

A Capacity-Achieving T -PIR Scheme Based On MDS Array Codes

Jingke Xu, Yaqian Zhang, Zhifang Zhang

KLMM, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
 School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
 Emails: xujingke14@mails.ucas.edu.cn, zhangyaqian15@mails.ucas.ac.cn, zfz@amss.ac.cn

Abstract—Suppose a database containing M records is replicated in each of N servers, and a user wants to privately retrieve one record by accessing the servers such that identity of the retrieved record is secret against any up to T servers. A scheme designed for this purpose is called a T -private information retrieval (T -PIR) scheme.

In this paper we focus on the field size of T -PIR schemes. We design a general capacity-achieving T -PIR scheme whose queries are generated by using some MDS array codes. It only requires field size $q \geq \sqrt[t]{N}$, where $\ell = \min\{t^{M-2}, (n-t)^{M-2}\}$, $t = T/\gcd(N, T)$, $n = N/\gcd(N, T)$ and has the optimal sub-packetization Nn^{M-2} . Comparing with existing capacity-achieving T -PIR schemes, our scheme has the following advantage, that is, its field size monotonically decreases as the number of records M grows. In particular, the binary field is sufficient for building a capacity-achieving T -PIR scheme as long as $M \geq 2 + \lceil \log_\mu \log_2 N \rceil$, where $\mu = \min\{t, n-t\} > 1$.

I. INTRODUCTION

Private information retrieval (PIR) is a canonical problem in the study of privacy issues that arise from the retrieval of information from public databases. Typically, a PIR model involves a database containing M records stored across N servers and a user who wants to privately retrieve one record by accessing the servers. Specifically, the privacy requirement means any colluding subset containing no more than T servers knows nothing about identity of the retrieved record. Since it is closely related to cryptography [1] and coding theory [2], PIR has become a central research topic in the computer science literature since it was first introduced by Chor et al. [3] in 1995.

The efficiency of PIR scheme is characterized by its *rate*. Specifically, the *rate* of a PIR scheme is measured as the ratio between the retrieved data size and the downloaded size, and the *capacity* is defined as the supremum of the rate over all PIR schemes. Recently, much work has been done on determining the capacity of PIR in various cases. Sun and Jafar derived that the capacity for the non-colluding servers (i.e., $T = 1$) is $\frac{1-1/N}{1-(1/N)^M}$ in [4] and further proved that the capacity for the colluding servers (i.e., $T > 1$) is $\frac{1-T/N}{1-(T/N)^M}$ in [5]. Moreover, the latter is called T -PIR. They also determined the capacity of PIR with symmetric privacy in [6]. The capacity of PIR with MDS coded non-colluding servers is determined in [7]. It remains an open problem to determine the capacity of PIR with MDS coded colluding servers. For non-MDS coded

storage, PIR schemes with colluding or non-colluding servers are presented in [9], [10].

In general, the capacity of PIR is achieved by dividing each record into multiple sub-packets and querying from each server specially designed combinations of these sub-packets. Therefore, both the number of sub-packets and the size of each sub-packet are important metrics for measuring the implementation complexity of a PIR scheme. As to the former, we call the number of sub-packets contained in each record as sub-packetization. The optimal sub-packetization for capacity-achieving PIR schemes has been determined in some cases [8], [12], [13]. As to the latter, since all existing PIR schemes are linear schemes over some finite fields, it is actually about the size of the field on which the PIR scheme can be built. The main concern of this work is to reduce the field size for T -PIR schemes while maintaining the rate achieving the capacity and the optimal sub-packetization.

In [5], it requires a field of size $q = \Omega(N^2 T^{M-2})$ for the capacity-achieving T -PIR scheme. The field size is reduced to $q = \Omega(N t^{M-2})$ for the capacity-achieving T -PIR scheme with optimal sub-packetization in [12], where $t = T/\gcd(N, T)$. The best known result of field size for capacity-achieving T -PIR scheme is $q = \Omega(N)$ in [14]. But the field size is still unfriendly with the growth of the number of servers.

The main contribution of this work consists of designing a T -PIR scheme that simultaneously achieves the capacity and the optimal sub-packetization Nn^{M-2} over a finite field \mathbb{F}_q for all possible parameters (N, T, M) , and it requires the field size $q \geq \sqrt[t]{N}$, where $\ell = \min\{t^{M-2}, (n-t)^{M-2}\}$, $n = N/\gcd(N, T)$, $t = T/\gcd(N, T)$. When $\ell = 1$, the constraint of the field size in our scheme degenerates into $q \geq N$, which is the same with that of the capacity-achieving T -PIR scheme in [14]. When $\ell > 1$, the binary field is sufficient for building a capacity-achieving T -PIR scheme provided $M \geq 2 + \lceil \log_\mu \log_2 N \rceil$, where $\mu = \min\{t, n-t\}$.

Reference	Field size(q)
Sun et al. [5]	$q \geq \max\{N^2 T^{M-2}, N^2 (N-T)^{M-2}\}$
Zhang et al. [11]	$q \geq \max\{N n t^{M-2}, N n (n-t)^{M-2}\}$
Zhang et al. [12]	$q \geq \max\{N t^{M-2}, N (n-t)^{M-2}\}$
Xu et al. [14]	$q \geq N$
This paper	$q \geq \sqrt[t]{N}, \ell = \min\{t^{M-2}, (n-t)^{M-2}\}$

Table 1: A list of all existing capacity-achieving T -PIR schemes with $T \geq 2$. And $n = \frac{N}{\gcd(N, T)}$, $t = \frac{T}{\gcd(N, T)}$.

Comparing with all existing capacity-achieving T -PIR schemes with $T \geq 2$ in [5], [11], [12], [14], as displayed in Table 1, the main difference in our scheme is to employ MDS array codes to generate queries, which is a key idea for reducing the field size. Moreover, an advantage of our scheme is that its field size monotonically decreases as the number of records M grows.

The rest of this paper is organized as follows. First, the T -PIR model is formally introduced and the MDS array code is defined in Section II. Then in Section III an example of the T -PIR scheme is presented to explain the design idea. The recovery property of MDS array codes is proved and the general descriptions of our scheme are given in Section IV. Finally, Section V concludes the paper.

II. PRELIMINARIES

A. Notations and the T -PIR model

For an integer $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, \dots, n\}$. For a vector $\mathbf{u} = (u_1, \dots, u_n)$ and a subset $\Gamma = \{i_1, \dots, i_m\} \subseteq [n]$, denote $\mathbf{u}_\Gamma = (u_{i_1}, \dots, u_{i_m})$. Most vectors in this paper are row vectors and they are denoted by the bold lowercase letters (eg. \mathbf{a}, \mathbf{b}). For a block matrix $A = (A^{(1)}, A^{(2)}, \dots, A^{(N)})$ and $\Gamma = \{i_1, \dots, i_m\} \subseteq [N]$, denote $A^\Gamma = (A^{(i_1)}, \dots, A^{(i_m)})$.

Suppose there are M records W_1, \dots, W_M and N servers $\text{Serv}^{(1)}, \dots, \text{Serv}^{(N)}$, each server stores all the M records. Moreover, the records are independent and each can be seen as an L -length vector over \mathbb{F}_q . Then suppose a user wants to privately retrieve W_θ for some $\theta \in [M]$. Formally, a T -PIR scheme consists of two phases:

- **Query phase.** Given $\theta \in [M]$, the user generates the query $\text{Que}(\theta, S) = (Q_\theta^{(1)}, \dots, Q_\theta^{(N)})$, and sends $Q_\theta^{(j)}$ to $\text{Serv}^{(j)}$ for $1 \leq j \leq N$, where S are some random resources privately chosen by the user. Note that $\text{Que}(\cdot, \cdot)$ is the *query function* defined by the scheme.
- **Response phase.** After receiving $Q_\theta^{(j)}$, the $\text{Serv}^{(j)}$ computes the answers $\text{Ans}^{(j)}(Q_\theta^{(j)}, W_{[M]}) = A_\theta^{(j)}$ for $1 \leq j \leq N$, and sends it back to the user, where $\text{Ans}^{(j)}(\cdot, \cdot)$ is the *answer function* defined by the scheme.

Moreover the functions $\text{Que}(\cdot, \cdot)$ and $\text{Ans}^{(j)}(\cdot, \cdot)$, $1 \leq j \leq N$ must satisfy the following two conditions:

- (1) **Correctness:** The user can reconstruct W_θ after collecting all answers from the N servers, i.e., $H(W_\theta | A_\theta^{[N]}, Q_\theta^{[N]}, S, \theta) = 0$, where $H(\cdot | \cdot)$ is the conditional entropy.
- (2) **Privacy:** For any $\Gamma \subseteq [N]$ with $|\Gamma| = T$, the servers in Γ can't obtain any information on θ even if they collude with each other, i.e., $I(\theta; Q_\theta^\Gamma, A_\theta^\Gamma, W_{[M]}) = 0$, where $I(\cdot; \cdot)$ denotes the mutual information.

Define the *rate* \mathcal{R} of a T -PIR scheme by

$$\mathcal{R} = \frac{H(W_\theta)}{\sum_{i=1}^N H(A_\theta^{(i)})},$$

that is, \mathcal{R} characterizes the amount of retrieved information per unit of downloaded data. Furthermore, the *capacity* of

T -PIR is defined by the largest rate over all achievable T -PIR schemes, denoted by $\mathcal{C}_{T\text{-PIR}}$. By [5], it has that $\mathcal{C}_{T\text{-PIR}} = \frac{1-T/N}{1-(T/N)^M}$.

B. MDS Array Codes

In this section we introduce MDS array code used in this paper and then give a method to construct such code over \mathbb{F}_q .

Suppose $N > T \geq 1$ and N, T are two positive integers. For a linear $[N, T]$ code \mathcal{C} over \mathbb{F}_{q^ℓ} , a codeword $\mathbf{c} = (c_1, c_2, \dots, c_N)$ can be seen as an $N\ell$ -length vector $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_N)$ over \mathbb{F}_q , i.e., for $i \in [N]$, the code block $\mathbf{c}_i = (c_{i,1}, c_{i,2}, \dots, c_{i,\ell}) \in \mathbb{F}_q^\ell$ denotes the ℓ -length vector corresponding to the symbol $c_i \in \mathbb{F}_{q^\ell}$. So we call the code \mathcal{C} a linear array code over \mathbb{F}_q , and refer to the code as an $(N, T; \ell)_q$ linear array code. Equivalently, an $(N, T; \ell)_q$ linear array code can be defined by a $T\ell \times N\ell$ full rank matrix G over \mathbb{F}_q as follows,

$$\mathcal{C} = \{\mathbf{c} = (\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_T)G : (\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_T) \in \mathbb{F}_q^{T\ell}\}.$$

The matrix G is called a generator matrix of the array code \mathcal{C} . Then the generator matrix G can be viewed as a block matrix

$$G = (G^{(1)}, G^{(2)}, \dots, G^{(N)}).$$

For $i \in [N]$, the $T\ell \times \ell$ sub-matrix $G^{(i)}$ is represented as the thick column associated with the i th code block in the codewords of \mathcal{C} .

Definition 1. (MDS Array Codes) A linear array code \mathcal{C} over \mathbb{F}_q is called an $(N, T; \ell)$ MDS array code if its generator matrix $G = (G^{(1)}, G^{(2)}, \dots, G^{(N)}) \in \mathbb{F}_q^{T\ell \times N\ell}$ has the following MDS property:

$$\forall \Gamma \subseteq [N] \text{ with } |\Gamma| = T, \text{rank}(G^\Gamma) = T\ell \quad (1)$$

where $G^{(i)} \in \mathbb{F}_q^{T\ell \times \ell}$ for $i \in [N]$ and $T < N$.

By the definition of $(N, T; \ell)$ MDS array code \mathcal{C} , it degenerates into a MDS code over \mathbb{F}_q for $\ell = 1$. Next we give a method to construct an $(N, T; \ell)$ MDS array code.

Suppose α is a primitive element of \mathbb{F}_{q^ℓ} , then $\mathbb{F}_{q^\ell} = \{\alpha^j : 0 \leq j \leq q^\ell - 2\} \cup \{0\}$. Suppose $m(x)$ is the minimal polynomial of α over \mathbb{F}_q . Let $C \in \mathbb{F}_q^{\ell \times \ell}$ be the companion matrix of $m(x)$ and $\mathbb{F} = \{C^j : j \in \mathbb{Z}\} \cup \{0\}$. Then \mathbb{F} is a finite field of size q^ℓ and the map which is defined by $\varphi(\alpha^j) = C^j$ and $\varphi(0) = \mathbf{0}$ is a field isomorphism from \mathbb{F}_{q^ℓ} to \mathbb{F} by [15]. Let $G = (\alpha_{i,j}) \in \mathbb{F}_{q^\ell}^{T \times N}$ be a generator matrix of an $[N, T]$ MDS code over \mathbb{F}_{q^ℓ} . Note that each symbol of \mathbb{F}_{q^ℓ} can be represented as an $\ell \times \ell$ matrix in \mathbb{F} over \mathbb{F}_q by using the field isomorphism $\varphi(\cdot)$, then the matrix G can be seen as an $T \times N$ block matrix G' , i.e., $G' = (\varphi(\alpha_{i,j}))_{i \in [T], j \in [N]}$, and each thick column is an $T\ell \times \ell$ matrix over \mathbb{F}_q . It is easy to verify that for any $\Gamma \subseteq [N]$ with $|\Gamma| = T$,

$$\det((\varphi(\alpha_{i,j}))_{i \in [T], j \in \Gamma}) = \varphi(\det((\alpha_{i,j})_{i \in [T], j \in \Gamma})) \neq 0.$$

Hence the linear array code which is defined by the generator matrix G' over \mathbb{F}_q is an $(N, T; \ell)$ MDS array code. Then we can directly obtain the following theorem.

Theorem 2. Suppose q is a power of a prime and $N, T, \ell \in \mathbb{N}$ with $N > T \geq 2$. If $q^\ell \geq N$, then there exists an $(N, T; \ell)$ MDS array code over \mathbb{F}_q .

Recall that for existing capacity-achieving T -PIR schemes in [5], [11], [12], some $[N\ell_k, T\ell_k]$ MDS codes over \mathbb{F}_q are used to construct the query. And all $N\ell_k$ symbols of each codeword are equally divided into N blocks. Then this MDS code can be seen as an $(N, T; \ell_k)$ MDS array code over \mathbb{F}_q . Based on this observation, we find a direction to reduce the field size. That is, we generate the query by using some MDS array codes over a smaller finite field rather than MDS codes. Moreover, the MDS array codes need to satisfy some special property that is determined by the correctness condition. To formally illustrate this idea, we will give an example in the next section.

III. EXAMPLE FOR $N = 5, T = 3, M = 3$

Before constructing our schemes, we first give an example by using the method described in [12]. And then, we explain how to reduce the field size by modifying this scheme.

Example 1. Suppose $M = 3, N = 5$ and $T = 3$. The field size $q \geq 15$ is enough and the sub-packetization of this case is $L = N^{M-1} = 25$, so each record can be seen as a 25-dimensional vector over \mathbb{F}_q , i.e., $W_1, W_2, W_3 \in \mathbb{F}_q^{25}$. WLOG, suppose W_1 is the desired record, i.e., $\theta = 1$.

Let S_1, S_2, S_3 be three matrices chosen by the user independently and uniformly from all 25×25 invertible matrices over \mathbb{F}_q . Actually, S_1, S_2 and S_3 are the random resources privately held by the user. Then, define

$$\begin{aligned} (a_1, a_2, \dots, a_{25}) &= W_1 S_1 \\ (b_1, b_2, \dots, b_{25}) &= W_2 S_2[:, (1 : 15)] G \\ (c_1, c_2, \dots, c_{25}) &= W_3 S_3[:, (1 : 15)] G \end{aligned} \quad (2)$$

where $S_i[:, (1 : 15)]$ denotes the 25×15 matrix formed by the first 15 columns of S_i and $G = \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix}$. Moreover, $G_1 \in \mathbb{F}_q^{9 \times 15}$ is a generator matrix of an $[15, 9]$ MDS code and $G_2 \in \mathbb{F}_q^{6 \times 10}$ is a generator matrix of an $[10, 6]$ MDS code over \mathbb{F}_q .

It can be seen that the answers are all sums of the symbols a_i, b_i, c_i in Fig.1. For each sum x in Fig.1, we define its support as a subset of $[3]$ and this subset is composed of the label of all terms in the sum x , denoted by $\text{supp}(x)$. For example, $\text{supp}(a_i) = \{1\}, \text{supp}(b_i + c_j) = \{2, 3\}$. For any $\Lambda \subseteq [3]$, a sum x in Fig.1 is called an Λ -type sum if $\Lambda = \text{supp}(x)$. For $\Lambda \subseteq [M] - \{\theta\}$, define $\underline{\Lambda} = \Lambda \cup \{\theta\}$ and call Λ -type sums as interference. Let $\gamma_k^{(i)}$ be the number of Λ -type sums in $\text{Serv}^{(i)}$ for each k -subset $\Lambda \subseteq [3]$ and $i \in [5]$.

Now we show the scheme satisfies the correctness condition and the privacy condition. Recall the sufficient conditions for the correctness (s1) requirement in [12], that is, for any $\Lambda \subseteq [M] - \{\theta\}$, the interference parts of all $\underline{\Lambda}$ -type sums can be linearly expressed by the Λ -type sums which appears in all servers. For $\Lambda \subseteq [M] - \{\theta\}$, we collect all Λ -type sums and the interference parts of all $\underline{\Lambda}$ -type sums to form a matrix and

Serv ⁽¹⁾	Serv ⁽²⁾	Serv ⁽³⁾	Serv ⁽⁴⁾	Serv ⁽⁵⁾
a_1	a_4	a_7	a_{10}, a_{11}, a_{12}	a_{13}, a_{14}, a_{15}
b_1	b_4	b_7	b_{10}, b_{11}, b_{12}	b_{13}, b_{14}, b_{15}
c_1	c_4	c_7	c_{10}, c_{11}, c_{12}	c_{13}, c_{14}, c_{15}
$a_2 + b_2$	$a_5 + b_5$	$a_8 + b_8$		
$a_3 + b_3$	$a_6 + b_6$	$a_9 + b_9$		
$a_{16} + c_2$	$a_{18} + c_5$	$a_{20} + c_{14}$		
$a_{17} + c_3$	$a_{19} + c_6$	$a_{21} + c_{15}$		
$b_{16} + c_{16}$	$b_{18} + c_{18}$	$b_{20} + c_{20}$		
$b_{17} + c_{17}$	$b_{19} + c_{19}$	$b_{21} + c_{21}$		
			$a_{22} + b_{22}$	$a_{24} + b_{24}$
			$+c_{22}$	$+c_{24}$
			$a_{23} + b_{23}$	$a_{25} + b_{25}$
			$+c_{23}$	$+c_{25}$

Fig. 1: Answers of the $(N = 5, T = 3, M = 3)$ PIR scheme for retrieving W_1 .

call this matrix as the distribution matrix of Λ -type sums. For example, for $\{2\}$ -type sums $\{b_i\}$, its distribution matrix has the following form,

$$\begin{pmatrix} \mathbf{b}_1 & \mathbf{b}_4 & \mathbf{b}_7 & \mathbf{b}_{10} & \mathbf{b}_{13} \\ b_2 & b_5 & b_8 & \mathbf{b}_{11} & \mathbf{b}_{14} \\ b_3 & b_6 & b_9 & \mathbf{b}_{12} & \mathbf{b}_{15} \end{pmatrix}. \quad (3)$$

where the bold symbols are all $\{2\}$ -type sums and the rest are the interference parts of all $\{2, 3\}$ -type sums. Similarly, the distribution matrix of $\{2, 3\}$ -type sums $\{b_i + c_j\}$ is

$$\begin{pmatrix} \mathbf{b}_{16} + \mathbf{c}_{16} & \mathbf{b}_{18} + \mathbf{c}_{18} & \mathbf{b}_{20} + \mathbf{c}_{20} & b_{22} + c_{22} & b_{24} + c_{24} \\ \mathbf{b}_{17} + \mathbf{c}_{17} & \mathbf{b}_{19} + \mathbf{c}_{19} & \mathbf{b}_{21} + \mathbf{c}_{21} & b_{23} + c_{23} & b_{25} + c_{25} \end{pmatrix}. \quad (4)$$

Then by the MDS property of G_1 and G_2 , the coordinates labeled by the bold symbols in (3) and (4) form an information set of G_1 and G_2 , respectively. That is, the rest symbols can be recovered by the bold symbols in (3) and (4), respectively. Note that for any k -subset $\Lambda \subseteq [M] - \{\theta\}$, the distribution matrix of Λ -type sums in Fig.1 has a similar form, as the matrix (3) or (4), so the interference parts of all $\underline{\Lambda}$ -type sums can be recovered by all Λ -type sums appeared in 5 columns. Hence this scheme satisfies the condition (s1) in [12], i.e., the correctness condition is guaranteed.

As to the privacy, recall the sufficient conditions for the privacy (s2) requirement in [12], it is sufficient to ensure that for any $\Lambda \subseteq [M] - \{\theta\}$, there are the same number of independent symbols contained in any 3 columns of Λ -type sums' distribution matrix (i.e., (3) or (4)). Actually this is guaranteed by the MDS property of the linear code which is used to generate such type interference. Thus the privacy condition is guaranteed.

Moreover the desired record consists of 25 symbols while the answers totally contain 49 symbols, so the scheme has rate $\frac{25}{49}$ attaining the capacity for this case.

The field size relies on the maximum length of the MDS codes used in this scheme, so it requires $q \geq 15$ in Example 1. Note that for any Λ -type interference, if its distribution matrix is a codeword of some MDS array code, then there are also the same number of independent symbols contained in any 3 columns of its distribution matrix. For example, suppose $\Lambda = \{2\}$, the matrix (3) can be viewed as a codeword of an

(5, 3; 3) MDS array code. Similarly, the [10, 6] MDS code also can be viewed as an (5, 3; 2) MDS array code. So if we adopt (5, 3; 3) and (5, 3; 2) MDS array codes rather than [15, 9] and [10, 6] MDS codes, then the new obtained scheme also satisfies the T -privacy condition.

However, there is a problem that how to guarantee the correctness condition. So the MDS array codes have to satisfy some property determined by the correctness condition. More precisely, for any k -subset Λ of $[M] - \{\theta\}$ and MDS array code corresponding to the Λ -type sums, denoted by $(5, 3; \ell_k)$, its generator matrix needs to have the following recovery property:

- (a1) for i th thick column, there are $\gamma_k^{(i)}$ columns which are used to generate the Λ -type sums, and the rest $\gamma_{k+1}^{(i)}$ columns are used to generate $\underline{\Lambda}$ -type sums, that is, $\ell_k = \gamma_k^{(i)} + \gamma_{k+1}^{(i)}$.
- (a2) All these $\sum_{i=1}^N \gamma_k^{(i)}$ columns have full column rank, that is, $\sum_{i=1}^N \gamma_k^{(i)} = T\ell_k$.

Now we give two admissible matrices $\tilde{G}_1 \in \mathbb{F}_2^{9 \times 15}$, $\tilde{G}_2 \in \mathbb{F}_2^{6 \times 10}$. That is,

$$\tilde{G}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix},$$

$$\tilde{G}_2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then one can verify that the columns labeled by $\{1, 4, 7, 10, 11, 12, 13, 14, 15\}$ in \tilde{G}_1 have full column rank and the columns labeled by $\{1, 2, 3, 4, 5, 6\}$ in \tilde{G}_2 also have full column rank. Hence \tilde{G}_1 and \tilde{G}_2 satisfy the recovery property. Actually, \tilde{G}_1 is obtained by applying the method in Theorem 2 to a generator matrix of an [5, 3] Generalized Reed-Solomon code over \mathbb{F}_{2^3} and rearranging the order of columns in each thick column by multiplying some 3×3 permutation matrix. Similarly, \tilde{G}_2 is obtained by using the same method to a generator matrix of a [5, 3] doubly-extended Generalized Reed-Solomon code over \mathbb{F}_{2^2} . Then the MDS property of \tilde{G}_1 and \tilde{G}_2 is also satisfied. Therefore, the new scheme obtained by using \tilde{G} to replace G in (2) is a capacity-achieving T -PIR scheme with optimal sub-packetization, where $\tilde{G} = \begin{pmatrix} \tilde{G}_1 & 0 \\ 0 & \tilde{G}_2 \end{pmatrix}$. Note that the field size is reduced to 2.

As displayed in the example, the main design idea behind our scheme is to make each $(N, T; \ell_k)$ MDS array code corresponding to Λ -type interference in the scheme satisfy the recovery property for any k -subset Λ of $[M] - \{\theta\}$ and some fixed $\gamma_k^{(1)}, \dots, \gamma_k^{(N)}$. Fortunately, we prove that every MDS array code trivially satisfies the recovery property by Lemma 3 in Section IV-A.

IV. THE GENERAL T -PIR SCHEME BASED ON MDS ARRAY CODES

In this section we first characterize the recovery property of MDS array codes and then describe our general capacity-achieving T -PIR scheme based on MDS array codes.

A. The Recovery property of MDS array code

Lemma 3. Suppose $G = (G^{(1)}, G^{(2)}, \dots, G^{(N)}) \in \mathbb{F}_q^{T\ell \times N\ell}$ is a generator matrix of an $(N, T; \ell)$ MDS array code over \mathbb{F}_q , where $G^{(i)} = (\mathbf{g}_1^{(i)}, \dots, \mathbf{g}_\ell^{(i)}) \in \mathbb{F}_q^{T\ell \times \ell}$ and $\mathbf{g}_j^{(i)}$ is a $T\ell$ -length column vector. Then for any $(m_1, \dots, m_N) \in \{0, 1, \dots, \ell\}^N$ with $\sum_{i=1}^N m_i = T\ell$, there exist N subsets $\Gamma_1, \dots, \Gamma_N$ of $[\ell]$ with $|\Gamma_i| = m_i$ such that

$$\text{rank}(G_{\Gamma_1}^{(1)}, G_{\Gamma_2}^{(2)}, \dots, G_{\Gamma_N}^{(N)}) = T\ell.$$

Proof. For any fixed (m_1, \dots, m_N) with $\sum_{i=1}^N m_i = T\ell$, there exist at least T nonzero numbers of them. Without loss of generality, we may assume that $m_i \neq 0, 1 \leq i \leq N$. Because that if $m_i = 0$, the new matrix obtained by deleting the thick block column $G^{(i)}$ is also a generator matrix of an $(N-1, T; \ell)$ MDS array code.

Let

$$b(m_1, m_2, \dots, m_N) = \max_{\substack{\Gamma_1, \dots, \Gamma_N \subseteq [\ell], \\ |\Gamma_i| = m_i, 1 \leq i \leq N}} \text{rank}(G_{\Gamma_1}^{(1)}, G_{\Gamma_2}^{(2)}, \dots, G_{\Gamma_N}^{(N)}).$$

Then there exist N subsets $\Gamma_1, \dots, \Gamma_N$ of $[\ell]$ with $|\Gamma_i| = m_i$ such that $b = \text{rank}(G_{\Gamma_1}^{(1)}, G_{\Gamma_2}^{(2)}, \dots, G_{\Gamma_N}^{(N)})$. Choose a maximum linearly independent subset of the vectors $\{G_{\Gamma_1}^{(1)}, G_{\Gamma_2}^{(2)}, \dots, G_{\Gamma_N}^{(N)}\}$, denoted by $\{G_{\Gamma_i^{(1)}}^{(i)} : i \in [N]\}$, then $b = \text{rank}(G_{\Gamma_1^{(1)}}^{(1)}, G_{\Gamma_2^{(1)}}^{(2)}, \dots, G_{\Gamma_N^{(1)}}^{(N)})$, where $\Gamma_i^{(1)} \subseteq \Gamma_i$ for $i \in [N]$ and $\sum_{i=1}^N |\Gamma_i^{(1)}| = b$. Then it is sufficient to show that $b = T\ell$.

On the contrary, we assume that $b < T\ell$. Let $\mathbf{V} = \text{Cspan}(G_{\Gamma_1}^{(1)}, G_{\Gamma_2}^{(2)}, \dots, G_{\Gamma_N}^{(N)})$, where $\text{Cspan}(\cdot)$ denotes the linear space spanned by all columns of the matrix over \mathbb{F}_q . Then $\dim \mathbf{V} = b < T\ell$ and $\{G_{\Gamma_i^{(1)}}^{(i)} : i \in [N]\}$ is a base of the vector space \mathbf{V} . To derive a contradiction, we assume the following claim has been proved.

Claim : if $b < T\ell$, then for $1 \leq f \leq T$, there exist f disjoint nonempty subsets X_1, \dots, X_f of $[N]$ such that $\forall u \in \bigcup_{i=1}^f X_i, \forall j \in [\ell], \mathbf{g}_j^{(u)} \in \mathbf{V}$.

Particularly, let $f = T$. Then it follows from the **Claim** that there exist T disjoint nonempty subsets X_1, \dots, X_T of $[N]$ such that $\forall u \in \bigcup_{i=1}^T X_i, \forall j \in [\ell], \mathbf{g}_j^{(u)} \in \mathbf{V}$. Hence $\text{rank}(G^{(u)} : u \in \bigcup_{i=1}^T X_i) \leq \dim \mathbf{V} = b$. On the other hand, note that $|\bigcup_{i=1}^T X_i| = \sum_{i=1}^T |X_i| \geq T$. Combining with the MDS property of G , then $\text{rank}(G^{(u)} : u \in \bigcup_{i=1}^T X_i) = T\ell$. So one can obtain that $T\ell \leq b < T\ell$, a contradiction.

To complete the proof, it remains to prove the **Claim**. Now we prove it by induction on f .

For $f = 1$, let $X_1 = \{i \in [N] : \Gamma_i^{(1)} \neq \emptyset\}$. Since $b < T\ell$, then $|X_1| \geq 1$. For any $u \in X_1$, it is sufficient to show that

for $j \in [\ell] - \Gamma_u$, $\mathbf{g}_j^{(u)} \in V$. Then choosing a m_u -subset Γ'_u of $[\ell]$ such that $\Gamma_u^{(1)} \cup \{j\} \subseteq \Gamma'_u$, one can obtain that

$$\mathbf{V} \subseteq \text{Cspan}(G_{\Gamma_1}^{(1)}, \dots, G_{\Gamma_{u-1}}^{(u-1)}, G_{\Gamma'_u}^{(u)}, G_{\Gamma_{u+1}}^{(u+1)}, \dots, G_{\Gamma_N}^{(N)}).$$

By the definition of $b(m_1, m_2, \dots, m_N)$, it holds that

$$\text{rank}(G_{\Gamma_1}^{(1)}, \dots, G_{\Gamma_{u-1}}^{(u-1)}, G_{\Gamma'_u}^{(u)}, G_{\Gamma_{u+1}}^{(u+1)}, \dots, G_{\Gamma_N}^{(N)}) \leq b = \dim \mathbf{V},$$

which implies that

$$\mathbf{V} = \text{Cspan}(G_{\Gamma_1}^{(1)}, \dots, G_{\Gamma_{u-1}}^{(u-1)}, G_{\Gamma'_u}^{(u)}, G_{\Gamma_{u+1}}^{(u+1)}, \dots, G_{\Gamma_N}^{(N)}).$$

Hence, $\mathbf{g}_j^{(u)} \in \mathbf{V}$.

Suppose that there exist $f - 1$ disjoint nonempty subsets X_1, \dots, X_{f-1} of $[N]$ such that

$$\forall u \in \bigcup_{i=1}^{f-1} X_i, \forall j \in [\ell], \mathbf{g}_j^{(u)} \in \mathbf{V}.$$

Consider the case f , note that $\text{rank}(G^{(u)} : u \in \bigcup_{i=1}^{f-1} X_i) \leq \dim \mathbf{V} = b < T\ell$. By the MDS property (1) of G , then the $|\bigcup_{i=1}^{f-1} X_i| < T$, which implies that the vectors $\{\mathbf{g}_j^{(u)} : u \in \bigcup_{i=1}^{f-1} X_i, j \in [\ell]\}$ are linearly independent over \mathbb{F}_q , so are the vectors $\{\mathbf{g}_j^{(u)} : u \in \bigcup_{i=1}^{f-1} X_i, j \in \Gamma_u\}$. Therefore the vectors can extend to be a base of \mathbf{V} . Then there exist $\Gamma_i^{(f)} \subseteq \Gamma_i$ for $1 \leq i \leq N$ and $\Gamma_u^{(f)} = \Gamma_u$ for $u \in \bigcup_{i=1}^{f-1} X_i$ such that $\{G_{\Gamma_i^{(f)}}^{(i)} : i \in [N]\}$ is a base of the vector space \mathbf{V} . Let $X_f = \{i \in [N] : \Gamma_i^{(f)} \neq \Gamma_i\}$. Then $X_f \neq \emptyset$, otherwise $\dim \mathbf{V} = \sum_{i=1}^N |\Gamma_i| = T\ell$. By the definition of X_f , one can obtain that $X_f \cap (\bigcup_{i=1}^{f-1} X_i) = \emptyset$, that is, such f subsets $X_i, i \in [f]$ are disjoint. Similarly, by using the same way in the case $f = 1$, one can obtain that $\forall u \in X_f, \forall j \in [\ell], \mathbf{g}_j^{(u)} \in \mathbf{V}$. \square

Remark 1. Using the notations introduced above, we may assume that for any fixed (m_1, m_2, \dots, m_N) with $\sum_{i=1}^N m_i = T\ell$, $\Gamma_i = \{1, 2, \dots, m_i\}, i \in [N]$ in a generator matrix of the $(N, T; \ell)$ MDS array code. This is because that we can rearrange the order of ℓ columns in each thick column by multiplying some $\ell \times \ell$ permutation matrix.

B. Formal Description of the general scheme

Our scheme can be obtained by modifying the capacity-achieving T -PIR schemes in [12]. As in Example 1, we replace $M - 1$ MDS codes with some $M - 1$ MDS array codes. Next we give these $M - 1$ desired MDS array codes.

Specially, for $1 \leq k \leq M - 1$, the k th MDS code defined by the generator matrix G_k in [12] has the parameters $[\frac{N}{T}(T\alpha_k + (N - T)\beta_k), T\alpha_k + (N - T)\beta_k]$ over \mathbb{F}_q , where α_k, β_k are defined as in the identities (35), (36) in [12]. Note that $T\alpha_k + (N - T)\beta_k = T(n - t)^{k-1}t^{M-1-k}$, and define $\ell_k = (n - t)^{k-1}t^{M-1-k}$, where $t = T/\gcd(N, T)$, $n = N/\gcd(N, T)$. Then for $1 \leq k \leq M - 1$, the k th MDS code can be viewed as an $(N, T; \ell_k)$ MDS array code. By Lemma 3, one can choose a generator matrix \tilde{G}_k of an $(N, T; \ell_k)$ MDS array code which has the recovery property for $(\gamma_k^{(1)}, \dots, \gamma_k^{(N)})$, where $\gamma_k^{(i)} = \alpha_k$ for $1 \leq i \leq T$ and $\gamma_k^{(i)} = \beta_k$ for $T + 1 \leq i \leq N$. Then these $M - 1$ matrices \tilde{G}_k are desired.

One can verify that the new scheme satisfies the correctness condition and T -privacy condition, which are guaranteed by the recovery property and MDS property of all $M - 1$ MDS array codes, respectively. Moreover, the new scheme doesn't change the sub-packetization of records and download size. Therefore the new scheme has the highest rate and the optimal sub-packetization. Note that there are $M - 1$ MDS array codes used in our scheme over \mathbb{F}_q , by Theorem 2 it only needs to requires that for $1 \leq k \leq M - 1$, $q^{\ell_k} \geq N$. That is, $q \geq \sqrt[\ell]{N}$, where $\ell = \min\{t^{M-2}, (n - t)^{M-2}\}$.

V. CONCLUSION

In this paper we build a general capacity-achieving T -PIR scheme based on MDS array codes over \mathbb{F}_q , that is, the queries are generated by using $M - 1$ MDS array codes rather than MDS codes. It requires the field size $q \geq \sqrt[\ell]{N}$ and has optimal sub-packetization. In particular, the binary field is enough to build our scheme as long as $M \geq 2 + \lceil \log_\mu \log_2 N \rceil$, where $\mu = \min\{t, n - t\} > 1$.

REFERENCES

- [1] A. Beimel, Y. Ishai, E. Kushilevitz, and I. Orlov, "Share conversion and private information retrieval," in *Proc. 27th Annu. Conf. Comput. Complex.*, pp. 258–268, Jun. 2012.
- [2] S. Yekhanin, "Locally Decodable Codes and Private Information Retrieval Schemes," *Ph.D. dissertation*, Massachusetts Institute of Technology, 2007.
- [3] B. Chor, E. Kushilevitz, O. Goldreich, M. Sudan, "Private information retrieval," *Proc. 36-th IEEE Symposium on Foundations of Computer Science*, pp.41–50, 1995.
- [4] H. Sun, S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. on Inf. Theory*, vol.63, no.7, pp. 4075–4088, Jul. 2017.
- [5] H. Sun, S. A. Jafar, "The capacity of private information retrieval with colluding databases," *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp.941–946, 2016.
- [6] H. Sun, S. A. Jafar, "The Capacity of Symmetric Private Information Retrieval," *IEEE Globecom Workshops (GC Wkshps)*, pp.1–5, 2016.
- [7] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans Inf Theory*, vol.64 no.3, pp.1945–1956, Mar. 2018.
- [8] H. Sun, S. A. Jafar, "Optimal Download Cost of Private Information Retrieval for Arbitrary Message Length", *IEEE Transactions on Information Forensics and Security*, vol.12, no.12, pp.2920–2932, 2017.
- [9] H.-Y. Lin, S. Kumar, E. Rosnes, and A. Graell i Amat, "An MDS-PIR Capacity-Achieving Protocol for Distributed Storage Using Non-MDS Linear Codes," *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, July 2018, pp. 966–970.
- [10] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, et al. "Private Information Retrieval Schemes Using Transitive Codes," *IEEE Transactions on Information Theory*, DOI 10.1109/TIT.2018.2871050.
- [11] Yiwei Zhang and Gennian Ge, "A general private information retrieval scheme for MDS coded databases with colluding servers," arXiv:1704.06785
- [12] Zhifang Zhang and Jingke Xu, "The Optimal Sub-Packetization of Linear Capacity-Achieving PIR Schemes with Colluding Servers," *IEEE Transaction on Information Theory*, DOI 10.1109/TIT.2018.2883283.
- [13] Jingke Xu and Zhifang Zhang, "On sub-packetization and access number of capacity-achieving PIR schemes for MDS coded non-colluding servers," *Science China Information Science*, 2018, Vol. 61 (10), pp.1–16.
- [14] Jingke Xu and Zhifang Zhang, "Building Capacity-Achieving PIR Schemes with Optimal Sub-Packetization over Small Fields," *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, July 2018, pp. 1749–1753.
- [15] R. Lidl and H. Niederreiter, "Introduction to finite fields and their applications," *Cambridge University Press, Cambridge, UK (1994)*.