Anonymous Fuzzy Identity-based Encryption for Similarity Search

Ye Zhang, Nikos Mamoulis, D.W.L. Cheung, S.M. Yiu, and W.K. Wong

Department of Computer Science, The University of Hong Kong

Abstract. In this paper, we consider the problem of predicate encryption and focus on the predicate for testing whether the hamming distance between the attribute X of a data item and a target V is equal to (or less than) a threshold t where X and V are of length m. Existing solutions either do not provide attribute protection or produce a big ciphertext of size $O(m2^m)$. For the equality version of the problem, we provide a scheme which is match-concealing (MC) secure and the sizes of the ciphertext and token are both $O(m)$. For the inequality version of the problem, we give two practical schemes. The first one, also achieving MC security, produces ciphertext with size $O(m^{t_{max}})$ if the maximum value of t, t_{max}, is known in advance and is a constant. We also show how to update the ciphertext if the user wants to increase t_{max} without constructing the ciphertext from scratch. On the other hand, in many real applications, the security requirement can be lowered from MC to MR (match-revealing). Our second scheme, which is MR secure, produces ciphertext of size $O(m)$ and token of size $O((t+1)m)$ only.

Key words: predicate encryption, anonymous fuzzy identity-based encryption, inner-product encryption

1 Introduction

It is getting more popular for a data owner to take advantage of the storage and computing resources of a data center to hold the data in encrypted form. Users will be given a token (by the owner) to access the data so that only authorized records can be retrieved and later be decrypted on the user site. Due to the privacy and security concern, it is obvious that the data will not be decrypted at the data center and checked against the criteria one by one. Thus computation is required to be carried out on encrypted data directly. Examples are retrieval of encrypted documents based on keyword matching, selection of encrypted audit logs using multi-dimensional range query on authorized IP addresses or port numbers, and hamming distance based similarity search on encrypted DNA sequence data. The problem, in fact, has received much attention from both database community [16, 3, 17, 13, 18, 26] and cryptography community [25, 4, 23, 10, 19].

In general, the problem can be stated as follows. For each data item M, there is an associated attribute value X (X may not be part of the record M)

and let $f : \{0,1\}^* \rightarrow \{0,1\}$ be a predicate which represents the computation we want to carried out on ciphertexts such that the data item M can be successfully decrypted if and only if $f(X) = 1$. Authorized users will obtain a token generated by the owner in order to perform the predicate evaluation. The predicate can take additional parameters, so a different token can be generated for a different parameter value which increases the flexibility of the data owner to provide different access power to different users. Here is an example. Each medicate record (M) is encrypted along with a selected region of the DNA sequence (X) of the person. When a research team is authorized to investigate the relationship between a certain DNA sequence V with diseases, this team would acquire a token which corresponds to the predicate f such that $f(X) = 1$ if and only if $HammingDist(X, V) \leq t$, say $t = 5$. By using the token, the research team would decrypt all medicate records for which the corresponding DNA sequence is similar to V. In the above motivating example, it is obvious that the research team should not infer any information on records for which the corresponding attribute X which is far away from V (i.e. $HammingDist(X, V) > 5$) since they are not authorized to do so. And it is desirable that the ciphertext $E(pk, I, M)$, where pk is the public key generated by the data owner, is the same for different V and t values such that the encryption of data items needs only to be done once. This emerging branch of encryption schemes are referred as *predicate encryption*.

Here we focus on the predicate f that tests whether the hamming distance between V and X is equal to (or less than) a certain threshold t, where V and X can be assumed as bit vectors of equal length m. Similarity search based on hamming distance[1] is an important searching criterion for record retrieval. This leads to many interesting applications in databases, bioinformatics, and other areas. Note that V and t can vary and will be given to the owner for the generation of a token independent of the ciphertext $E(pk, I, M)$.

The security of predicate encryption [19] can be classified into (1) protecting the data item only; and (2) protecting both the data item and attributes. Attribute protection is usually referred as *anonymous* in general and can be further classified into two levels: *match-revealing (MR)* [23] and *match-concealing*[2] *(MC)* [10, 19]. The difference between MR and MC is that attributes will remain hidden in MC level even if it satisfies the predicate. In our "medicate record" example, we sometimes require the encryption scheme to be anonymous such that the DNA sequence is protected. It depends on applications whether we require MC or MR level of security. So far, the predicate encryption scheme supporting this predicate is the one in [24], called "Fuzzy Identity-Based Encryption". However, it does not provide the property of anonymity (i.e., attribute protection). In this paper, we propose "anonymous fuzzy identity-based encryption" schemes to handle both the equality threshold and the inequality threshold (less than or equal to) versions of the predicate.

[1] It is well known that hamming distance of two bit vectors can provide a good necessary condition for the corresponding edit distance [11, 2] which would be useful in many database applications

[2] In [19], match-concealing is called attribute-hiding.

It is not trivial how to make the scheme in [24] anonymous. On the other hand, there is a straight-forward solution [10] (see Appendix B) that can support the predicate we study with the property of anonymity and is MC secure. Their scheme provides a general construction to support any polynomial computable predicate. However, their scheme embeds (pre-computes for) every possible value of V and t in the ciphertext even for the equality threshold version of the problem (the same applies to the inequality version), thus the size of each ciphertext is $O(m2^m)$ which is impractical even for moderate m although the token size is constant.

1.1 Our contributions

For the equality threshold version, we provide an anonymous fuzzy identity-based encryption scheme achieving the MC level of security with both the sizes of ciphertext and token equal to $O(m)$. The construction is based on an inner-product encryption scheme in [19]. The core idea is to represent the hamming distance computation as an inner product such that X and V can be separated into the ciphertext and the token, respectively, so that V can be given only when the token is needed to be generated.

For the inequality threshold version, we provide two practical schemes to solve the problem. In many applications (e.g. in bioinformatics applications), $t << m$. Even assuming that we know the maximum value of t (t_{max}) in advance and is a constant, the size of the ciphertext produced by the solution based on [10] is still $O(2^m)$. In our first scheme, also achieving the MC security level, the sizes of ciphertext is only $O(m^{t_{max}})$ (precisely, $\sum_{i=0}^{t_{max}+1} \binom{m}{i}$) which is much smaller than $O(2^m)$ if $t_{max} << m$. The core of this scheme is to come up with an inner product expression with a total number of $\sum_{i=0}^{t+1} \binom{m}{i}$ terms to express whether $HammingDist(X, V) \leq t$ and modifying the scheme in [19] to a new primitive to support our encryption scheme. We also show how to update the ciphertext if the user wants to increase the value of t_{max}.

On the other hand, in many applications (in particular for those where the attribute X is part of the data item M), we only require the schemes to be MR secure. By lowering the security requirement to MR, we provide another scheme in which the sizes of ciphertext and token are only $O(m)$ and $O((t+1)m$, respectively which is attractive for real applications.

1.2 Related Works

The predicate that was studied in the very beginning is "exact keyword matching". That is, whether the value hidden by the token is equal to the attribute value hidden in the ciphertext. Schemes that only provide data item security are basically "Identity-based encryption" [22, 6]. Schemes protecting both the data item and the attributes were initialed by Song *et al.* [25] in the private-key setting and by Boneh *et al.* [5] in the public-key setting. Relationship between [5] and "Anonymous Identity-based encryption" [9, 14] was revisited in [1].

Then, range query as the predicate was also considered. Boneh *et al.* devised an Augmented Broadcast Encryption [8] which allows checking if the attribute value falls within a range on encrypted data. Their scheme also provides attribute protection. Then, Boneh and Waters [10] extended it to multi-dimensional range query. Shi *et al.* [23] also devised another scheme for multi-dimensional range query, but the scheme is MR secure.

The predicate investigated in this paper was initialed by [24] which only protects the data item. This predicate is powerful and has many applications other than those stated in [24]. However, there is no practical scheme supporting this predicate with attribute protection in a public-key setting. Park *et al.* [21] investigated this problem in the private-key setting and is IND2-CKA secure. Liesdonk [20] also investigated this problem in his master thesis. His scheme is in a public-key setting. However, the scheme requires the threshold value t to be fixed in the setup time.

From the technical point of view, the most related work is [19]. They provided schemes for handling predicates represented as inner products. As we will show, their formulation of using inner products with bounded disjunction is powerful and is used as the framework to build our encryption schemes for the hamming distance similarity comparison predicate.

1.3 Paper Organization

The rest of this paper is organized as follows. Section 2 introduces the framework of the encryption scheme, the security models and the hard problem assumption. Section 3 presents the scheme for the equality threshold version (i.e., $HammingDist(V, X) = t$) of the problem and Section 4 deals with the inequality threshold version (i.e., $HammingDist(V, X) \leq t$) of the problem. We conclude the paper in Section 5.

2 Preliminaries

We assume that the attribute X is represented as a bit vector. The attribute V (referred as the *target attribute*) provided by the user to generate the token is also a bit vector of the same length as X. In the rest of the paper, for simplicity, we focus on predicate-only encryption, that is, we assume that we only have X without M. So, the scheme will output "1" to indicate the decryption is successful ($f(X) = 1$) and "0" otherwise. Note that extending solutions for predicate-only encryption to include the data item M can be done easily [19]. Also, there exist applications that we only need to encrypt the attribute X and based on the decryption result to retrieve the corresponding records separately.

2.1 Framework

An anonymous fuzzy identity-based encryption scheme Π consists of the following four probabilistic polynomial-time (PPT) algorithms.

- $\mathtt{Setup}(1^n)$: On an unary string input 1^n where n is a security parameter, it produces the public-private key pair (pk, sk).
- $\mathtt{Encrypt}(pk, X)$: Taking the public key pk and the attribute vector X, it outputs the ciphertext C.
- $\mathtt{GenTK}(pk, sk, V, t)$: The token generation algorithm takes the public key pk, private key sk, outputs the token TK for the vector V and threshold t.
- $\mathtt{Test}(pk, TK, C)$: Given the ciphertext C, the token TK, and the public key pk, it outputs "1" if the hamming distance between the vector X associated with C and the vector V associated with TK is equal to t (is less than or equal to t for the inequality version); "0" otherwise.

2.2 Security models

We define MR and MC security in the Selective-ID [12, 10, 23, 19] model as follow.

Definition 1. (Selective-ID secure in the match-concealing model) *An anonymous fuzzy identity-based encryption scheme* $\Pi = (\mathtt{Setup}, \mathtt{Encrypt}, \mathtt{GenTK}, \mathtt{Test})$ *is MC secure if for all probabilistic polynomial-time Turing machine (adversary)* $\mathcal{A}$*, the advantage of* $\mathcal{A}$ *in the following game is negligible.*

Setup: Adversary $\mathcal{A}(1^n)$ outputs two possible equal-length vectors X_0 and X_1 to challenger $\mathcal{C}$. The challenger $\mathcal{C}$ takes a security parameter n and runs $\mathtt{Setup}$ to generate pk and sk. $\mathcal{C}$ sends pk to $\mathcal{A}$.
Challenge: The challenger $\mathcal{C}$ picks a random bit $b \in \{0, 1\}$, computes and returns $C^* \stackrel{\$}{\leftarrow} \mathtt{Encrypt}(pk, X_b)$ to adversary $\mathcal{A}$.
Phase 1: Adversary $\mathcal{A}$ may adaptively request polynomially bounded numbers of tokens ("TK") for any (V_i, t_i), with the restriction that $t_i = HammingDist(V_i, X_j)$ for both $j = 0, 1$ or $t_i \neq HammingDist(V_i, X_j)$ for both $j = 0, 1$ (for inequality threshold, $t_i < HammingDist(V_i, X_j)$ for both $j = 0, 1$ or $t_i \geq HammingDist(V_i, X_j)$ for both $j = 0, 1$).
Guess: The adversary $\mathcal{A}$ output a guess bit b'. The advantage $\mathbf{Adv}^{\mathsf{MC}}_{\Pi,\mathcal{A}}(n)$ of $\mathcal{A}$ is defined as $|\Pr[b' = b] - \frac{1}{2}|$.

Definition 2. (Selective-ID secure in the match-revealing model) *An anonymous fuzzy identity-based encryption scheme* $\Pi = (\mathtt{Setup}, \mathtt{Encrypt}, \mathtt{GenTK}, \mathtt{Test})$ *is MR secure if for all probabilistic polynomial-time Turing machine (adversary)* $\mathcal{A}$*, the advantage of* $\mathcal{A}$ *in the following game is negligible.*

Setup: Adversary $\mathcal{A}(1^n)$ outputs two possible equal-length vectors X_0 and X_1. The challenger $\mathcal{C}$ takes a security parameter n and runs $\mathtt{Setup}$ to generate pk and sk.
Challenge: The challenger $\mathcal{C}$ picks a random bit $b \in \{0, 1\}$, computes and returns $C^* \leftarrow \mathtt{Encrypt}(pk, X_b)$ to adversary $\mathcal{A}$.
Phase 1: Adversary $\mathcal{A}$ may adaptively request polynomially bounded number of token ("TK") for any (V_i, t_i) with the restriction that $t_i \neq HammingDist(V_i, X_j)$ for both $j = 0, 1$ (for inequality threshold, $t_i < HammingDist(V_i, X_j)$ for both $j = 0, 1$).

Guess: The adversary $\mathcal{A}$ output a guess bit b'. The advantage $\mathbf{Adv}^{\mathsf{MR}}_{\Pi,\mathcal{A}}(n)$ of $\mathcal{A}$ is defined as $|\Pr[b' = b] - \frac{1}{2}|$.

2.3 The Hard Problem Assumption

The hard problem used in this paper is introduced by [19] and has been shown to "hold in generic bilinear groups of composite order $N = pqr$ as long as finding a non-trivial factor of N is hard".

Let $\mathcal{G}$ be a group generator which takes security parameter n as input and (randomly) output the group we use (i.e. $(p, q, r, \mathbb{G}, \mathbb{G}_T, e)$), where $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is a bilinear pairing which can be computed efficiently. We call the groups $\mathbb{G}$ and $\mathbb{G}_T$ bilinear groups. $\mathbb{G}$ and $\mathbb{G}_T$ are cyclic and share the same composite order $N = pqr$ where p, q and r are three large primes. We define Assumption 1 as follows.

Definition 3. *We say that $\mathcal{G}$ satisfies "Assumption 1" if for any probabilistic polynomial-time Turing machine $\mathcal{A}$, the advantage of $\mathcal{A}$,* $|\Pr[\mathcal{A}(\bar{Z}, T_1 = g_p^{b^2 s} R_3) = 1] - \Pr[\mathcal{A}(\bar{Z}, T_2 = g_p^{b^2 s} Q_3 R_3) = 1]|$, *is negligible in security parameter n, where $\bar{Z}$ is defined as:*

$$(p, q, r, \mathbb{G}, \mathbb{G}_T, e) \xleftarrow{\$} \mathcal{G}(1^n), N = pqr, g_p \xleftarrow{\$} \mathbb{G}_p, g_q \xleftarrow{\$} \mathbb{G}_q, g_r \xleftarrow{\$} \mathbb{G}_r$$
$$Q_1, Q_2, Q_3 \xleftarrow{\$} \mathbb{G}_q, R_1, R_2, R_3 \xleftarrow{\$} \mathbb{G}_r, a, b, s \xleftarrow{\$} \mathbb{Z}_p \text{ and outputs}$$
$$\bar{Z} = \{g_p, g_r, g_q R_1, h_p = g_p^b, k_p = g_p^{b^2}, g_p^a g_q, g_p^{ab} Q_1, g_p^s, g_p^{bs} Q_2 R_2\}$$

3 Scheme for Equality Threshold

In this section, we describe our scheme for handling the equality threshold version of the hamming distance predicate. Recall that both the target attribute V and the threshold t will only be known when the user wants to obtain a token from the owner and can vary for different users. The core step is to represent the hamming distance (see the following lemma) as an inner product so that the attribute of the data item X and the target attribute V can be encrypted separately into the ciphertext and token.

Lemma 1. *Given two vectors X and V of equal length m, $HammingDist(X, V)$ equals $\sum_{i=1}^{m} x_i(1 - 2v_i) + 1 \times \sum_{i=1}^{m} v_i$, where $X = x_1 \dots x_m$ and $V = v_1 \dots v_m$.*

Based on [19], we can generate a ciphertext C based on X and a token TK based on V such that given C and X, we can compute $e(g, g)^{s[\sum_{i=1}^{m} x_i v_i]}$, where s is a random number, which gives $1_{\mathbb{G}_T}$ only when $\sum_{i=1}^{m} x_i v_i = 0$ (i.e., the hamming distance is 0), or a random number otherwise. For evaluating whether $HammingDist(X, V) = t$, we can simply check if $e(g, g)^{s[\sum x_i(1-2v_i)+1\times(\sum v_i - t)]}$ equals $1_{\mathbb{G}_T}$ or not. The details of the scheme are as follow.

– `Setup`(1^n). It first randomly selects $\{h_{1,i}, h_{2,i}\}_{i\in[1,m]}, h_3$ and h_4 from $\mathbb{G}_p$, and then randomly selects $R, \{R_{1,i}, R_{2,i}\}_{i\in[1,m]}, R_3$ and R_4 from $\mathbb{G}_r$. It outputs

$$pk = \{g_p, g_r, Q = g_q R, [H_{1,i} = h_{1,i} R_{1,i}, H_{2,i} = h_{2,i} R_{2,i}]_{i\in[1,m]}, \\ H_3 = h_3 R_3, H_4 = h_4 R_4\}$$

and

$$sk = \{p, q, r, g_q, [h_{1,i}, h_{2,i}]_{i\in[1,m]}, h_3, h_4\}$$

- `Encrypt`$(pk, X = x_1...x_i...x_m)$. The encryption algorithm first randomly selects s, α, β from $\mathbb{Z}_N$ and $\{R'_{1,i}, R'_{2,i}\}_{i\in[1,m]}, R'_3, R'_4$ from $\mathbb{G}_r$. Then, it outputs the ciphertext C:

$$\{C_0 = g_p^s, [C_{1,i} = H_{1,i}^s Q^{\alpha x_i} R'_{1,i}, C_{2,i} = H_{2,i}^s Q^{\beta x_i} R'_{2,i}]_{i\in[1,m]}, \\ C_3 = H_3^s Q^{\alpha} R'_3, C_4 = H_4^s Q^{\beta} R'_4\}$$

- `GenTK`$(pk, sk, V = v_1...v_i...v_m, t)$. It randomly selects $\{r_{1,i}, r_{2,i}\}_{i\in[1,m]}, r_3, r_4$ and f_1, f_2 from $\mathbb{Z}_N$. Then, it randomly selects Q'' and R'' from $\mathbb{G}_q$ and $\mathbb{G}_r$ respectively. It outputs the token TK:

$$\{K_0 = Q''R''h_3^{-r_3} h_4^{-r_4} \prod_{i=1}^{m} h_{1,i}^{-r_{1,i}} h_{2,i}^{-r_{2,i}}, \\ [K_{1,i} = g_p^{r_{1,i}} g_q^{f_1(1-2v_i)}, K_{2,i} = g_p^{r_{2,i}} g_q^{f_2(1-2v_i)}]_{i\in[1,m]}, \\ K_3 = g_p^{r_3} g_q^{f_1(\sum v_i - t)}, K_4 = g_p^{r_4} g_q^{f_2(\sum v_i - t)}\}$$

- `Test`(pk, TK, C). It outputs 1 if $r = 1_{\mathbb{G}_T}$ and 0 otherwise, where $r = e(C_3, K_3)e(C_4, K_4)e(C_0, K_0) \prod_{i=1}^{m} e(C_{1,i}, K_{1,i})e(C_{2,i}, K_{2,i})$.

From above scheme, it is easily shown that the sizes of both ciphertext and token are $O(m)$.

Correctness analysis: Our construction is based on Lemma 1 to express the hamming distance as an inner product and then uses the inner-product encryption in [19], so the correctness can be guaranteed by the correctness of the inner-product encryption. The details of the correctness proof can be found in Appendix A.

Security analysis: Our encryption scheme can be proved to be MC secure using a similar proof as in [19] (Interested reader may refer to Section 4.3 and 4.4 of [19] for more details). The proof is based on a reduction as follows. Assume that there exits an adversary $\mathcal{A}_1$ that can win the MC game of our scheme with non-negligible advantage, we can use $\mathcal{A}_1$ as a subroutine to construct an adversary A_2 that can win the MC game of the scheme in [19] with non-negligible advantage. The idea of the construction is as follows. Since $HammingDist(X, V) = t$ (or $\neq t$) is corresponding to $\sum x_i(1-2v_i) + 1 \times (\sum v_i - t) = 0$ (or $\neq 0$), $\mathcal{A}_2$ could call $\mathcal{A}_1$ to generate two vectors to be challenged, then convert the vectors into inner-products. Then, based on the tokens obtained by $\mathcal{A}_1$, $\mathcal{A}_2$ can easily transform it to answer the challenger for the scheme in [19]. We omit the proof in this paper.

4 Schemes for Inequality Threshold

Recall that there is a straight-forward solution for solving the case of inequality threshold by using the idea from [10] which can be shown to be MC secure. The details of this straight-forward solution are given in Appendix B. The ciphertext of this solution is of size $O(m2^m)$ although the token size is constant which is not practical. In the following, we provide two practical schemes to handle the inequality threshold version.

4.1 Scheme with known t_{max}

If we can know the maximum value for the threshold t, t_{max}, in advance, we can have a scheme which is better than the straight-forward solution. The sizes of the ciphertext can be reduced to $O(\sum_{i=0}^{t_{max}+1}\binom{m}{i})$. In some applications, $t_{max} << m$ and is a constant. In that case, the size becomes $O(m^{t_{max}})$. The restriction on setting t_{max} seems to be quite stringent. At the end of this section, we show the how one can update the ciphertext if the user decides to increase t_{max} without computing ciphertext from scratch. We first present the scheme for known t_{max}.

The idea behind our construction is based on the observation that Hamming distance $H \leq t$ if and only if $H(H-1)\cdot ... \cdot(H-t) = 0$. Therefore, we try to design an encryption scheme whose decryption result is in the form of $e(g,g)^{sH(H-1)\cdot ... \cdot(H-t)}$ as what we have done (i.e. $e(g,g)^{s(H-t)}$) in the scheme for the equality threshold. We use the same technique to design this new scheme. We also note that the result of the decryption only reveals the information of whether $H \leq t$ which makes the scheme MC-secure.

Let $H(H-1)\cdot ... \cdot(H-t) = a_{t+1}H^{t+1} + \ldots + a_1 H$ assuming that all coefficients a_l can be determined. Hence, the problem becomes how to express H^k using x_i and v_i. We notice that $H^k = (\sum x_i(1-2v_i) + \sum v_i)^k$ can be expanded using Binomial theorem and therefore $H^k = \binom{k}{0}(\sum x_i(1-2v_i))^k + ... + \binom{k}{j}(\sum x_i(1-2v_i))^{k-j}(\sum v_i)^j + ... + \binom{k}{k}(\sum v_i)^k$. We also notice that $(\sum x_i(1-2v_i))^l = (x_1(1-2v_1) + x_2(1-2v_2) + ... + x_m(1-2v_m))^l$ which can be expanded by Multinomial theorem. To sum them up, we have

$$
\begin{aligned}
&H(H-1)(H-2)\cdot ... \cdot(H-t) = \\
&\textstyle\sum_{k_1+...+k_m=t+1}[\frac{(t+1)!}{k_1!k_2!...k_m!}(a_{t+1}\binom{t+1}{0})(1-2v_1)^{k_1}(1-2v_2)^{k_2}...(1-2v_m)^{k_m}]x_1^{k_1}x_2^{k_2}\cdots x_m^{k_m} + \\
&...+ \\
&\textstyle\sum_{k_1+...+k_m=l}[\frac{l!}{k_1!k_2!...k_m!}(a_{t+1}\binom{t+1}{t+1-l}(\sum v_i)^{t+1-l} + ... + a_l\binom{l}{0})(1-2v_1)^{k_1}...(1-2v_m)^{k_m}]x_1^{k_1}...x_m^{k_m} + \\
&...+ \\
&\textstyle\sum_{k_1+...+k_m=2}[\frac{2!}{k_1!k_2!...k_m!}(a_{t+1}\binom{t+1}{t-1}(\sum v_i)^{t-1} + ... + a_2\binom{2}{0})(1-2v_1)^{k_1}...(1-2v_m)^{k_m}]x_1^{k_1}...x_m^{k_m} + \\
&\textstyle\sum_{k_1+...+k_m=1}[(a_{t+1}\binom{t+1}{t}(\sum v_i)^t + ... + a_1\binom{1}{0})(1-2v_i)]x_i + \\
&\textstyle(a_{t+1}\binom{t+1}{t+1}(\sum v_i)^{t+1} + ... + a_1\binom{1}{1}\sum v_i)
\end{aligned}
$$

We note that $x_i^{k_i} = x_i$ in the above formula because each $x_i \in \{0,1\}$. Therefore, the number of terms in above formula can be future reduced because those terms involving $x_i^{k_i}$ can be incorporated into corresponding term involving x_i:

$H(H-1)(H-2) \cdot ... \cdot (H-t) = a_{t+1}(\sum v_i)^{t+1} + a_t(\sum v_i)^t + ... + a_1(\sum v_i) +$
$\sum_{1 \le j \le m} (\sum_{1 \le k_1 \le t+1} b_{k_1}(1-2v_j)^{k_1})x_j +$
$\sum_{1 \le j_1 < j_2 \le m}(\sum_{k_1+k_2 \le t+1; k_i \ge 1} \frac{(k_1+k_2)!}{k_1!k_2!} b_{k_1+k_2}(1-2v_{j_1})^{k_1}(1-2v_{j_2})^{k_2})x_{j_1}x_{j_2} + ... +$
$\sum_{1 \le j_1 < ... < j_l \le m} (\sum_{k_1+...+k_l \le t+1; k_i \ge 1} \frac{(k_1+...+k_l)!}{k_1!...k_l!} b_{k_1+...+k_l}(1-2v_{j_1})^{k_1}...(1-2_{j_l})^{k_l})x_{j_1}...x_{j_l} +$
$... + \sum_{1 \le j_1 < ... < j_{t+1} \le m} ((t+1)! b_{t+1}(1-2v_{j_1})...(1-2_{j_l}))x_{j_1}...x_{j_{t+1}}$

where $b_j = a_{t+1}\binom{t+1}{t+1-j}(\sum v_i)^{t+1} + ... + a_j\binom{j}{0}$ in the above formula. The total number of different terms is $\sum_{i=0}^{t+1}\binom{m}{i}$. We separate those elements involving $x_1, x_2, \ldots, x_m$ and $v_1, v_2, \ldots, v_m$ into ciphertext and token, respectively.

Since the number of terms in the above expression is decided by the threshold $t \le t_{max}$, the token size, therefore the decryption cost, can be reduced. The reason is that when we generate the token, we know what t is, and therefore, according to our explanation on computing $H(H-1)...(H-t)$, only $\sum_{i=0}^{t+1}\binom{m}{i}$ terms of token are needed, hence the token size can be reduced. To support our discussion here, we describe a new encryption scheme $\Pi_2 = (\mathsf{Setup}, \mathsf{Encrypt}, \mathsf{GenTK}, \mathsf{Test})$ slightly different from [19].

- `Setup`(1^n). It first randomly selects $\{h_{1,i}, h_{2,i}\}_{i \in [1, l_{max}]}$ from $\mathbb{G}_p$, and then randomly selects $R, \{R_{1,i}, R_{2,i}\}_{i \in [1, l_{max}]}$ from $\mathbb{G}_r$. It outputs

$$pk = \{g_p, g_r, Q = g_q R, [H_{1,i} = h_{1,i}R_{1,i}, H_{2,i} = h_{2,i}R_{2,i}]_{i \in [1, l_{max}]}\}$$

  and

$$sk = \{p, q, r, g_q, [h_{1,i}, h_{2,i}]_{i \in [1, l_{max}]}\}$$

- `Encrypt`$(pk, X = x_1...x_i...x_{l_{max}})$. The encryption algorithm first randomly selects s, α, β from $\mathbb{Z}_N$ and $\{R'_{1,i}, R'_{2,i}\}_{i \in [1, l_{max}]}$ from $\mathbb{G}_r$. Then, it outputs ciphertext C:

$$\{C_0 = g_p^s, [C_{1,i} = H_{1,i}^s Q^{\alpha x_i} R'_{1,i}, C_{2,i} = H_{2,i}^s Q^{\beta x_i} R'_{2,i}]_{i \in [1, l_{max}]}\}$$

- `GenTK`$(pk, sk, V = v_1...v_i...v_t)$. Note that $t \le m$. It randomly selects $\{r_{1,i}, r_{2,i}\}_{i \in [1,t]}$ and f_1, f_2 from $\mathbb{Z}_N$. Then, it randomly selects Q'' and R'' from $\mathbb{G}_q$ and $\mathbb{G}_r$ respectively. It outputs token TK:

$$\{K_0 = Q''R'' \prod_{i=1}^{t} h_{1,i}^{-r_{1,i}} h_{2,i}^{-r_{2,i}},\ [K_{1,i} = g_p^{r_{1,i}} g_q^{f_1 v_i}, K_{2,i} = g_p^{r_{2,i}} g_q^{f_2 v_i}]_{i \in [1,t]}\}$$

- `Test`(pk, TK, C). It computes $r = e(C_0, K_0)\prod_{i=1}^{t} e(C_{1,i}, K_{1,i}) e(C_{2,i}, K_{2,i})$. If $r = 1_{\mathbb{G}_T}$, it will output 1; otherwise it outputs 0.

Based on this new encryption primitive and above inner-product expression to test $HammingDist(X, V) \le t$, we construct our encryption scheme as follows:

- `Setup`(1^n): It first randomly selects $\{h_{1,l,i}, h_{2,l,i}\}_{l \in [1, t_{max}+1], i \in [1, \binom{m}{l}]}$ from $\mathbb{G}_p$. Then it randomly selects h_3, h_4 from $\mathbb{G}_p$. It also randomly selects $R, \{R_{1,l,i}, R_{2,l,i}\}_{l \in [1, t_{max}+1], i \in [1, \binom{m}{l}]}, R_3, R_4$ from $\mathbb{G}_r$. It outputs

$$pk = \{g_p, g_r, Q = g_q R, [H_{1,l,i} = h_{1,l,i} R_{1,l,i}, H_{2,l,i} = h_{2,l,i} R_{2,l,i}]_{l\in[1,t_{max}+1], i\in[1,\binom{m}{l})]}, H_3 = h_3 R_3, H_4 = h_4 R_4\}$$

and

$$sk = \{p, q, r, g_q, [h_{1,l,i}, h_{2,l,i}]_{l\in[1,t_{max}+1], i\in[1,\binom{m}{l})]}, h_3, h_4\}$$

– `Encrypt`$(pk, X = x_1...x_m)$: Encryption algorithm first randomly selects s, α, β from $\mathbb{Z}_N$ and $\{R'_{1,l,i}, R'_{2,l,i}\}_{l\in[1,t_{max}+1], i\in[1,\binom{m}{l})]}, R'_3, R'_4$ from $\mathbb{G}_r$. Then it outputs ciphertext C:

$$\{C_0 = g_p^s, [C_{1,l,i} = H_{1,l,i}^s Q^{\alpha x_{j_1}...x_{j_l}} R'_{1,l,i}, C_{2,l,i} = H_{2,l,i}^s Q^{\beta x_{j_1}...x_{j_l}} R'_{2,l,i}]_{l\in\{1...t_{max}+1\}; 1\le j_1<...<j_l\le m}, C_3 = H_3^s Q^{\alpha} R'_3, C_4 = H_4^s Q^{\beta} R'_4\}$$

– `GenTK`$(pk, sk, V = v_1...v_m, t)$: It randomly selects $\{r_{1,l,i}, r_{2,l,i}\}_{l\in[1,t+1], i\in[1,\binom{m}{i})]}, r_3, r_4$ and f_1, f_2 from $\mathbb{Z}_N$. Then, it randomly selects Q'' and R'' from $\mathbb{G}_q$ and $\mathbb{G}_r$ respectively. It outputs token TK:

$$\{K_0 = Q'' R'' h_3^{-r_3} h_4^{-r_4} \prod_{l=1}^{t+1} \prod_{i=1}^{\binom{m}{l}} h_{1,l,i}^{-r_{1,l,i}} h_{2,l,i}^{-r_{2,l,i}}$$

$$\begin{bmatrix} K_{1,1,i} = g_p^{r_{1,2,i}} g_q^{f_1[\sum_{1\le k_1\le t+1} b_{k_1}(1-2v_i)^{k_1}]} \\ K_{2,1,i} = g_p^{r_{2,2,i}} g_q^{f_2[\sum_{1\le k_1\le t+1} b_{k_1}(1-2v_i)^{k_1}]} \\ ... \\ K_{1,l,i} = g_p^{r_{1,l,i}} g_q^{f_1[\sum_{1\le k_1+...+k_l\le t+1; k_i\ge 1} \frac{(k_1+...+k_l)!}{k_1!k_2!...k_l!} b_{k_1+...+k_l}(1-2v_{j_1})^{k_1}...(1-2v_{j_l})^{k_l}]} \\ K_{2,l,i} = g_p^{r_{2,l,i}} g_q^{f_2[\sum_{1\le k_1+...+k_l\le t+1; k_i\ge 1} \frac{(k_1+...+k_l)!}{k_1!k_2!...k_l!} b_{k_1+...+k_l}(1-2v_{j_1})^{k_1}...(1-2v_{j_l})^{k_l}]} \\ ... \\ K_{1,t+1,i} = g_p^{r_{1,t+1,i}} g_q^{f_1[(t+1)! b_{t+1}(1-2v_{j_1})...(1-2v_{j_{t+1}})]} \\ K_{2,t+1,i} = g_p^{r_{2,t+1,i}} g_q^{f_2[(t+1)! b_{t+1}(1-2v_{j_1})...(1-2v_{j_{t+1}})]} \end{bmatrix}_{i\in[1,\binom{m}{l})]},$$

$$K_3 = g_p^{r_3} g_q^{f_1(a_{t+1}(\sum v_i)^{t+1} + a_t(\sum v_i)^t) + ... + a_1(\sum v_i)},$$

$$K_4 = g_p^{r_4} g_q^{f_2(a_{t+1}(\sum v_i)^{t+1} + a_t(\sum v_i)^t) + ... + a_1(\sum v_i)}\}$$

– `Test`(pk, sk, TK, C) : It outputs 1 if $r = 1_{\mathbb{G}_T}$ and 0 otherwise, where $r = e(K_0, C_0)e(K_3, C_3)e(K_4, C_4) \prod_{l=1}^{t+1} \prod_{i=1}^{\binom{m}{l}} e(K_{1,l,i}, C_{1,l,i}) e(K_{2,l,i}, C_{2,l,i})$.

Size of ciphertext in above scheme is $O(\sum_{l=0}^{t_{max}} \binom{m}{l})$ and size of token is $O(\sum_{l=0}^{t+1} \binom{m}{l})$ for threshold t. It can be shown that $r = e(g_q, g_q)^{(\alpha f_1 + \beta f_2) H(H-1)(H-2)...(H-t)}$, the correctness follows. The security of the scheme is stated in Theorem 1 and proved in appendix C.

Theorem 1. *Our construction* Π_1 *in Section 4.1 is Selective-ID secure in the math-concealing model under Assumption 1.*

Lastly, to show that it is feasible to compute the coefficients a_l, we have implemented an algorithm to calculate a_l. In fact, it can automatically calculate the coefficient appeared on the exponent of g_q for each term of token. It is written in C++. For example, with input $m = 100$ and $t = 3$, it took about 16 seconds to calculate all coefficients on an Intel Core 2 Due E6750 2.66GHz CPU platform.

Increasing t_{max}: It may be possible that the user wants to increase t_{max} to T', the following shows the idea of how to update the ciphertext to support T' provided the values α, β and s which are generated in the `Encrypt` procedure are kept without generating the ciphertext from scratch. The idea is based on the observation that the ciphertext only involves $x_1, x_2, ..., x_m$. More specifically, ciphertext contains $x_i = x_{j_1}, x_{j_1}x_{j_2}, \dots, x_{j_1}x_{j_2}\dots x_{j_{t_{max}+1}}$ where $1 \le j_1 < \dots < j_\ell \le m$, $\ell = 1, \dots, t_{max}+1$. Therefore, when we want to support T', we can just compute the additional terms involving $x_{j_1}x_{j_2}\dots x_{j_{t_{max}+2}}$, $x_{j_1}x_{j_2}\dots x_{j_{t_{max}+3}}$ etc. This update procedure can be shown to be MC-secure. Roughly speaking, once x_{j_1} is fixed, all terms including the one to be generated due to the increase in t_{max} have been fixed although they are not computed yet. In other words, an adaptive attack will not work since the attacker has no way to modify how the missing terms are generated no matter what T' it provides. Therefore, if the scheme for t_{max} is secure, the the update procedure is also secure. The proof and the details of how to perform this update can be found in Appendix E.

Note that in the worst case, $t_{max} = m$, the size of the ciphertext (and token) becomes $O(2^m)$. Although it is better than $O(m2^m)$ for the solution in Appendix B, it is not practical. So, this scheme should be used when t_{max} is small.

4.2 Scheme for Inequality Threshold with MR security

In this section, we consider another practical situation in which a lower security level (the MR security) is required. For many real applications, especially for those where the attribute X is part of the data item M, MR security is reasonable. Based on this security level, we present a practical scheme in which the size of the ciphertext is only $O(m)$ and the size of the token is only $O((t+1)m)$.

The idea behind our construction is simple. In MR security, as long as $HammingDist(X, V) \le t$, we allow the user to have full information about $HammingDist(X, V)$. So, we can make use of the scheme for the equality version presented in Section 3 and generate tokens for each $j = 0, 1, \dots, t$. Then, check if $HammingDist(X, V) = j$ for each j. Denote the scheme in Section 3 as $\Pi_1 = (\texttt{Setup}', \texttt{Encrypt}', \texttt{GenTK}', \texttt{Test}')$, we describe our scheme, $\Pi = (\texttt{Setup} = \texttt{Setup}', \texttt{Encrypt} = \texttt{Encrypt}', \texttt{GenTK}', \texttt{Test}')$, with MR security as follows.

- $\texttt{GenTK}(pk, sk, V)$: For each $j \in \{0, ..., t\}$, it runs $TK_j = \texttt{GenTK}'(pk, sk, V, j)$ and returns $TK = (TK_0, ..., TK_t)$.

- $\texttt{Test}(pk, TK, C)$: For each TK_j in TK, it runs $\texttt{Test}'(pk, TK_j, C)$ and checks if any of the results for any j is $1_{\mathbb{G}_T}$, it returns 1, otherwise it returns 0 indicating that $HammingDist(X, V) > t$.

It easily see that size of ciphertext in above scheme is $O(m)$ and size of token for threshold t is $O((t+1)m)$.The correctness of the scheme follows directly from the correctness of the scheme in Section 3. The security of the scheme is given in the following theorem. The proof can be found in Appendix D based on a simple reduction. Recall that the scheme in section 3 is MC secure under Assumption 1.

Theorem 2. *Scheme in Section 4.2 is selective-ID secure in match-revealing model if scheme in Section 3 is selective-ID secure in match-concealing model.*

5 Conclusion

In this paper, we investigated the problem of predicate encryption with attribute protection and focus on a hamming distance similarity comparison predicate. We consider both the equality and inequality versions on a user-specific threshold t.For the equality version, we provide a MC-secure scheme with both the sizes of ciphertext and token equal to $O(m)$ where m is the length of the attribute vector. For the inequality version, we provide two practical schemes, one works for the situation when the maximum value of t (t_{max}) is known and is MC secure. The other works for applications which require only MR security. The schemes provided in the paper should be applicable to real applications thus allowing the application communities to perform more useful computation on encrypted data.

The sizes of the ciphertext in our MC-secure scheme for the inequality threshold is $\sum_{i=0}^{t_{max}+1} \binom{m}{i}$. We leave it as an open problem whether it could be improved to $O((t+1)m)$ as in the MR-secure scheme.

References

1. M. Abdalla, M. Bellare, D. Catalano, E. Kiltz, T. Kohno, T. Lange, J. Malone-Lee, G. Neven, P. Paillier and H. Shi.: Searchable Encryption Revisited: Consistency Properties, Relation to Anonymous IBE, and Extensions. In Proceedings of CRYPTO'05.
2. A. Arasu, V. Ganti and R. Kaushik.: Efficient Exact Set-Similarity Joins. In Proceedings of VLDB'06.
3. R. Agrawal, J. Kiernan, R. Srikant and Y. Xu.: Order Preserving Encryption for Numeric Data. In Proceedings of SIGMOD'04.
4. M. Bellare, A. Boldyreva and A. O'neill.: Deterministic and Efficiently Searchable Encryption. In Proceedings of CRYPTO'07.
5. D. Boneh, G.D. Crescenzo, R. Ostrovsky, G. Persiano.: Public Key Encryption with keyword Search. In Proceedings of EUROCRYPT'04.
6. D. Boneh and M. Franklin.: Identity-Based Encryption from the Weil Pairing. In Proceedings of CRYPTO'01.
7. D. Boneh, E.J. Goh and K. Nissim.: Evaluating 2-DNF Formulas on Ciphertexts. In Proceedings of TCC'05.
8. D. Boneh and B. Waters.: A Fully Collusion Resistant Broadcast, Trace, and Revoke System. In Proceedings of CCS'06.

9. X. Boyen and B. Waters.: Anonymous Hierarchical Identity-Based Encryption (Without Random Oracles). In Proceedings of CRYPTO'06.
10. D. Boneh and B. Waters.: Conjunctive, Subset, and Range Queries on Encrypted Data. In Proceedings of TCC'07.
11. S. Chaudhuri, V. Ganti and R. Kaushik.: A Primitive Operator for Similarity Joins in Data Cleaning. In Proceedings of ICDE'06.
12. R. Canetti, S. Halevi and J. Katz.: A forward-secure public-key encryption scheme. In Proceedings of EUROCRYPT'03.
13. E. Damiani, S.D.C. Vimercati, S. Jajodia, S. Paraboschi and P. Samarati.: Balancing Confidentiality and Efficiency in Untrusted Relational DBMSs. In Proceedings of CCS'03.
14. C. Gentry.: Practical Identity-Based Encryption Without Random Oracles. In Proceedings of EUROCRYPT'06.
15. V. Goyal, O. Pandey, A. Sahai and B. Waters.: Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data. In Proceedings of CCS'06.
16. H. Hacigumus, B. Iyer and S. Mehrotra.: Providing Database as a Service. In Proceedings of ICDE'02.
17. H. Hacigumus, B. Iyer, C. Li and S. Mehrotra.: Executing SQL over Encrypted Data in the Database-Service-Provider Model. In Proceedings of SIGMOD'02.
18. B. Hore, S. Mehrotra and G. Tsudik.: A Privacy-Preserving Index for Range Queries. In Proceedings of VLDB'04.
19. J. Katz, A. Sahai and B. Waters.: Predicate Encryption Supporting Disjunctions, Polynomial Equations, and Inner Products. In Proceedings of EUROCRYPT'08.
20. P.P. van Liesdonk.: Anonymous and Fuzzy Identity-Based Encryption. Master thesis, Eindhoven University.
21. H.A. Park, B.H. Kim, D.H. Lee, Y.D. Chung and J. Zhang.: Secure Similarity Search. In Proceedings of IEEE International Conference on Granular Computing'07.
22. A. Shamir.: Identity-Based Cryptosystems and Signature Schemes. In Proceedings of CRYPTO'84.
23. E. Shi, J. Bethencourt, T-H.H. Chan, D. Song and A. Perrig.: Multi-Dimensional Range Query over Encrypted Data. In Proceedings of IEEE Symposium on Security and Privacy'07.
24. A. Sahai and B. Waters.: Fuzzy Identity-Based Encryption. In Proceedings of EUROCRYPT'05.
25. D.X. Song, D. Wagner and A. Perrig.: Practical Techniques for Searches on Encrypted Data. In Proceedings of IEEE Symposium on Security and Privacy'00.
26. W.K. Wong, D.W. Cheung, B. Kao and N. Mamoulis.: Secure k-NN Computation on Encrypted Databases. In Proceedings of SIGMOD'09.

Appendix

A Correctness analysis of the scheme for equality threshold

One key property behind the scheme in [19] is that if two elements a and b come from two different (prime) subgroups of $\mathbb{G}$, then $e(a, b) = 1_{\mathbb{G}_T}$. To simplify our correctness analysis, since the token of encryption scheme in Section 3 only involves elements from $\mathbb{G}_p$ and $\mathbb{G}_q$, we investigate consider the value of $r = \texttt{Test}(pk, TK, C)$ in subgroup $\mathbb{G}_p$ and $\mathbb{G}_q$ respectively only.

$$r_q =$$
$$e(g_q^{\alpha}, g_q^{-f_1(\sum v_i - t)})e(g_q^{\beta}, g_q^{-f_2(\sum v_i - t)}) \cdot 1 \cdot \prod e(g_q^{\alpha x_i}, g_q^{f_1(1-2v_i)})e(g_q^{\beta x_i}, g_q^{f_2(1-2v_i)})$$
$$= e(g_q, g_q)^{(\alpha f_1 + \beta f_2)(\sum v_i - t + \sum (1-2v_i)x_i)}$$
$$= e(g_q, g_q)^{(\alpha f_1 + \beta f_2)(HammingDist(x_1 \ldots x_i \ldots x_m, v_1 \ldots v_i \ldots v_m) - t)}$$

and

$$r_p = e(h_3^s, g_p^{r_3})e(h_4^s, g_p^{r_4})e(g_p^s, h_3^{-r_3})e(g_p^s, h_4^{-r_4}) \prod e(g_p^s, h_{1,i}^{-r_{1,i}})e(g_p^s, h_{2,i}^{-r_{2,i}}) \cdot$$
$$\prod e(h_{1,i}^s, g_p^{r_{1,i}})e(h_{2,i}^s, g_p^{r_{2,i}}) = 1$$

B A baseline construction from [10]

The main idea of this baseline construction is that we generate a ciphertext $(C_0, C_2, \ldots C_{t_{max}})$ for each possible $V \in \{0,1\}^m$ where we assume that the vector size is m. For each $j \in \mathbb{Z}_{t_{max}+1}$, if $HammingDist(X, V) \leq j$, then C_j will be an encrypted message for "true" based on an IND-CPA secure encryption scheme; otherwise, C_j will be an encrypted message for "false". When we $\texttt{Test}()$ for a certain (V, t), we can simply find the ciphertext for V and decrypt the t-th element C_t in that ciphertext. If $HammingDist(X, V) \leq t$, the decryption result should be 1. More specifically, we define the encryption scheme as follows. Let $(\texttt{G}, \texttt{E}, \texttt{D})$ be an IND-CPA secure encryption scheme.

- $\texttt{Setup}(1^n)$: Run $\texttt{G}(1^n)$ for $(t_{max}+1)2^m$ times to generate $(pk_{l,j}, sk_{l,j})_{\{l \in \{0,1\}^m, j \in \mathbb{Z}_{t_{max}+1}\}}$. Return $\{pk_{l,j}\}_{\{l \in \{0,1\}^m, j \in \mathbb{Z}_{t_{max}+1}\}}$ as the public-key pk and $\{sk_{l,j}\}_{\{l \in \{0,1\}^m, j \in \mathbb{Z}_{t_{max}+1}\}}$ as the secret key sk.

- $\texttt{Encrypt}(pk, X = x_1 \ldots x_m)$: For each $l \in \{0,1\}^m$, return $(C_0, C_1, \ldots, C_{t_{max}})_l$ where

$$C_j = \begin{cases} \texttt{E}_{pk_{l,j}}(\text{``}true'') & \text{if } HammingDist(X, l) \leq j; \\ \texttt{E}_{pk_{l,j}}(\text{``}false'') & \text{otherwise.} \end{cases}$$

- $\texttt{GenTK}(pk, sk, V, t)$: Return $sk_{V,t}$ as the token.

– $\texttt{Test}(pk, TK, C)$: It first finds $(C_0, C_1, ..., C_m)_V$ and computes $r = \mathtt{D}_{TK}(C_t)$. If r is equal to "true" then return 1; otherwise return 0.

The security of above solution comes from the IND-CPA secure encryption scheme we used [3], see appendix A of [10] for more details. We can rearrange the ciphertexts so that $C_{l,j}$ in our solution is corresponding to $C_{(t_{max}+1)l+j}$ in [10]'s proof and the rest of the proof is the same. We should also note that $(t_{max}+1)2^m$ should be $\leq poly(n)$ in security parameter n because all algorithms given above should be polynomial-time.

C Proof of Theorem 1

Definition 4. *The encryption scheme* $\Pi_2 = (\texttt{Setup}, \texttt{Encrypt}, \texttt{GenTK}, \texttt{Test})$ *defined in Section 4.1 is MC secure if for all probabilistic polynomial-time Turing machine (adversary)* $\mathcal{A}$*, the advantage of* $\mathcal{A}$ *in the following game is negligible.*

Setup: The adversary $\mathcal{A}(1^n)$ outputs two possible equal-length (l_{max}-length) vectors X_0 and X_1 to the challenger $\mathcal{C}$. The challenger $\mathcal{C}$ takes a security parameter n and runs *Setup* to generate pk and sk. $\mathcal{C}$ sends pk to $\mathcal{A}$.
Challenge: The challenger $\mathcal{C}$ picks a random bit $b \in \{0, 1\}$, computes and returns $C^* \stackrel{\$}{\leftarrow} \texttt{Encrypt}(pk, X_b)$ to adversary $\mathcal{A}$.
Phase 1: The adversary $\mathcal{A}$ may adaptively request polynomially bounded numbers of tokens ("TK") for any $V_i = v_{i,1}...v_{i,t_i}$ where $t_i \leq l_{max}$, subject to the restriction that $0 = \sum_{k=1}^{t_i} x_{j,k}v_{i,k}$ for both $j = 1$ or 0 or $0 \neq \sum_{k=1}^{t_i} x_{j,k}v_{i,k}$ for both $j = 0, 1$.
Guess: The adversary $\mathcal{A}$ outputs a guess bit b'. The advantage $\mathbf{Adv}^{\mathsf{MC}}_{\Pi_2,\mathcal{A}}(n)$ of $\mathcal{A}$ is defined as $|\Pr[b' = b] - \frac{1}{2}|$.

Lemma 2. *Our construction* Π_1 *in Section 4.1 is Selective-ID secure in the math-concealing model if* $\Pi_2 = (\texttt{Setup}, \texttt{Encrypt}, \texttt{GenTK}, \texttt{Test})$ *in Section 4.1 is secure under definition 4.*

Proof. This proof is by contradiction. We will show that if there exist Adversary $\mathcal{A}_1$ with non-negligible adversary ϵ with our construction Π_1, then we can construct an Adversary $\mathcal{A}_2$ also with non-negligible with scheme Π_2 under definition 4.
Setup: The adversary $\mathcal{A}_2(1^n)$ runs $\mathcal{A}_1(1^n)$. $\mathcal{A}_1$ outputs two equal-length vector $X_0 = x_{0,1}...x_{0,k}...x_{0,m}$ and $X_1 = x_{1,1}...x_{1,k}...x_{1,m}$. $\mathcal{A}_1$ passes X_0 and X_1 to $\mathcal{A}_2$. $\mathcal{A}_1$ also submits $t_{max} \leq m$ to $\mathcal{A}_2$.

The adversary $\mathcal{A}_2$ calculates $\tilde{X}_j = (1, x_{j,1}, x_{j,2}, ..., x_{j,m},$
$x_{j,1}x_{j,2}, ..., x_{j,s_1}x_{j,s_2}, ..., x_{j,m-1}x_{j,m}, ..., x_{j,1}x_{j,2}...x_{j,l}, ..., x_{j,s_1}x_{j,s_2}...x_{j,s_l}, ..., x_{m-l+1}...x_m,$
$..., x_{j,1}x_{j,2}...x_{j,t_{max}+1}, ..., x_{j,s_1}x_{j,s_2}...x_{j,s_{t_{max}+1}}, ..., x_{j,m-t_{max}}x_{j,m-t_{max}+1}...x_m)$ for

[3] The main idea is that we cannot distinguish $E(pk_i, \text{"}true\text{"})$ from $E(pk_i, \text{"}false\text{"})$ in the case of $HammingDist(X_0, V) \leq t$ but $HammingDist(X_1, V) > t$.

both $j = 0, 1$. Then, $\mathcal{A}_2$ submits $\tilde{X}_0$ and $\tilde{X}_1$ to challenger $\mathcal{C}$. Note that $l_{max} = \sum_{l=0}^{t_{max}+1} \binom{m}{l}$.
Challenge: The challenger $\mathcal{C}$ runs $\texttt{Setup}(1^n)$ to generate pk and sk; $\mathcal{C}$ sends pk to adversary $\mathcal{A}_2$. $\mathcal{A}_2$ rearranges pk into pk' according to the description of *Setup*() in scheme Π_1. $\mathcal{A}_2$ passes pk' to $\mathcal{A}_1$.

The challenger picks a random bit $b \in \{0, 1\}$. $\mathcal{C}$ computes and returns $C^* \xleftarrow{\$} \texttt{Encrypt}(pk, \tilde{X}_b)$ to $\mathcal{A}_2$. Adversary $\mathcal{A}_2$ rearranges C^* and sends it to $\mathcal{A}_1$.
Phase 1: The adversary $\mathcal{A}_1$ may adaptively request polynomially bounded number of tokens for any (V_i, t_i) subject to the restriction that $t_i < HammingDist(V_i, X_j)$ for both $j = 0, 1$ or $t_i \le HammingDist(V_i, X_j)$ for both $j = 0, 1$. When receiving valid $(V = v_1 ... v_m, t)$, $\mathcal{A}_2$ will calculate a $\tilde{t} = \sum_{l=0}^{t+1} \binom{m}{l}$-length vector $\tilde{V} = (a_{t+1}(\sum v_i)^{t+1} + ... + a_1(\sum v_i), ..., \sum_{1 \le k_1 + ... + k_l \le t+1; k_i \ge 1} \frac{(k_1 + ... + k_l)!}{k_1! ... k_l!} b_{k_1 + ... + k_l} (1 - 2v_{s_1})^{k_1} ... (1 - 2v_{s_l})^{k_l}, ..., (t+1)! b_{t+1} (1 - 2v_{m-t})(1 - 2v_{m-t+1}) ... (1 - 2v_m))$ We note that $HammingDist(V, X_j) \le t$ if and only if $\sum_{k=1}^{\tilde{t}} \tilde{X}_{j,k} \tilde{V}_k = 0$ as we explained in Section 4.1. $\mathcal{A}_2$ submits $\tilde{V}$ to the challenger $\mathcal{C}$ to acquire a token TK. $\mathcal{A}_2$ rearranges TK and sends it to $\mathcal{A}_1$.
Guess: $\mathcal{A}_1$ outputs a bit b'. And $\mathcal{A}_2$ passes b' to the challenger as its output.

Recall that $HammingDist(V, X_j) \le t$ if and only if $\sum_{k=1}^{\tilde{t}} \tilde{X}_{j,k} \tilde{V}_k = 0$, therefore valid token requests for Π_1 are still valid in Π_2. And $\mathbf{Adv}^{\mathsf{MC}}_{\Pi_2, \mathcal{A}_2}(n) = \mathbf{Adv}^{\mathsf{MC}}_{\Pi_1, \mathcal{A}_1}(n)$. That completes our proof.

Lemma 3. $\Pi_2 = (\mathsf{Setup}, \mathsf{Encrypt}, \mathsf{GenTK}, \mathsf{Test})$ *in Section 4.1 is secure under Definition 4 under Assumption 1.*

Proof. This proof is quite similar to the proof in [19]. For simplicity, we denote $T = l_{max}$. Given two equal-length vector $X = x_1 ... x_i ... x_T$ and $Y = y_1 ... y_i ... y_T$, loosely speaking, we try to prove $(X, X) \overset{c}{\equiv} (X, 0) \overset{c}{\equiv} (X, Y) \overset{c}{\equiv} (0, Y) \overset{c}{\equiv} (Y, Y)$ in the game defined by Definition 1. Where 0 stands for a T-length vector $(0, 0, ..., 0)$ and $\overset{c}{\equiv}$ is "computationally indistinguishable".

Let us prove $(X, X) \overset{c}{\equiv} (X, 0)$ first. Given $\{\bar{Z}, T\}$ where T may be equal to $T_1 = g_p^{b^2 s} R_3$ or $T_2 = g_p^{b^2 s} Q_3 R_3$, the challenger $\mathcal{C}$ answers $\texttt{Setup}(1^n)$ as follow: It randomly selects $\omega_{1,i}$ and $\omega_{2,i}$ from $\mathbb{Z}_p$. Then, it computes $h_{1,i} = h_p^{x_i} g_p^{\omega_{1,i}} = g_p^{bx_i + \omega_{1,i}}$ and $h_{2,i} = k_p^{x_i} g_p^{\omega_{2,i}} = g_p^{b^2 x_i + \omega_{2,i}}$ for $i = 1, ..., T$. It outputs pk to adversary $\mathcal{A}$: where $R_{1,i}$ and $R_{2,i}$ are randomly selected from $\mathbb{Z}_r$.

$$pk = \{g_p, g_r, Q = g_p R_1, [H_{1,i} = h_{1,i} R_{1,i}, H_{2,i} = h_{2,i} R_{2,i}]_{i=1,...,T}\}$$

The challenging ciphertext C^* is generated as follow: challenger $\mathcal{C}$ first randomly selects $R'_{1,i}$ and $R'_{2,i}$ from $\mathbb{G}_r$.

$$C = \{C_0 = g_p^s, [C_{1,i} = (g_p^{bs} Q_2 R_2)^{x_i} (g_p^s)^{\omega_{1,i}} R'_{1,i}, C_{2,i} = (T)^{x_i} (g_p^s)^{\omega_{2,i}} R'_{2,i}]_{i=1,...,T}\}$$

We note that $C_{1,i} = (g_p^{bs} Q_2 R_2)^{x_i} (g_p^s)^{\omega_{1,i}} R'_{1,i} = (g_p^{bx_i + \omega_{1,i}})^s Q_2^{x_i} R_2^{x_i} R'_{1,i} = h_{1,i}^s Q^{\alpha x_i} R_2^{x_i} R'_{1,i}$. where we denote $Q_2 = g_q^{\alpha}$. And $C_{2,i} = (T)^{x_i} (g_p^s)^{\omega_{2,i}} R'_{2,i} = (g_p^{b^2 x_i + \omega_{2,i}})^s Q^{\beta x_i} R_3^{x_i} R'_{2,i}$
$= h_{2,i}^s Q^{\beta x_i} R_3^{x_i} R'_{2,i}$ where $\beta = 0$ if $T = T_1$ and β is random from $\mathbb{Z}_N$ if $T = T_2$.

When receiving $V = v_1...v_i...v_t$ from adversary, challenger $\mathcal{C}$ generates corresponding token as follows: It firstly randomly selects $\tilde{f_1}$ and $\tilde{f_2}$ from $\mathbb{Z}_N$. It also randomly chooses $r'_{1,i}$ and $r'_{2,i}$ form $\mathbb{Z}_N$. Then, it calculates $K_{1,i} = (g_p^a g_q)^{\tilde{f}_1 v_i}(g_p^{ab}Q_1)^{-\tilde{f}_2 v_i} g_p^{r'_{1,i}} = g_p^{a\tilde{f}_1 v_i - ab\tilde{f}_2 v_i + r'_{1,i}} g_q^{\tilde{f}1 v_i - d\tilde{f}_2 v_i}$. (We denote $Q_1 = g_q^d$.) We denote $r_{1,i} = a\tilde{f}_1 v_i - ab\tilde{f}_2 v_i + r'_{1,i}$ and $f_1 = \tilde{f}1 v_i - d\tilde{f}_2 v_i$.

It also calculates $K_{2,i} = (g_p^a g_q)^{\tilde{f}_2 v_i} g_p^{r'_{2,i}} = g_p^{a\tilde{f}_2 v_i + r'_{2,i}} g_q^{\tilde{f}_2 v_i}$ where we denote $r_{2,i} = a\tilde{f}_2 v_i + r'_{2,i}$ and $f_2 = \tilde{f}_2$.

K_0 is calculated by $K_0 = Q''R'' \prod_{i=1}^t h_{1,i}^{-r_{1,i}} h_{2,i}^{-r_{2,i}}$
$= QR \prod_{i=1}^t g_p^{-(bx_i+\omega_{1,i})(a\tilde{f}_1 v_i - ab\tilde{f}_2 v_i + r'_{1,i})} g_p^{-(b^2 x_i + \omega_{2,i})(a\tilde{f}_2 v_i + r'_{2,i})}$
$= QR \prod_{i=1}^t g_p^{-(ab\tilde{f}_1 x_i v_i - ab^2\tilde{f}_2 x_i v_i + br'_{1,i}x_i + a\tilde{f}_1\omega_{1,i}v_i - ab\tilde{f}_2\omega_{1,i}v_i + r'_{1,i}\omega_{1,i} + ab^2\tilde{f}_2 v_i x_i + b^2 r'_{2,i}x_i + a\tilde{f}_2\omega_{2,i}v_i + r'_{2,i}\omega_{2,i})}$
$= QR \prod_{i=1}^t g_p^{-a(\tilde{f}_1\omega_{1,i}v_i + \tilde{f}_2\omega_{2,i}v_i) - ab(\tilde{f}_1 x_i v_i + \tilde{f}_2\omega_{1,i}v_i) - b(r'_{1,i}x_i) - b^2(r'_{2,i}x_i) + ab^2(-\tilde{f}_2 v_i x_i + \tilde{f}_2 v_i x_i) - (r'_{1,i}\omega_{1,i} + r'_{2,i}\omega_{2,i})}$
$= QR \prod_{i=1}^t [(g_p^a g_q)^{-(\tilde{f}_1\omega_{1,i}v_i + \tilde{f}_2\omega_{2,i}v_i)} \cdot (g_p^{ab}Q_1)^{-(\tilde{f}_1 x_i v_i + \tilde{f}_2\omega_{1,i}v_i)} \cdot h_p^{-(r'_{1,i}x_i)} \cdot k_p^{-(r'_{2,i}x_i)} \cdot g_p^{-(r'_{1,i}\omega_{1,i} + r'_{2,i}\omega_{2,i})}]$

Now, let us prove $(X, 0) \stackrel{c}{\equiv} (X, Y)$. The challenger $\mathcal{C}$ answers $\texttt{Setup}(1^n)$ as follows: It first randomly selects $\omega_{1,i}$ and $\omega_{2,i}$ from $\mathbb{Z}_p$. It randomly selects $R_{1,i}$ and $R_{2,i}$ from $\mathbb{G}_r$ Then, it calculates $h_{1,i} = h_p^{x_i} g_p^{\omega_{1,i}} = g_p^{bx_i+\omega_{1,i}}$ and $h_{2,i} = k_p^{y_i} g_p^{\omega_{2,i}} = g_p^{b^2 y_i + \omega_{2,i}}$. It outputs pk to adversary:

$$pk = \{g_p, g_r, Q = g_q R_1, [H_{1,i} = h_{1,i}R_{1,i}, H_{2,i} = h_{2,i}R_{2,i}]_{i=1...T}\}$$

The challenging ciphertext C^* is generated as follows: challenger $\mathcal{C}$ first randomly select $R'_{1,i}$ and $R'_{2,i}$ from $\mathbb{Z}_r$.

$$C = \{C_0 = g_p^s, [C_{1,i} = (g_p^{bs}Q_2R_2)^{x_i}(g_p^s)^{\omega_{1,i}}R'_{1,i}, C_{2,i} = (T)^{y_i}(g_p^s)^{\omega_{2,i}}R'_{2,i}]_{i=1...T}\}$$

We note that $C_{1,i} = (g_p^{bx_i+\omega_{1,i}})^s Q_2^{x_i} R_2^{x_i} R'_{1,i} = h_{1,i}^s Q^{\alpha x_i} R_2^{x_i} R'_{1,i}$ And $C_{2,i} = (g_p^{b^2 y_i + \omega_{2,i}})^s Q_3^{y_i} R_3^{y_i} R'_{2,i} = h_{2,i}^s Q^{\beta y_i} R_3^{y_i} R'_{2,i}$ where we denote $Q_3 = g_q^\alpha$. $\beta = 0$ if $T = T_1$ and β is random number in $\mathbb{Z}_N$ if $T = T_2$.

When receiving $V = v_1...v_i...v_t$ from adversary $\mathcal{A}$, the challenger $\mathcal{C}$ generates corresponding token as follows. According to Definition 4, V should be subject to (i) $\sum_{i=1}^t y_i v_i = 0 = \sum_{i=1}^t x_i v_i$ or (ii) $\sum_{i=1}^t y_i v_i \neq 0$ and $\sum_{i=1}^t x_i v_i \neq 0$. We handle token generation in this two conditions separately.

Case (i) $\sum_{i=1}^t y_i v_i = 0 = \sum_{i=1}^t x_i v_i$:

The challenger first randomly selects $\tilde{f_1}$, $\tilde{f_2}$ and $r'_{1,i}$, $r'_{2,i}$ from $\mathbb{Z}_N$. Then, it calculates $K_{1,i} = (g_p^a g_q)^{\tilde{f}_1 v_i} g_p^{r'_{1,i}} = g_p^{a\tilde{f}_1 v_i + r'_{1,i}} g_q^{\tilde{f}_1 v_i}$. We denote $r_{1,i} = a\tilde{f}_1 v_i + r'_{1,i}$ and $f_1 = \tilde{f}_1$. It also calculates $K_{2,i} = (g_p^a g_q)^{\tilde{f}_2 v_i} g_p^{r'_{2,i}} = g_p^{a\tilde{f}_2 v_i + r'_{2,i}} g_q^{\tilde{f}_2 v_i}$ where we denote $r_{2,i} = a\tilde{f}_2 v_i + r'_{2,i}$ and $f_2 = \tilde{f}_2$.

K_0 is calculated by $K_0 = Q''R'' \prod_{i=1}^t h_{1,i}^{-r_{1,i}} h_{2,i}^{-r_{2,i}}$
$= QR \prod_{i=1}^t g_p^{-(bx_i+\omega_{1,i})(a\tilde{f}_1 v_i + r'_{1,i})} g_p^{-(b^2 y_i + \omega_{2,i})(a\tilde{f}_2 v_i + r'_{2,i})}$

$= QR\prod_{i=1}^{t} g_p^{-(ab\tilde{f}_1 x_i v_i + br'_{1,i} x_i + a\omega_{1,i}\tilde{f}_1 v_i + r'_{1,i}\omega_{1,i} + ab^2\tilde{f}_2 y_i v_i + b^2 r'_{2,i} y_i + a\tilde{f}_2\omega_{2,i} v_i + r'_{2,i}\omega_{2,i})}$
$= QR\prod_{i=1}^{t} g_p^{-a(\omega_{1,i}\tilde{f}_1 v_i + \tilde{f}_2\omega_{2,i} v_i)} g_p^{-b(r'_{1,i} x_i)} g_p^{-b^2(r'_{2,i} y_i)} g_p^{-(r'_{1,i}\omega_{1,i} + r'_{2,i}\omega_{2,i})}$ since $\sum_{i=1}^{t} y_i v_i = 0 = \sum_{i=1}^{t} x_i v_i$. Then, $K_0 = QR\prod_{i=1}^{t} (g_p^a g_q)^{-(\omega_{1,i}\tilde{f}_1 v_i + \tilde{f}_2\omega_{2,i} v_i)} h_p^{-(r'_{1,i} x_i)} k_p^{-(r'_{2,i} y_i)} g_p^{-(r'_{1,i}\omega_{1,i} + r'_{2,i}\omega_{2,i})}$

Case (ii) $\sum_{i=1}^{t} y_i v_i = c_y \neq 0$ and $\sum_{i=1}^{t} x_i v_i = c_x \neq 0$:

The challenger firstly randomly chooses $\tilde{f}_1$, $\tilde{f}_2$ and $r'_{1,i}$, $r'_{2,i}$ from $\mathbb{Z}_N$. Then, it calculates $K_{1,i} = (g_p^a g_q)^{\tilde{f}_1 v_i} (g_p^{ab} Q_1)^{-c_y \tilde{f}_2 v_i} g_p^{r'_{1,i}} = g_p^{a\tilde{f}_1 v_i - abc_y\tilde{f}_2 v_i + r'_{1,i}} g_q^{\tilde{f}_1 v_i - dc_y\tilde{f}_2 v_i}$ where we denote $Q_1 = g_q^d$. And we denote $r_{1,i} = a\tilde{f}_1 v_i - abc_y\tilde{f}_2 v_i + r'_{1,i}$ and $f_1 = \tilde{f}_1 - dc_y\tilde{f}_2$. It also calculates $K_{2,i} = (g_p^a g_q)^{c_x\tilde{f}_2 v_i} g_p^{r'_{2,i}} = g_p^{ac_x\tilde{f}_2 v_i + r'_{2,i}} g_q^{c_x\tilde{f}_2 v_i}$. Where we denote $r_{2,i} = ac_x\tilde{f}_2 v_i + r'_{2,i}$ and $f_2 = c_x\tilde{f}_2$.

K_0 is generated by $K_0 = Q''R''\prod_{i=1}^{t} h_{1,i}^{-r_{1,i}} h_{2,i}^{-r_{2,i}}$
$= QR\prod_{i=1}^{t} g_p^{-(bx_i+\omega_{1,i})(a\tilde{f}_1 v_i - abc_y\tilde{f}_2 v_i + r'_{1,i})} g_p^{-(b^2 y_i + \omega_{2,i})(ac_x\tilde{f}_2 v_i + r'_{2,i})}$
$= QR\prod_{i=1}^{t} g_p^{-(ab\tilde{f}_1 x_i v_i - ab^2 c_y\tilde{f}_2 x_i v_i + br'_{1,i} x_i + a\tilde{f}_1\omega_{1,i} v_i - abc_y\tilde{f}_2\omega_{1,i} v_i + r'_{1,i}\omega_{1,i} + ab^2 c_x\tilde{f}_2 y_i v_i + b^2 r'_{2,i} y_i + ac_x\tilde{f}_2 v_i\omega_{2,i} + r'_{2,i}\omega_{2,i})}$
$= QR\prod_{i=1}^{t} g_p^{-a(\tilde{f}_1\omega_{1,i} v_i + c_x\tilde{f}_2 v_i\omega_{2,i})} g_p^{-ab(\tilde{f}_1 x_i v_i - c_y\tilde{f}_2\omega_{1,i} v_i)} g_p^{-b(r'_{1,i} x_i)} g_p^{-b^2(r'_{2,i} y_i)} g_p^{-(r'_{1,i}\omega_{1,i} + r'_{2,i}\omega_{2,i})}$
$= QR\prod_{i=1}^{t} (g_p^a g_q)^{-(\tilde{f}_1\omega_{1,i} v_i + c_x\tilde{f}_2 v_i\omega_{2,i})} (g_p^{ab} Q_1)^{-(\tilde{f}_1 x_i v_i - c_y\tilde{f}_2\omega_{1,i} v_i)} h_p^{-(r'_{1,i} x_i)} k_p^{-(r'_{2,i} y_i)} g_p^{-(r'_{1,i}\omega_{1,i} + r'_{2,i}\omega_{2,i})}$

According to the symmetric property of scheme Π_2, $(X, Y) \stackrel{c}{\equiv} (0, Y)$ and $(0, Y) \stackrel{c}{\equiv} (Y, Y)$ can be proved similarly.

D Proof of Theorem 2

Proof. We first assume that the above encryption scheme Π is not selective-ID secure in the MR model. That is there exists an PPT Adversary $\mathcal{A}_1$ with non-negligible advantage ϵ in the game of Definition 2. Now we construct a PPT adversary $\mathcal{A}_2$ which acts as Challenger interacting with $\mathcal{A}_1$ and show that it can win the game of Definition 1 also with non-negligible advantage with the scheme $\Pi_1 = (\mathsf{Setup}', \mathsf{Encrypt}', \mathsf{GenTK}', \mathsf{Test}')$.
Setup: The adversary $\mathcal{A}_2(1^n)$ runs adversary $\mathcal{A}_1(1^n)$. Adversary $\mathcal{A}_1$ outputs two possible equal-length vector X_0 and X_1 to adversary $\mathcal{A}_2$.

Adversary $\mathcal{A}_2$ passes X_0 and X_1 to the challenger $\mathcal{C}$.

The challenger $\mathcal{C}$ takes a security parameter n and runs `Setup` to generate pk and sk; C sends pk to adversary $\mathcal{A}_2$. $\mathcal{A}_2$ passes pk to $\mathcal{A}_1$.
Challenge: The challenger $\mathcal{C}$ picks a random bit $b \in \{0, 1\}$, computes and returns $C^* \stackrel{\$}{\leftarrow} \mathtt{Encrypt}(pk, X_b)$ to adversary $\mathcal{A}_2$. Adversary $\mathcal{A}_2$ passes it to adversary $\mathcal{A}_1$.
Phase 1: The adversary $\mathcal{A}_1$ adaptively requests polynomially bounded tokens for any (V_i, t_i) subject to the restriction that $t_i < HammingDist(V_i, X_i)$ for both $j = 0, 1$. When receiving the request (V_i, t_i), adversary $\mathcal{A}_2$ generates t_i token requests to the challenger $\mathcal{C}$ that $(V_i, 0), ..., (V_i, t_i)$ and receives token $TK_0, ..., TK_{t_i}$. Adversary $\mathcal{A}_2$ answers adversary $\mathcal{A}_1$ with $(TK_0, ..., TK_{t_i})$.

Guess: Adversary $\mathcal{A}_1$ returns with the output bit b'. The adversary $\mathcal{A}_2$ passes b' as its output to the challenger $\mathcal{C}$.

Since we have the restriction in MR definition that $t_i < HammingDist(V_i, X_i)$ for both $j = 0, 1$, $j \neq HammingDist(V_i, X_i)$ for each $j \in \{0, ..., t_i\}$. And therefore, $\mathcal{A}_2$'s token requests satisfy the requirement in MC definition that $t_i = HammingDist(V_i, X_j)$ for both $j = 0, 1$ or $t_i \neq HammingDist(V_i, X_j)$ for both $j = 0, 1$.

The advantage $\mathcal{A}_2$ in the above MC game $|\Pr[b' = b] - \frac{1}{2}| = \mathbf{Adv}^{\mathsf{MR}}_{\Pi,\mathcal{A}_1}(n) = \epsilon$ which is non-negligible in security parameter n. However, according to our security analysis in Section 3, the non-negligible advantage is impossible. This completes our security proof.

E Secure t_{max} update

To support t_{max} update, we need do some modifications on the scheme in Section 4.1. The new encryption scheme allowing t_{max} update consists of five PPT algorithms $\Pi = (\texttt{Setup}, \texttt{Encrypt}, \texttt{UpdateCipher}, \texttt{GenTK}, \texttt{Test})$:

- $\texttt{Setup}(1^n)$: It first randomly selects $\{h_{1,l,i}, h_{2,l,i}\}$ from $\mathbb{G}_p$ where $l \in [1, m+1]$ and $i \in [1, \binom{m}{i}]$. (So, the total number of terms is 2^m) Then it randomly selects h_3, h_4 from $\mathbb{G}_p$. It also randomly selects $R, \{R_{1,l,i}, R_{2,l,i}\}_{l\in[1,m+1], i\in[1,\binom{m}{i}]}, R_3, R_4$ from $\mathbb{G}_r$. It outputs

$$pk = \{g_p, g_r, Q = g_q R, \\ [H_{1,l,i} = h_{1,l,i}R_{1,l,i}, H_{2,l,i} = h_{2,l,i}R_{2,l,i}]_{l\in[1,m+1], i\in[1,\binom{m}{i}]}, \\ H_3 = h_3R_3, H_4 = h_4R_4\}$$

  and

$$sk = \{p, q, r, g_q, [h_{1,l,i}, h_{2,l,i}]_{l\in[1,m+1], i\in[1,\binom{m}{l}]}, h_3, h_4\}$$

- $\texttt{Encrypt}(pk, X)$: Rather than outputting ciphertext C, it also outputs state $S = \{\alpha, \beta, s\}$ used.
- $\texttt{GenTK}(pk, sk, V)$: The same as the original algorithm.
- $\texttt{UpdateCipher}(pk, T', X, C, S)$: It randomly selects $\{R'_{1,l,i}, R'_{2,l,i}\}_{l\in[t_{max}+2, T'+1], i\in[1,\binom{m}{i}]}$ from $\mathbb{G}_r$. Then, it outputs δ:

$$\{[C_{1,l,i} = H^s_{1,l,i} Q^{\alpha x_{j_1} \cdots x_{j_l}} R'_{1,l,i}, C_{1,l,i} = \\ H^s_{2,l,i} Q^{\beta x_{j_1} \dots x_{j_l}} R'_{2,l,i}]_{\{l\in\{t_{max}+2, T'+1\}, 1\leq j_1<j_2<\ldots<j_l\leq m\}}\}$$

- $\texttt{Test}(pk, sk, TK, C)$ The same as the original algorithm.

Then, we also need to define a proper security definition concerning t_{max} updates.

Definition 5. (Selective-ID secure in match-concealing mode with t_{max} update capability) *The encryption scheme* $\Pi = (\texttt{Setup}, \texttt{Encrypt}, \texttt{UpdateCipher}, \texttt{GenTK}, \texttt{Test})$ *is MC secure if for all probabilistic polynomial-time Turing machine (adversary)* $\mathcal{A}$*, the advantage of* $\mathcal{A}$ *in the following game is negligible.*

Setup: The adversary $\mathcal{A}(1^n)$ outputs two possible equal-length (m-length where $2^m = poly(n)$) vectors X_0 and X_1 to the challenger $\mathcal{C}$. The challenger $\mathcal{C}$ takes a security parameter n and runs `Setup` to generate pk and sk. Adversary $\mathcal{A}$ is given pk.

Challenge: The challenger $\mathcal{C}$ picks a random bit $b \in \{0,1\}$, computes $(C^*, S^*) \xleftarrow{\$}$ $\texttt{Encrypt}(pk, X_b)$. $\mathcal{C}$ returns C^* to adversary $\mathcal{A}$.
Phase 1: The adversary $\mathcal{A}$ may adaptively request polynomially bounded number of queries. The types of queries allowed are described as below:

- `GenTK`: Adversary $\mathcal{A}$ can request $\mathcal{C}$ to compute and return tokens of any $(V_i = v_{i,1}...v_{i,m}, t_i)$ where $t_i \leq t_{max}$ and V_i subject to the restriction that $t_i < HammingDist(V_i, X_j)$ for both $j = 0$ and 1, or $t_i \geq HammingDist(V_i, X_j)$ for both $j = 0$ and 1.
- `UpdateCipher`: Adversary $\mathcal{A}$ outputs T' ($t_{max} < T' \leq m$) to the challenger $\mathcal{C}$. The challenger $\mathcal{C}$ computes and returns $\delta \xleftarrow{\$} UpdateCipher(pk, T', X_b, C^*, S^*)$ to adversary $\mathcal{A}$. The challenger also records T' as the new t_{max}.

Guess: Adversary $\mathcal{A}$ outputs a guess bit b'. The advantage of $\mathcal{A}$ is defined as $|\Pr[b' = b] - \frac{1}{2}|$.

Theorem 3. *The encryption scheme* $\Pi_{\mathsf{H}} = (\texttt{Setup}, \texttt{Encrypt}, \texttt{UpdateCipher}, \texttt{GenTK}, \texttt{Test})$ *defined here is Selective-ID secure in match-concealing mode with* t_{max} *update capability.*

Proof. The above theorem is proved in two steps. We first modify the inner-product encryption defined in Section 4.1 to a scheme supporting ciphertext updates. Similar wto the method in Lemma 2, we prove in Lemma 4 that if this modified inner-product is Selective-ID secure in match-concealing mode with t_{max} update capability, then, the encryption scheme $\Pi = (\texttt{Setup}, \texttt{Encrypt}, \texttt{UpdateCipher}, \texttt{GenTK}, \texttt{Test})$ defined here is also Selective-ID secure in match-concealing mode with t_{max} update capability.

The second step is that we prove this modified inner-product is Selective-ID secure in match-concealing mode with t_{max} update if inner-product encryption defined in Section 4.1 is MC secure (Definition 4). It is proved by Lemma 5.

To prove the first stage, we first describe the modified inner-product encryption scheme which is also consists of five PPT algorithms
$\Pi = (\texttt{Setup}, \texttt{Encrypt}, \texttt{UpdateCipher}, \texttt{GenTK}, \texttt{Test})$:

- $\texttt{Setup}(1^n)$. It first randomly selects $\{h_{1,i}, h_{2,i}\}_{i \in [1, L_{max}]}$ from $\mathbb{G}_p$, and then randomly selects $R, \{R_{1,i}, R_{2,i}\}_{i \in [1, L_{max}]}$ from $\mathbb{G}_r$. It outputs L_{max} along with public key pk and secret key sk:

$$pk = \{g_p, g_r, Q = g_q R, [H_{1,i} = h_{1,i} R_{1,i}, H_{2,i} = h_{2,i} R_{2,i}]_{i \in [1, L_{max}]}\}$$

and

$$sk = \{p, q, r, g_q, [h_{1,i}, h_{2,i}]_{i \in [1, L_{max}]}\}$$

– `Encrypt`$(pk, X = x_1...x_i...x_{l_{max}})$: Rather than outputting ciphertext C, it also outputs state $S = \{\alpha, \beta, s\}$ used.
– `UpdateCipher`$(pk, X' = x'_1...x'_{|X'|}, X, C, S)$: X' is decided by $|X'|$, l_{max} and X. It randomly selects $\{R'_{1,i}, R'_{2,i}\}_{i\in[1,|X'|]}$ from $\mathbb{G}_r$. Then, it outputs

$$\delta = \{C_{1,i+t_{max}} = H^s_{1,i+t_{max}} Q^{\alpha x'_i} R'_{1,i}, C_{2,i+t_{max}} = H^s_{2,i+t_{max}} Q^{\beta x'_i} R'_{2,i}\}_{i\in[1,|X'|]}$$

– `GenTK`$(pk, sk, V = v_1...v_i...v_t)$: The same as the original algorithm.
– `Test`(pk, TK, C): The same as the original algorithm.

We define its security definition as follows.

Definition 6. (Selective-ID secure in Match Concealing mode with t_{max} update capability)

Setup: The adversary $\mathcal{A}(1^n)$ outputs two possible equal-length (l_{max}-length) vectors X_0 and X_1. The challenger $\mathcal{C}$ takes a security parameter n and runs `Setup` to generate pk and sk. $\mathcal{C}$ sends pk to $\mathcal{A}$. `Setup` also outputs $L_{max} = poly(n)$.

Challenge: The challenger picks a random bit $b \in \{0, 1\}$, computes $(C^*, S^*) \xleftarrow{\$}$ `Encrypt`(pk, X_b). C^* is given to adversary $\mathcal{A}$.
Phase 1: The adversary $\mathcal{A}$ may adaptively request polynomially bounded number of queries. The types of queries allowed are described as below:

– `GenTK`: Adversary $\mathcal{A}$ may adaptively request challenger $\mathcal{C}$ of tokens for any $V_i = v_{i,1}v_{i,2}...v_{i,t_i}$ where $t_i \leq l_{max}$ and V_i subject to the restriction that $\sum_{k=1}^{t_i} x_{j,k}v_{i,k} = 0$ for both $j = 0$ and 1 or $\sum_{k=1}^{t_i} x_{j,k}v_{i,k} \neq 0$ for both $j = 0$ and 1.
– `UpdateCipher`: Adversary $\mathcal{A}$ outputs two equal-length (L'-length) vectors X'_0 and X'_1 where X'_0 is determined by L', l_{max} and X_0 ; X'_1 is determined by L', l_{max} and X_1. And $l_{max} + L' <= L_{max}$ where T_{max} is decided at `Setup` time. Adversary $\mathcal{A}$ sends X'_0 and X'_1 to challenger $\mathcal{C}$. The challenger $\mathcal{C}$ computes and returns $\delta \xleftarrow{\$} UpdateCipher(pk, X'_b, X_b, C^*, S^*)$ to adversary $\mathcal{A}$. $\mathcal{C}$ also record $l_{max} + L'$ as the new l_{max}.

Guess: The adversary $\mathcal{A}$ outputs a bit b' to guess b. The advantage of $\mathcal{A}$ is defined as $|\Pr[b' = b] - \frac{1}{2}|$.

Lemma 4. *The Anonymous Fuzzy Identity-Based Encryption scheme allowing t_{max} updates is secure under Definition 5 if Inner-product Encryption scheme allowing t_{max} updates is secure under Definition 6.*

Proof. We first assume that there exists an adversary $\mathcal{A}_1$ who wins Definition 5 with non-negligible advantage. Then, we try to construct an adversary $\mathcal{A}_2$ who can win Definition 6 also with non-negligible.
Setup: Adversary $\mathcal{A}_2(1^n)$ runs $\mathcal{A}_1(1^n)$. $\mathcal{A}_1(1^n)$ outputs two possible equal-length (m-length) vector X_0 and X_1 to $\mathcal{A}_2$.

A_2 calculates $\sum_{i=0}^{t_{max}+1}\binom{m}{i}$-length vectors $\tilde{X}_0$ and $\tilde{X}_1$ to challenger $\mathcal{C}$. The Challenger $\mathcal{C}$ takes a security parameter n and runs `Setup` to generate pk and sk. Set $L_{max} = 2^m$. $\mathcal{C}$ sends pk to adversary $\mathcal{A}_2$. $\mathcal{A}_2$ rearrange pk and send it to $\mathcal{A}_1$.
Challenge: The challenger $\mathcal{C}$ picks a random bit $b \in \{0,\}$, computes $(C^*, S^*) \xleftarrow{\$}$ `Encrypt`(pk, X_b). $\mathcal{C}$ sends C^* to adversary $\mathcal{A}_2$. $\mathcal{A}_2$ rearranges C^* and sends to it $\mathcal{A}_1$.
Phase 1: Adversary $\mathcal{A}_1$ may adaptively request polynomially bounded `GenTK` and `UpdateCipher` queries.

– `GenTK`: The adversary $\mathcal{A}_1$ requests tokens to $\mathcal{A}_2$ for any $(V_i = v_{i,1}...v_{i,m}, t_i)$ where $t_i \le t_{max}$ and V_i subject to the restriction that $t_i < HammingDist(V_i, X_j)$ for both $j = 0$ and 1 or $HammingDist(V_i, X_j) \le t_i$ for both $j = 0$ and 1.
When receiving (V, t), adversary A_2 calculates $\tilde{V} = \tilde{v_1}, ..., \tilde{v}_{\sum_{i=0}^{t+1}\binom{m}{i}}$ such that $\sum \tilde{v_i}\tilde{x_{j,i}} = 1$ if and only if $HammingDist(V, X_j) \le t$. $\mathcal{A}_2$ submits $\tilde{V}$ to challenger $\mathcal{C}$ and get token.
$\mathcal{A}_2$ rearranges the token and sends it to $\mathcal{A}_1$.
– `UpdateCipher`: The adversary $\mathcal{A}_1$ outputs $m \ge T' > t_{max}$ to $\mathcal{A}_2$. $\mathcal{A}_2$ based on X_0 and X_1 and T' to calculate two equal-length ($L' = \sum_{i=t_{max}+2}^{T'+1}\binom{m}{i}$)-length) vectors X'_0 and X'_1 (base on the inner-product formula in Section 4.1). Adversary $\mathcal{A}_2$ passes X'_0 and X'_1 to challenger $\mathcal{C}$.
The challenger $\mathcal{C}$ returns $\delta \xleftarrow{\$} UpdateCipher(pk, X'_b, X_b, C^*, S^*)$ to adversary $\mathcal{A}_2$. Challenger $\mathcal{C}$ also updates l_{max} as $L' + l_{max}$. $\mathcal{A}_2$ rearranges δ and sends to $\mathcal{A}_1$.

Guess: The adversary $\mathcal{A}_1$ outputs a bit b' to guess b. $\mathcal{A}_2$ passes b' to challenger $\mathcal{C}$ as its output.

Then, we prove Lemma 5 to complete the proof for Theorem 3.

Lemma 5. *The scheme* $\Pi = (\mathsf{Setup}, \mathsf{Encrypt}, \mathsf{UpdateCipher}, \mathsf{GenTK}, \mathsf{Test})$ *is secure under Definition 6 if* $\Pi = (\mathsf{Setup}, \mathsf{Encrypt}, \mathsf{GenTK}, \mathsf{Test})$ *is secure under Definition 4.*

Proof. This proof is by contradiction. We first assume that there exists an adversary $\mathcal{A}_1$ who wins Selective-ID game in match-concealing mode with t_{max} update capability with non-negligible advantage ϵ, then we can construct an adversary $\mathcal{A}_2$ who wins Definition 4 also with non-negligible advantage ϵ.
Setup: The adversary $\mathcal{A}_2(1^n)$ runs $\mathcal{A}_1(1^n)$. $\mathcal{A}_1(1^n)$ outputs two possible equal-length (l_{max}-length) vectors X_0 and X_1 to $\mathcal{A}_2$. Note that $l_{max} \le L_{max}$.

The adversary $\mathcal{A}_2$ calculates and outputs two L_{max}-length vectors $X_0^{(max)}$ and $X_1^{(max)}$ to challenger $\mathcal{C}$ where $X_0^{(max)}$ is decided by X_0 and L_{max}; $X_1^{(max)}$ is decided by X_1 and L_{max}.

The challenger $\mathcal{C}$ takes a security parameter n and runs `Setup` to generate pk and sk (for L_{max}-length). $\mathcal{C}$ returns pk to adversary $\mathcal{A}_2(1^n)$. The adversary $\mathcal{A}_2$ passes pk to $\mathcal{A}_1$.

Challenge: The challenger $\mathcal{C}$ picks a random bit $b \in \{0,1\}$, computes and returns $C^* \xleftarrow{\$} \texttt{Encrypt}(pk, X_b^{(max)})$ to adversary $\mathcal{A}_2$.

The adversary $\mathcal{A}_2$ takes the first l_{max} components of ciphertext C^* and returns it (denoted as $C^*_{l_{max}}$) to $\mathcal{A}_1$.

Phase 1: The adversary $\mathcal{A}_1$ may adaptively request polynomially-bounded number of `GenTK` and `UpdateCipher` queries.

- `GenTK`: The adversary $\mathcal{A}_1$ may adaptively request tokens for any $V_i = v_{i,1}v_{i,2}...v_{i,t_i}$ where $t_i \leq l_{max}$ and V_i subject to the restriction that $\sum_{k=1} t_i x_{j,k} v_{i,k} = 0$ for both $j = 0$ and 1 or $\sum_{k=1} t_i x_{j,k} v_{i,k} \neq 0$ for both $j = 0$ and 1. $\mathcal{A}_1$ sends these requests to $\mathcal{A}_2$.
  The adversary $\mathcal{A}_2$ passes these requests to the challenger $\mathcal{C}$. Note that $t_i \leq l_{max} \leq L_{max}$. The challenger generates tokens and adversary $\mathcal{A}_2$ passes them to $\mathcal{A}_1$.
- `UpdateCipher`: The adversary $\mathcal{A}_1$ outputs two equal-length (L'-length) vectors X_0' and X_1' where X_0' is determined by L',l_{max} and X_0 ; X_1' is determined by L',l_{max} and X_1. And $l_{max} + L' \leq L_{max}$. The adversary $\mathcal{A}_1$ passes X_0' and X_1' to $\mathcal{A}_2$.
  The adversary $\mathcal{A}_2$ passes $\delta = C^*[l_{max}+1, l_{max}+L']$ to $\mathcal{A}_1$. The adversary $\mathcal{A}_2$ also record $l_{max} + L'$ as the new l_{max}.

Guess: The adversary $\mathcal{A}_1$ outputs a bit b' to $\mathcal{A}_2$. And $\mathcal{A}_2$ passes b' to the challenger $\mathcal{C}$ as its output. Note that $|\Pr[b' = b] - \frac{1}{2}| = \epsilon$. This completes our proof.