A_3\mu_v + A_3\tilde{v}$.

As the constant component $A_2\mathbf{1} + A_3\mu_v$ can be canceled by subtracting any known plaintext/ciphertext record pair (x_0, y_0) from the pairs (x_i, y_i), the problem is reduced to estimate A_1 with the presence of the noise component $A_3\tilde{v}$ and the known plaintext records P and the corresponding ciphertext records Q. The standard method for learning A_1 is regression analysis [9]. In fact, the Gauss-Markov theorem [16] tells the least square regression method gives the minimum variance unbiased estimator, which helps us identify the lower bound of APC.

Because eventually we want to use the whole set of ciphertext C to predict the plaintext M, for easier manipulation, we further transform the equation to the canonical regression problem that have the "responses" M on the left side of the equation

$$M = (A_1^T A_1)^{-1} A_1^T C - (A_1^T A_1)^{-1} A_1^T A_3 \tilde{v}, \tag{5}$$

assuming A is selected so that $A_1^T A_1$ is invertible. Let's consider one dimension only. Let M_i be the i-th row of M, and β be the i-th row of $(A_1^T A_1)^{-1} A_1^T$ and U_i be i-th row of $-(A_1^T A_1)^{-1} A_1^T A_3 \tilde{v}$. The equation is simplified to

$$M_i = \beta C + U_i. \tag{6}$$

We further simplify this representation to focus on the dimensional value. Considering the dimensional plaintext value x_{ij} of the j-th dimension for the plaintext record x_i, from Eq. 6 we have $x_{ij} = \beta y_i + u_j$, where y_i is x_i's ciphertext and u_j is some random value drawn from the distribution of U_i. In practice, the attacker can only use $\hat{x}_{ij} = \hat{\beta} y_i$ to estimate x_{ij}, where $\hat{\beta}$ is the estimation of β learned with regression analysis and the known plaintext P and ciphertext Q. According to [16], the expected squared prediction error of the estimator $\hat{x}_{ij} = \hat{\beta} y_i$ is

$$\begin{aligned} E(x_{ij} - \hat{\beta} y_i)^2 &= var(u_j) + MSE(\hat{\beta} y_i) \\ &= var(u_j) + E^2(x_{ij} - \hat{\beta} y_i) + var(\hat{\beta} y_i). \end{aligned} \tag{7}$$

Because $var(x_{ij} - \hat{\beta} y_i) = E(x_{ij} - \hat{\beta} y_i)^2 - E^2(x_{ij} - \hat{\beta} y_i)$, it follows $var(x_{ij} - \hat{x}_{ij}) = var(u_j) + var(\hat{\beta} y_i)$. Regardless of the variance of the estimator $var(\hat{\beta} y_i)$, the noise component $var(u_j)$ is irreducible [16], for which we can conclude that the $var(u_j)$ gives a very conservative lower bound for $var(x_{ij} - \hat{x}_{ij})$.

We further analyze the closed form of $var(u_j)$ to better understand the lower bound. Let c_i be the i-th element of the vector $(A_1^T A_1)^{-1} A_1^T A_3$. It follows that $var(u_j) = c_i^2 var(\tilde{v}) = c_i^2 \sigma_v^2$. Therefore, the variance $var(u)$ is co-determined by the key matrix A and the variance of the original noise v.

Based on this result, we can derive the lower bound of APC for RASP encryption. Assume the prediction error $x_{ij} - \hat{\beta} y$ follows a normal distribution, we choose the span $\delta = 2$ to cover more than 95% of the error population. Let the domain length of the dimension i be D_i. We consider the lower bound of APC for RASP encryption for dimension i

$$\min(APC\_RASP(i)) = 2\delta \sqrt{var(u)} / D_i = 4 c_i \sigma_v / D_i. \tag{8}$$

To remove the effect of different domain lengths, we can standardize all the dimensions before performing the encryption. With the help of OPE, we can tune the domain distributions to make all domains have normal distributions. Thus, all D_i can be set to 4 to cover more than 95% of the normalized values. With this preparation, it is easy to derive encryption settings that satisfy a user-defined lower bound APC_0 for all dimensions - we can control the generation of A and σ_v so that $\min\{c_i\sigma_v, i = 1 \dots d\} > APC_0$.

4 Related Work

The order preserving encryption (OPE) [1] preserves the dimensional value order after encryption. Thus, it can be used in most database operations, such as indexing and range query. Boldyreva et al. [2, 3] has formally analyzed the security of OPE. As we demonstrated in the paper, any OPE schemes are vulnerable to distributional attacks. Crypto-Index is 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. However, the bucketization scheme leaks a lot of information. Thus, a bucket-diffusion scheme [17] was proposed to introduce noise records into the results to improve the security, which, however, has to sacrifice the precision of query results. Secure keyword search on encrypted documents [24, 14, 12, 4, 7] is another cluster of utility preserving encryption methods. They allow the server to scan each encrypted document in the database and find the documents containing the keyword. There have been rigid security analysis on this line of research [12, 7].

5 Conclusion and Future Work

In this paper we thoroughly analyzed the data security of the utility preserving RASP encryption. The result shows that the RASP encryption does not satisfy IND-CPA. Thus, we introduce a relaxed definition, which is based on learning and estimation theory. A concrete estimation attack - the regression attack utilizing the known plaintext-ciphertext pairs to estimate the unknown plaintexts - is analyzed under the new security definition. The result show that the RASP encryption parameters can be possibly selected to meet different level of requirements on data confidentiality. Due to the space limitation, we have to focus on our analysis on data confidentiality. Another interesting problem is the security of the RASP's query transformation algorithm. Although the initial analysis shows its resilience to various attacks, a formal analysis may reveal possible weaknesses or uncover nice security properties.

References

1. R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Order preserving encryption for numeric data," in *Proceedings of ACM SIGMOD Conference*, 2004.
2. A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill, "Order preserving symmetric encryption," in *Proceedings of EUROCRYPT conference*, 2009.
3. A. Boldyreva, N. Chenette, and A. ONeill, "Order-preserving encryption revisited: Improved security analysis and alternative solutions," in *CRYPTO*, 2011.

4. D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public-key encryption with keyword search," in *Proceedings of Advances in Cryptology, (EUROCRYPT)*. Springer, 2004.
5. G. Casella and R. L. Berger, *Statistical Inference*. Duxbury Press, 2001.
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.
7. R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in *Proceedings of the 13th ACM conference on Computer and communications security*. New York, NY, USA: ACM, 2006, pp. 79–88.
8. J. Domingo-Ferrer, *Inferance Control in Statistical Databases*. Springer, 2002.
9. N. R. Draper and H. Smith, *Applied Regression Analysis*. Wiley, 1998.
10. M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*. Wiley, 2000.
11. C. Gentry, "Fully homomorphic encryption using ideal lattices," in *STOC '09: Proceedings of the 41st annual ACM symposium on Theory of computing*. New York, NY, USA: ACM, 2009, pp. 169–178.
12. E.-J. Goh, "Secure indexes," Cryptology ePrint Archive, Report 2003/216, 2003.
13. O. Goldreich, *Foundations of Cryptography*. Cambridge University Press, 2001.
14. P. Golle, J. Staddon, and B. Waters, "Secure conjunctive keyword search over encrypted data," in *ACNS 04: 2nd International Conference on Applied Cryptography and Network Security*. Springer-Verlag, 2004, pp. 31–45.
15. H. Hacigumus, B. Iyer, C. Li, and S. Mehrotra, "Executing sql over encrypted data in the database-service-provider model," in *Proceedings of ACM SIGMOD Conference*, 2002.
16. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer-Verlag, 2001.
17. B. Hore, S. Mehrotra, and G. Tsudik, "A privacy-preserving index for range queries," in *Proceedings of Very Large Databases Conference (VLDB)*, 2004.
18. A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.
19. J. Katz and Y. Lindell, *Introduction to Modern Cryptography*. Chapman and Hall/CRC, 2007.
20. F. Li, M. Hadjieleftheriou, G. Kollios, and L. Reyzin, "Dynamic authenticated index structures for outsourced databases," in *Proceedings of ACM SIGMOD Conference*, 2006.
21. C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, 2000.
22. T. Mitchell, *Machine Learning*. McGraw Hill, 1997.
23. R. Sion, "Query execution assurance for outsourced databases," in *Proceedings of Very Large Databases Conference (VLDB)*, 2005.
24. D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *IEEE Symposium on Security and Privacy*. Washington, DC, USA: IEEE Computer Society, 2000, p. 44.
25. M. Xie, H. Wang, J. Yin, and X. Meng, "Integrity auditing of outsourced data," in *VLDB*, 2007, pp. 782–793.