Privacy Preserving Search on Multimedia
Part IV

Prof. Ja-Ling Wu
(Cited from Prof. Min Wu’s work)
V. SECURITY ANALYSIS

One of the key features of modern information retrieval is the separation of data from search indexes, which enables efficient search capability and convenient data management. Similarly, in the proposed secure retrieval schemes, image data themselves are encrypted separately from search indexes by cryptographic ciphers before being stored on the server. Built on top of established cryptographic primitives, it is computationally difficult for an adversary to infer image content by breaking the cryptographic ciphers.
Therefore, the security of the entire retrieval schemes is determined by the security of the encryption schemes for features or indexes. In this section, we focus our analysis on the security of the proposed feature/index encryption schemes and potential information leakage under two attack models, namely, ciphertext only attack (COA) and known-plaintext attack (KPA).
We assume that the adversary is semi-honest, i.e., the adversary will follow the execution requirement of the protocol but may use what they see during the execution to infer additional information. Such a semi-honest model is applicable to such scenarios as third-party service providers. We provide security definition and proofs for the proposed secure retrieval schemes under the COA model and compare their security levels in the KPA model.
A. Ciphertext Only Attack (COA)

Ciphertext only attack model assumes an adversary has access only to the ciphertext. Since the goal of confidentiality preserving multimedia retrieval is to protect content privacy against adversaries, such as malicious intruders and untrustworthy service providers who have access to encrypted images and indexes, it is necessary for any confidentiality preserving retrieval scheme to be secure under the COA model. We first discuss the potential information leakage and define the security objective in the COA model.
Suppose $M$ images $\{I_1, \ldots, I_M\}$ are stored in the database in their encrypted forms $\{\widetilde{I}_i\}, i = 1, \ldots, M$, we denote the corresponding encrypted index or encrypted features as $\mathcal{E}(I_i)$. Trying to identify the database content, the adversary can compare $\mathcal{E}(I_i)$ for all $i$ and divide the database into several groups such that images in each group have similar encrypted indexes or features and are likely to have similar content.
As will be shown subsequently through a rigorous security analysis, if we assume for the proposed schemes that the adversary has no prior knowledge about the database content and that the features and indexes are properly encrypted, the adversary will not be able to infer the plaintext image content using just the grouping information.
During retrieval, the encrypted index or feature $\varepsilon(Q)$ of the query image is sent to the server and the encrypted versions of images similar to the query image are returned. The adversary can thus obtain information about the retrieval history in terms of which images are retrieved each time, but no content information will be revealed.
A secure retrieval scheme should be able to prevent the adversary from learning the following: (1) the plaintext versions of the database images $\{I_i\}$ and the query image $Q$; (2) the secret key $K$ used in the encryption; and (3) any function of the images $f(I_i)$ and $f(Q)$, such as the plaintext features of the images. We provide a security definition formulated using the concept of content indistinguishability, which is adapted from the indistinguishability definition [34] widely used in cryptanalysis. We show that most of the proposed schemes are provably secure under this definition.
We begin with a review of the concept of negligible function. It is used in security analysis to characterize the probability that a computationally-bounded adversary successfully breaks a computationally secure encryption scheme [29]. A function $f(\cdot)$ is negligible if for every polynomial $p(\cdot)$, there exists an $N$ such that for all integers $n > N$ it holds that $f(n) < 1/p(n)$.
According to modern cryptography [29], events that occur with negligible probability are unlikely to occur and they can be ignored for all practical purposes. In security analysis, the parameter $n$ typically determines the length of the secret key and the security of the encryption scheme. With an increase in $n$, the probability that an adversary successfully breaks the encryption scheme decays faster than the inverse of any polynomial function of $n$. 
We now provide the definition of content indistinguishability for characterizing the security of the proposed schemes. **Content indistinguishability:** Consider the following security experiment: an adversary chooses two images $I_0$ and $I_1$. The content owner chooses a secret key $K_0$ and encrypts the two images using a feature or index encryption scheme $\mathcal{E}$ to obtain the encrypted indexes $\mathcal{E}(I_0, K_0)$ and $\mathcal{E}(I_1, K_0)$. The content owner then randomly chooses a value $b$ from $\{0,1\}$ with equal probability and sends the encrypted index $\mathcal{E}(I_b, K_0)$ to the adversary.
Using any probabilistic polynomial time algorithm, the adversary outputs a number $b'$ as an estimate of $b$. The retrieval scheme satisfies the property of content indistinguishability if

$$| \Pr(b' = b) - \frac{1}{2} | \leq \text{negl}(n),$$

for every choice of $I_0, I_1$ by the adversary, where $\text{negl}(n)$ is a negligible function of the security parameter $n$. The probability is taken over all possible random choices by the adversary and in the experiment, such as the secret key $K_0$ and the value of $b$. 
The **content indistinguishability** definition essentially states that a computationally bounded adversary cannot distinguish between two encrypted features or indexes even if he/she has knowledge of the plaintext features or indexes. Under the COA model where the attacker is assumed to have knowledge of the ciphertext only, if a feature/index encryption scheme satisfies the above definition, the adversary will not be able to distinguish any two encrypted images in terms of their content, and thus confidentiality is preserved under such a COA model.
To prove that the proposed encryption schemes satisfy the above security definition, we carry out reduction to relate the security of the entire scheme to the security of some basic cryptographic building blocks, such as pseudorandom functions and pseudorandom permutations [29]. Since these cryptographic primitives are considered hard to break by any probabilistic polynomial time algorithm, we can prove that an encryption scheme satisfies the above security definition by showing that breaking the scheme is equivalent to breaking the cryptographic primitives.
A1. Proof for random projection and randomized unary encoding: We analyze the security of feature protection schemes based on random projection and randomized unary encoding together, both of which have projection onto random vectors as the last step. Given any two images \( I_0 \) and \( I_1 \), we denote their corresponding feature vectors as \( h_0 \) and \( h_1 \), which are normalized so that for some constant \( c \):

\[
\| h_0 \|_2 = \| h_1 \|_2 = c
\]
For the random projection based scheme, the content owner will choose a secret key $K_0$ and generate a pseudorandom matrix $R_0$ whose elements are independent Gaussian variables from $N(0, 1)$. Encryption by random projection is denoted as

$$\tilde{f}_0 \triangleq \mathcal{E}(f_0, K_0) = R_0 \cdot f_0 = (r_1 \cdot f_0, r_2 \cdot f_0, \ldots, r_m \cdot f_0),$$

$$\tilde{f}_1 \triangleq \mathcal{E}(f_1, K_0) = R_0 \cdot f_1 = (r_1 \cdot f_1, r_2 \cdot f_1, \ldots, r_m \cdot f_1),$$

where $r_i$ is the $i^{th}$ row of the matrix $R_0$. 
Assuming now that the matrix $R_0$ is truly random with each element as independent standard Gaussian variables, the encrypted features $\tilde{f}_0, \tilde{f}_1$ can be considered as vectors chosen uniformly at random from the distribution of vectors whose components are independent Gaussian variables $N(0, c^2)$, because $r_i \cdot f_b$ are independent Gaussian $N(0, c^2)$ for any $i \in \{1, \cdots, m\}$ and $b \in \{0, 1\}$. The conditional probability of the encrypted feature given the plaintext feature only depends on the value of $c$. Since all plaintext features are normalized to have the same value of $c$, we have $\Pr[\tilde{f}_b | f_b] = \Pr[\tilde{f}_b]$, for any $f_b$, which satisfies the definition of perfect secrecy [35].
A cryptosystem satisfies **perfect secrecy** if the posterior probability of the ciphertext given the plaintext is exactly the same as the prior probability of the ciphertext, for all ciphertexts and plaintexts. Therefore, the probability that any probabilistic polynomial time algorithm can distinguish $\tilde{f}_0$ and $\tilde{f}_1$ is exactly $1/2$. 
After replacing the truly random projection matrix $R_0$ with a cryptographically secure pseudorandom matrix, we denote the probability that $\tilde{f}_0$ and $\tilde{f}_1$ are distinguished by a probabilistic polynomial time attacker as $\Pr(b' = b) = \frac{1}{2} + \varepsilon(n)$. 
If the function $\varepsilon(n)$ is not negligibly small, it would imply that there exists a polynomial time algorithm that can distinguish a truly random sequence from a pseudorandom sequence, which contradicts the definition of cryptographically secure pseudorandom sequence of numbers [29]. Thus $\varepsilon(n)$ must be negligibly small implying that feature protection schemes based on random projection or on randomized unary encoding satisfy the definition of content indistinguishability.
Although the security analysis of the random projection scheme and the randomized unary encoding scheme are the same under the COA model, we will show in Section V-B that the additional encryption stage in the randomized unary encoding scheme makes it more secure than the random projection scheme in the KPA model. Proof for feature protection scheme based on bit-plane randomization and index encryption using secure min-hash can be carried out in a similar fashion, by showing that the distribution of the encrypted features does not depend on the plaintext features.
A2. Security for inverted index based scheme:

As described in Section III-B1, the inverted index based scheme encrypts the inverted index by pseudorandom permutation and order preserving encryption. Although the order statistics are scrambled by pseudorandom permutation, the variance of the index may still be approximately preserved by order preserving encryption. Therefore, in the security experiment of content indistinguishability, an adversary can choose two images whose bag-of-words representations have very different variance and such variance can be preserved after encryption, allowing the adversary to distinguish the two images.
Since order preserving encryption (OPE) reveals some information about the plaintext in terms of its variance, the retrieval scheme based on OPE does not satisfy the content indistinguishability definition. A proper security definition for OPE with clear practical implications would be desirable but is a challenging task. Despite the security limitations, order preserving encryption can enable efficient indexing and range query in the encrypted domain, which would otherwise be much more difficult. We refer readers to [36] for a detailed security analysis of OPE and recommend that OPE based scheme should be used with caution in practice.
A3. **Retrieval performance using a wrong key**: To demonstrate the security of our secure retrieval schemes under the COA model, we perform attacks by first extracting and encrypting the features or indexes from some plaintext images using a randomly chosen key, and then searching the database using these encrypted features as a query and analyzing the retrieved encrypted images.
For a secure retrieval scheme, the query index encrypted by a wrong key is equally like to be closest to any encrypted index in the database. Therefore, retrieval from an encrypted database using a wrong key is equivalent to picking images randomly from the database. For verification, we perform retrieval using every image in the database as the query but encrypt the query index or features with a key different from the one used in encryption. The precision-recall curves for all the schemes are shown in Fig. 12, Fig. 13, and Fig. 14.
Since the database has 100 images in each of the 10 categories, random selection from the database would imply a precision value around 0.1 for all recall values, which is confirmed in the above figures. Although the server can search the database using any plaintext images, it cannot learn anything about the image content from the ranking of the returned images, as the returned images are essentially a random selection from the database.
Fig. 12: Retrieval using a wrong key for feature protection schemes
Fig. 13: Inverted index scheme using wrong key
Fig. 14: Min-Hash scheme using wrong key
B. Known Plaintext Attack (KPA)

In the KPA model, the attacker is assumed to know some plaintext images $I$ and their corresponding encrypted features or indexes $\mathcal{E}(I, K_0)$, where $K_0$ is the secret key used by the content owner during encryption and is unknown to the attacker. In the semi-honest adversary model, the known-plaintext attack is a less serious concern as compared to the ciphertext-only attack. Hence, we provide an informal security analysis for the proposed schemes under the KPA model, as a supplement to our formal security analysis in the COA model.
One immediate consequence of the information available to the attacker under the KPA model is that the attacker can search the database using the known \( \mathcal{E}(I, K_0) \), and the top images returned from the retrieval would have similar visual appearance to the known image \( I \) if the distance between their encrypted features is small. For example, the adversary can search the database using a known image of sky with the corresponding encrypted features or indexes and identify how many images in the database might be visually similar to the sky image based on the retrieval results. If the attacker has access to an increasing number of plaintext images with diverse content and their encrypted features, more information about the database content would be revealed.
To evaluate the security of our proposed retrieval schemes under the KPA model, we compare the number of plaintext-ciphertext pairs required for an attacker to recover the secret key used to encrypt features or indexes. To illustrate the trade-off between security and computational complexity, we provide analysis for the two feature protection schemes based on random projection and unary encoding. We shall see that more encryption stages can bring higher security at a higher computational cost. Security analysis for the other schemes can be carried out in a similar fashion.
B1. **Analysis for random projection**: Assume that the attacker knows $k$ pairs of plaintext features and their encrypted $\tilde{f}_i, i = 1, 2, \ldots, k$. We can write $\tilde{F} = R \cdot F$, where $F$ and $\tilde{F}$ have $f_i$ and $\tilde{f}_i$ as their $i^{th}$ columns, respectively. The key-dependent projection matrix $R$ can be uniquely determined if $F$ is invertible. To make $F$ invertible, the attacker needs at least $n$ linearly independent plaintext features. Overall, to break the random projection scheme, the attacker needs $O(n)$ pairs of plaintext and ciphertext. Similar analysis has been used for evaluating the security of different image hashing schemes [37].
B2. Analysis for randomized unary encoding: In randomized unary encoding, the feature vector of the image $f = \{f_1, \ldots, f_n\}$ is first converted by unary encoding to a binary string $U(f) = [U(f_1), \ldots, U(f_n)]$, whose length is $nM$ where $M$ is the maximum possible feature value. This binary string $U(f)$ is XORed with a random binary string, then randomly permuted, and finally projected to give the encrypted feature $\tilde{f} = R \cdot \mathcal{E}(U(f))$, where $\mathcal{E}(\cdot)$ here represents the random XOR and permutation.
If the attacker knows the intermediate stage, then $\mathcal{O}(nM)$ pairs of are required to deduce the XOR and permutation pattern, and similarly, $\mathcal{O}(nM)$ pairs of are required to obtain the random projection matrix $R$. However, due to the concatenation of the two encryption stages, the randomized unary encoding achieves higher security than the single encryption of random projection.
Using similar ideas in differential cryptanalysis [38], in the best case for the attacker, if the known feature vectors satisfy some special property, for example, $U(f_1)$ and $U(f_2)$ differ only in 1 bit, $\varepsilon(U(f_1))$ and $\varepsilon(U(f_2))$ obtained after XOR and random permutation will have only one different component, so the attacker can obtain individual columns of the projection matrix $R$. 
However, due to the pseudorandom permutation, given all the columns of $R$, there are $O((nM)!!)$ possibilities for the matrix $R$. Furthermore, the system $\tilde{f} = R \cdot \varepsilon(U(f))$ is typically underdetermined.
Thus knowing $R$ will not help the attacker uniquely determine $\mathcal{E}(U(f))$. Because the projection matrix $R$ is typically non-invertible, the concatenation of the two encryption stages here is resilient to the meet-in-the-middle attack which undermines the security increase of double encryption using DES from a squared key space to only a doubled key space [39]. A formal analysis on the security increase from the multiple encryption is desirable but is outside the scope of the current work.
In summary, randomized unary encoding exhibits much better security than the simple random projection scheme. The higher security of the randomized unary encoding comes from the use of two encryption stages, which reflects the trade-off between security and computational complexity. Similar arguments also hold for the secure min-hash scheme and secure inverted index scheme, where the two step encryption in secure min-hash scheme provides better security than one-step encryption schemes such as random projection and secure inverted index.
The chosen plaintext attack (CPA) model, where the attacker can obtain encrypted features for any chosen plaintext image, will be a severe attack for the proposed secure retrieval schemes, because the attacker in the CPA model can essentially query the database using any plaintext images and infer the database content based on the retrieval results.