Multi-media Security Final Project
An Attack of MRSE

網媒二 r01944007 簡伯宇
資工三 b00902064 宋昊恩

Abstract

MRSE is first mentioned in [Privacy-preserving multi-keyword ranked search over encrypted cloud data][1], and the same team kept using this method in their newest paper [Verifiable Privacy-Preserving Multi-keyword Text Search in the Cloud Supporting Similarity-based Ranking][2] cited in 2013.

Introduction

We purposed two attacks to MRSE method cited in the above-said two papers. One attack is for the intuitive method, and the other one is for the MRSE itself. We will show that MRSE shares the same level of security as the original, intuitive method. And thus, we prove both of them are not safe.

Attack of intuitive method

I. Problem Formalization

The intuitive method uses k-dimention vectors to present the keyword space for documents and queries.

To be specific, client will hold a $k \times k$ secret encryption matrix R. Furthermore, there are m plain documents $F = (F_1, F_2, \dots, F_m)$, which will be encrypted into $\tilde{F} = (\tilde{F}_1, \tilde{F}_2, \dots, \tilde{F}_m)$, and n plain queries $Q = (\mathrm{Q}_1, Q_2, \dots, Q_n)$, which will be encrypted into $\tilde{Q} = (\tilde{Q}_1, \tilde{Q}_2, \dots, \tilde{Q}_n)$.

II. Key Point of Attack

[1] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," in Proc. IEEE Conf. Comput. Commun. (INFOCOM), Apr. 2011.
[2] S. W, B. Wang, N. Cao, M. Li, W. Lou, Y. Hou, H. Li, "Verifiable Privacy-Preserving Multi-keyword Text Search in the Cloud Supporting Similarity-based Ranking", in IEEE Trans, Nov.2013.

If we have many encrypted documents, for example, $k+1$ of them, then we can perform attacks upon them.

Let $\tilde{F}_M^{k\times(k+1)} = [\tilde{F}_1, \tilde{F}_2, \dots, \tilde{F}_k, \tilde{F}_{k+1}]$. By calculating the rank of this matrix, it is obvious that some of the encrypted documents will have dependency. We just call them $\tilde{F}_{dep} = [\tilde{F}_1, \dots, \tilde{F}_p]$. And, since the secret matrix R is invertible, $F_{dep} = (R^T)^{-1}\tilde{F}_{dep}$ will have the same dependency. That is to say, this intuitive method is not secure.

Attack of MRSE_I

I. Problem Formalization

MRSE_I divides k-dimention vectors into $2k$-dimention vectors in order to destroy linearity of encryption. However, we give an attack to show that we can recover this linearity of MRSE_I, once we've collected enough data.

Formally, assume the MRSE_I has a 3-tuple key $\{S, M, N\}$. It encrypts m documents $F = \{f_1, f_2, \dots f_m\}$ into $\tilde{F} = \{\tilde{f}_1, f_2, \dots \tilde{f}_{\mathrm{m}}\}$ and n queries $\mathrm{Q} = \{q_1, q_2, \dots q_n\}$ into $\tilde{Q} = \{\tilde{q}_1, \tilde{q}_2, \dots \tilde{q}_{\mathrm{m}}\}$.

Our attack will take $\widetilde{D}$ and $\tilde{Q}$ as input, and output a pair of $k\times 2k$ matrices A, B such that there exists a pseudo key R satisfied:

$$A\tilde{f} = R^T f$$
$$B\tilde{q} = R^{-1} q$$

Where f and q are any possible document and query, $\tilde{f}$ and $\tilde{q}$ are corresponding encrypted form. That is, after multiplying a matrix (A or B) on encrypted document or query, the resulting k-dimention vector is same with client's important data encrypted by the intuitive method, which is shown unsafe as mentioned, with key R.

II. Key Point of Attack

The first and key defect of MRSE is that, whether the bit of S is $0/1$ will change the dimension of encrypted data differently.

Let's look at MRSE_I encryption closer. First, it divides a document f_i into a $2k$-dimension vector $[f_i'; f_i'']$ based on S, and then it multiplies a matrix on it:

$$\tilde{f} = \begin{bmatrix} \widetilde{f_i'} \\ \widetilde{f_i''} \end{bmatrix} = \mathcal{M}\begin{bmatrix} f_i' \\ f_i'' \end{bmatrix}, \mathcal{M} = \begin{bmatrix} M^T & 0 \\ 0 & N^T \end{bmatrix}$$

Let $\wp$ be the vector space of all possible $[f_i'; f_i'']$, and $\widetilde{\wp}$ be the vector space of all possible $\tilde{f}$. If j-th bit of S is 0, then $f_i'[j] = f_i''[j] = f_{\mathrm{i}}[j]$, which only contributes one to the dimension of $\wp$; On the other hand, If j-th bit of S is 0, then $f_i'[j]$, $f_i''[j]$ are two random numbers satisfied $f_i'[j] + f_i''[j] = f_{\mathrm{i}}[j]$, which contributes two to the dimension of $\wp$.

Since $\mathcal{M}$ is invertible, the dimension of $\widetilde{\wp}$ must equal to the dimension of $\wp$. Therefore, once the server contains encrypted documents $\tilde{F}$ that are enrich enough to span $\widetilde{\wp}$, then we can tell the number of bit 1 in S.

Similar work can also apply on queries instead of documents. Let $\mathbb{Q}$ and $\widetilde{\mathbb{Q}}$ are the vector spaces of all possible $[q_i'; q_i'']$ and $\tilde{q}$. Once the server collects queries $\tilde{Q}$ that are enrich enough, we can reconstruct $\widetilde{\mathbb{Q}}$.

Theorem 1:

$dim(\widetilde{\wp}) = k + \ell_0$, $dim(\widetilde{\mathbb{Q}}) = k + \ell_1$, where ℓ_0, ℓ_1 are the number of bit 0, 1 in S.

And that's all we can learn about S. Since the client can swap two entries without changing encrypted data, we can't learn about which bit of S is 0 or 1. Look on the bright side, instead, the ability of swap two entries also allow us to assume which bit of S is 0 or 1 freely, so let's assume $S[1, \dots \ell_0] = 0, S[\ell_0 + 1, \dots k] = 1$.

III. Subspace Dividing

Now, we have to decompose M, N into vectors. Let $M^T = [m_1, m_2, \dots m_k]$, $M^{-1} = [\widetilde{m}_1, \widetilde{m}_2, \dots \widetilde{m}_k]$, $N^T = [n_1, n_2, \dots n_k]$ and $N^{-1} = [\tilde{n}_1, \tilde{n}_2, \dots \tilde{n}_k]$. We have $m_i^T \widetilde{m_j} = n_i^T \tilde{n}_j = \delta_{ij}$, where δ_{ij} equals 0 if i ≠ j and 1 otherwise.

Let v be a k-dimension vector,$[v;\ 0]$ denote the $2k$-dimension vector whose first k entries are v, and the last k entries are 0; $[0;\ v]$ denote the $2k$-dimension vector whose first k entries are 0, and the last k entries are v.

Despite to the restriction of the space limit, we don't provide proofs to the theorems in remaining part of these projects. If the reader has any problem of the proof, it's please to contact the authors.

Theorem 2:

The vectors $[m_i;\ 0], 1 \le i \le \ell_0$ and $[0;\ n_i], 1 \le i \le \ell_0$ and $[m_i;\ n_i], \ell_0 < i \le k$ is a basis of $\widetilde{\wp}$.

Theorem 3:

Let vector space $U = \{[x; 0] | x \in \mathbb{R}^k\}$, then $[m_i;\ 0], 1 \le i \le \ell_0$ span $\widetilde{\wp} \cap U$. Similarly, Let vector space $D = \{[0; x] | x \in \mathbb{R}^k\}$, then $[0;\ n_i], 1 \le i \le \ell_0$ span $\widetilde{\wp} \cap D$.

Theorem 4:

There exists a basis of keywords such that the vectors $[m_i;\ n_i], \ell_0 < i \le k$ are orthogonal to subspaces $\widetilde{\wp} \cap U$ and $\widetilde{\wp} \cap D$. That is, $[m_i;\ n_i], \ell_0 < i \le k$ span $\widetilde{\wp} \cap (\widetilde{\wp} \cap U)^{\perp} \cap (\widetilde{\wp} \cap D)^{\perp}$.

Given $\widetilde{\wp}$, we can compute its three subspaces $\widetilde{\wp}_U = \widetilde{\wp} \cap U, \widetilde{\wp}_D = \widetilde{\wp} \cap D$, and $\widetilde{\wp}_M = \widetilde{\wp} \cap (\widetilde{\wp} \cap U)^{\perp} \cap (\widetilde{\wp} \cap D)^{\perp}$. It gives us some information of m_i, n_i. However, we cannot determine m_i, n_i yet.

IV. Compute M, N under a Keyword Basis

Recall that our goal is to merge a $2k$-dimension vector back to a k-dimension vector. Hence, we have to recognize corresponding pairs of entries, i.e., recognize corresponding n_i to each m_i.

However, we cannot do it by $\widetilde{\wp}$ only. At the time, $\widetilde{\mathbb{Q}}$ must be introduced. The following theorem show that the basis of $\widetilde{\wp}_M$ and $\widetilde{\mathbb{Q}}_M$ can be determined arbitrary.

Theorem 5:

Let $\{s_1, \dots, s_{\ell_1}\}$ is an arbitrary basis of $\widetilde{\wp}_M$ and $\{t_1, \dots, t_{\ell_0}\}$ is an arbitrary basis of $\widetilde{\mathbb{Q}}_M$, Then there exist a basis of keywords such that:

a. $[m_i;\ n_i], \ell_0 < i \leq k$ are orthogonal to subspaces $\widetilde{\wp} \cap U$ and $\widetilde{\wp} \cap D$
b. $[m_i;\ n_i] = s_{i-\ell_0}, \ell_0 < i \leq k$
c. $[\widetilde{m_\iota};\ \widetilde{n_\iota}] = t_i, 1 \leq \mathrm{i} \leq \ell_0$

Therefore, we can generate arbitrary basis of $\widetilde{\wp}_M$ and $\widetilde{\mathbb{Q}}_M$, and treat it as $[m_i;\ n_i]$, $\ell_0 < i \leq k$ and $[\widetilde{m_\iota};\ \widetilde{n_\iota}] = t_i, 1 \leq \mathrm{i} \leq \ell_0$. For these vectors, m_i and n_i are mapped correctly. The other vector pairs can be determined by following theorem:

Theorem 6:

Let P be an invertible $k \times k$ matrix, $P^T = [p_1, p_2, \dots p_k]$, $P^{-1} = [\tilde{p}_1, \tilde{p}_2, \dots \tilde{p}_k]$, and there's a ℓ such that $p_i^T p_j = 0, 1 \leq i \leq \ell < j \leq k$. If only $\tilde{p}_i, 1 \leq \mathrm{i} \leq \ell$ and $p_i, \ell < i \leq k$ are known, we can still using these to determine P.

After theorem 5, we determine the basis of keywords in order to fit our basis of $\widetilde{\wp}_M$ and $\widetilde{\mathbb{Q}}_M$, also obtain some pairs of m_i and n_i. Using theorem 6, we can find out the other mapping of m_i and n_i. Therefore, under these basis of keywords, we can finally compute the keys M and N.

V. Obtain A, B

Now we have 3-tuple key (S, M, N) under some keyword basis. We can

$$[(M^{-1})^T \quad (N^{-1})^T]\widetilde{f_i} = (M^{-1})^T \widetilde{f_i'} + (N^{-1})^T \widetilde{f_i''} = f_i' + f_i''$$

Denote $\mathrm{A}' = [(M^{-1})^T \quad (N^{-1})^T]$, the previous equation shows that $A'\tilde{f}_\mathrm{i}$ is close

to f_i.

Hence we can modify A' as

$$A[i][j] = \begin{cases} A'[i][j], if\ 1 \le i \le \ell_0 \\ A'[i][j]/2, if\ \ell_0 < i \le k \end{cases}, 1 \le j \le 2k$$

Then $A'\tilde{f}_{\mathrm{i}} = f_i$ is what we want. The matrix B can be computed in a similar way.

Conclusion and discussion

I. Security of MRSE_I

During the previous section, the only restriction of attack is that we have to collect enough documents and queries that span $\widetilde{\wp}$ and $\widetilde{\mathbb{Q}}$ respectively. The dimension of $\widetilde{\wp}$ and $\widetilde{\mathbb{Q}}$ is no more than $2k$, hence if the content of documents and queries are random enough, the server can attack the data when the amount of documents and queries are more than $2k$.

II. Security of MRSE_II

In the same paper it also proposes MRSE_II encryption which adds some random entries. But it doesn't help to increase security from MRSE_I. The random entries are just actually low-frequency pseudo keywords which may either appear in all documents once or appear in all queries once. Under the assumption of randomness of documents of queries, if the client adds w extra pseudo keywords on k keywords, it will let the server needs to collect only $2w$ more documents and queries to attack. However, it will also let the query time become $O((k+w)^2)$ from $O(k^2)$. It shows that adding pseudo words may not be worth.

III. Repair of MRSE_I

The key defect of MRSE_I is that the dimension of documents space reveals some information of S. The (possible) repair is that we can eliminate the relationship. For example, let the random number $f_i'[j]$, $f_i''[j]$ reserves some dependency, or redesign the encrypt method when bit of S is 0 instead of $f_i'[j] = f_i''[j] = f_{\mathrm{i}}[j]$.