Two-Level Private Information Retrieval

Ruida Zhou, *Student Member, IEEE* Chao Tian, *Senior Member, IEEE*, Hua Sun *Member, IEEE*, and James S. Plank

***Abstract*—In the conventional robust T-colluding private information retrieval (PIR) system, the user needs to retrieve one of the possible messages while keeping the identity of the requested message private from any T colluding servers. Motivated by the possible heterogeneous privacy requirements for different messages, we consider the $(N, T_1 : K_1, T_2 : K_2)$ two-level PIR system with a total of K_2 messages in the system, where $T_1 \geq T_2$ and $K_1 \leq K_2$. Any one of the K_1 messages needs to be retrieved privately against T_1 colluding servers, and any one of the full set of K_2 messages needs to be retrieved privately against T_2 colluding servers. We obtain a lower bound to the capacity by proposing two novel coding schemes, namely the non-uniform successive cancellation scheme and the non-uniform block cancellation scheme. A capacity upper bound is also derived. The gap between the upper bound and the lower bounds is analyzed, and shown to vanish when $T_1 = T_2$. Lastly, we show that the upper bound is in general not tight by providing a stronger bound for a special setting.**

***Index Terms*—Colluding, information retrieval, privacy**

I. Introduction

Capacity characterizations of the canonical private information retrieval (PIR) system and its variants have drawn considerable attention recently in the information and coding theory community, for which novel code constructions and impossibility results have been discovered.

In the canonical PIR model, user privacy needs to be preserved during message retrieval from replicated servers, i.e., the identity of the desired message should not be revealed to any single server. Specifically, the user is required to retrieve one of the K messages from N servers, each of which stores a copy of K messages, such that the identity of the desired message is not revealed to any single server. In the PIR capacity characterization problem, the goal is to identify the minimum download cost, i.e., the minimum amount of download per-bit of the desired message, the inverse of which is referred to as the capacity of PIR. The PIR capacity was characterized in [1] through a code construction and a matching converse bound. The code construction recursively exploits three key elements: server symmetry, message symmetry, and side information; the converse bound recursively reduces the problem scale by utilizing the privacy constraint.

The canonical PIR problem formulation is to some extent idealized and possesses abundant symmetry and homogeneity (both in the servers and messages), which were judiciously exploited in the code construction proposed in [1]. Going forward, it is imperative to enrich the canonical model to make it more heterogenous and comprehensive so that 1) practical constraints that arise naturally in diverse applications are incorporated and tackled, and 2) the extendability and limitation of the capacity results [1] are better understood. Along this line, the following aspects that generalize the canonical model have been studied in the literature.

1) Colluding pattern: Privacy is guaranteed against each single server in the canonical model, which has been generalized to any set of T colluding servers in [2]. The T-colluding privacy constraint was further generalized to the fully heterogeneous model where each colluding set of servers can be an arbitrary subset of all servers [3], [4]. Interestingly, while server symmetry appears to be broken, the recursively constructed MDS coded queries can still be allocated according to a linear program, and furthermore, this elegant solution was shown to be optimal [4].
2) Download per server: As the message size is allowed to approach infinity in capacity characterizations, the download size per server can be made the same through symmetrization in the canonical model [5]. However, if other metrics are considered such as message size [5]–[7] or physical constraints that limit the communication link between each server and the user [8], schemes with heterogeneous downloads per server are useful and sometimes necessary. While server symmetry is lost, the iterative construction from [1] can proceed with the two remaining elements in a similar manner [6], [8], [9].
3) Message size: The K messages are assumed to have equal length and allowed to approach infinity in the canonical model. The generalization to arbitrary different lengths was considered in [10] and the iterative construction from [1] was applied to truncated subsets of messages with the same length [10]. The minimum message sizes for capacity-achieving codes were considered in [5], [7] where server symmetry and side information were utilized in the code constructions.
4) Server storage: Each server has the same storage capability and stores all K messages in the canonical model. The storage system at the servers has been generalized to MDS coded [11], [12] or coded by a given linear code [13] for each message, and arbitrarily uncoded [14] with

R. Zhou and C. Tian are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station TX 77843. H. Sun is with the Department of Electrical Engineering, University of North Texas, Denton TX 76203. J. S. Plank is with the Department of Electrical Engineering and Computer Science, University of Tennessee Knoxville, Knoxville TN 37996.

The work of R.-D Zhou and C. Tian was supported in part by the National Science Foundation under Grants CCF-1816546 and CCF-2007067. The work of H. Sun was supported in part by the National Science Foundation under Grants CCF-2007108 and CCF-2045656. The work of J. S. Plank was supported in part by the National Science Foundation under Grant CCF-1816518.

This work was presented in part at 2021 IEEE International Symposium on Information Theory, Melbourne, Victoria, Australia, July 2021.

heterogeneous capabilities [15], [16]. In these settings the iterative construction from [1] is still largely compatible with the storage structure. However, for the general model where all messages can be jointly coded, the tradeoff between the storage constraint and the download cost is far from being fully understood [17]–[20].

The main motivation of this work is a crucial aspect that has not been previously addressed – the heterogeneity of the privacy constraints on the messages. That is, in all existing works, each message is required to be equally private in the sense that any single server [1], or any colluding set of T servers [2], is completely ignorant of the desired message identity. However, the sensitivities of different types of information are commonly different in practice. To be more concrete, let us consider the following simple example setting.

Example 1. *There are a total of four short videos, which are replicated on six storage servers. The first two videos are political campaign videos from two opposing political parties, while the other two videos are non-political music videos. Given the sensitivity of revealing one's political view, as well as the requirement of protecting the user's privacy in a general sense, the user may wish to assure the following privacy protection when retrieving one of these videos:*

- *Any one of the servers will not be able to infer any information regarding which message is being requested;*
- *Any three of the servers jointly will not be able to infer any information regarding which one of the first two messages is being requested.*

Consider the following several scenarios: 1) When any video is retrieved, any one of the server will not infer any information regarding which was being requested, and any four or more servers may collude to infer exactly which was being retrieved, 2) When the user retrieves one of the political campaign videos, any two or three servers may collude to infer that the retrieved video is indeed a political campaign video, but they will not be able to infer which one it is, thus protecting the user's political view; 3) When a non-political video is retrieved, any two or three servers may collude to infer exactly which video is retrieved. Therefore, the user's political view is indeed protected in a stronger manner than his preference among general contents. It should be noted that the user is not enforcing a stronger privacy protection against the fact that a political video is retrieved in general, since this fact alone does not reveal any sensitive information about the user's political preference: only the information on exactly which political video is retrieved will reveal such information.

Motivated by the consideration above, we formulate the problem of multilevel private information retrieval problem. Specifically, the *privacy level* of a message set is defined as the maximum allowed number of colluding servers that the identity of a desired message is kept private among that message set. We focus on the two-level PIR system, where some K_1 messages out of the K_2 messages have a higher privacy level of T_1, i.e., any colluding set of T_1 servers do not learn anything about which one of the K_1 messages is desired, while all the K_2 messages together have a lower privacy level of T_2, any colluding set of T_2 servers do not learn anything about which one of the K_2 messages is desired.

Characterizing the capacity of the two-level PIR system turns out to be rather challenging. A naive approach, which can be used as a baseline, is to treat the system as if it were a homogeneous T_1-colluding private information retrieval system. However, the crux of the two-level PIR hinges on how to leverage the less stringent privacy requirement for some messages. Towards this end, we must treat the two sets of messages with distinct privacy levels differently, i.e., message symmetry cannot be taken for granted. Without message symmetry, the iterative construction breaks since message symmetry is the key step that enables the connection between the layers, and we have to delve deeper into the code structure to adjust the parameters of the MDS coded queries in a heterogeneous manner. As a result, we discover two general schemes that can outperform the naive baseline scheme. For the converse direction, we first apply the iterative induction technique to obtain a general upper bound, and analyze the gap between the upper bound and the lower bound. We then show that this bound is strictly sub-optimal by deriving a tighter bound for a special case. This implies that the induction technique must be combined with more delicate consideration on the heterogeneous nature of the system. This observation may shed some light on other open settings, where it is not known if similar symmetric reduction based converse bounds are tight [8], [21], [22].

Notations: We adopt the notation $i:j \triangleq \{i, i+1, \ldots, j-1, j\}$. Denote vector $a_{\mathcal{N}} \triangleq (a_i)_{i\in\mathcal{N}}$ for any sequence $(a_1, a_2, \ldots)$ and $\mathcal{N} \subset \mathbb{N}$. We use $X \sim Y$ to indicate that the random variables X and Y follow an identical distribution. For any matrix $A[:,:]$, the first coordinate is for row indices and the second coordinate is for column indices.

II. Problem Formulation

There are K_2 mutually independent messages $W_{1:K_2} = (W_1, W_2, \ldots, W_{K_2})$ in the system. Each message is uniformly distributed over $\mathbb{F}_q^L$, where $\mathbb{F}_q$ is a large enough finite field and L is the number of q-ary symbols in the message (i.e., the message length). This is equivalent to

$$H(W_1) = H(W_2) = \cdots = H(W_{K_2}) = L, \tag{1}$$

$$H(W_{1:K_2}) = K_2 L, \tag{2}$$

where (and in the rest of this work) we take base-q logarithm for simplicity. There are N servers in the system, each of which stores a copy of all the K_2 messages. Let $k^* \in 1:K_2$ be the identity of the desired message. The process to retrieve message W_{k^*}, for any $k^* \in 1:K_2$, involves three steps:

1. (Query) The user sends a randomized query $Q_n^{[k^*]}$ to server n for each $n \in 1:N$;
2. (Answer) Each server n, where $n \in 1:N$, returns an answer $A_n^{[k^*]}$ to the user;
3. (Recovery) The user recovers the message as $\hat{W}_{k^*}$, using the queries $Q_{1:N}^{[k^*]}$ to all the servers and the answers $A_{1:N}^{[k^*]}$ from all the servers.

Denote the set of all possible queries sent to server n as $\mathcal{Q}_n$. $Q_n^{[k^*]} \in \mathcal{Q}_n$ is a random variable, whose superscript $[k^*]$

indicates that the query is for retrieving message W_{k^*}. The user has no knowledge of $W_{1:K_2}$, and thus the queries are independent of the messages, that is

$$I(Q_{1:N}^{[k^*]}; W_{1:K_2}) = 0, \quad \forall k^* \in 1 : K_2. \tag{3}$$

Each symbol of the answer $A_n^{[k^*]}$, the answer to the query $Q_n^{[k^*]}$, is a sequence of symbols in $\mathbb{F}_q$; denote the number of symbols of $A_n^{[k^*]}$ as $\ell_n^{[k^*]}$. The answer $A_n^{[k^*]}$ is a deterministic function of the query $Q_n^{[k^*]}$ and the messages $W_{1:K_2}$, that is

$$H(A_n^{[k^*]} | Q_n^{[k^*]}, W_{1:K_2}) = 0, \quad \forall k^* \in 1 : K_2,\ n \in 1 : N. \tag{4}$$

The recovered message $\hat{W}_{k^*}$ depends on the queries $Q_{1:N}^{[k^*]}$ as well as the answers $A_{1:N}^{[k^*]}$, that is

$$H(\hat{W}_{k^*} | A_{1:N}^{[k^*]}, Q_{1:N}^{[k^*]}) = 0, \quad \forall k^* \in 1 : K_2. \tag{5}$$

The message should be retrieved correctly, i.e., $W_{k^*} = \hat{W}_{k^*}$ for all $k^* \in 1 : K_2$. Additionally, the system has certain privacy requirements. To measure user privacy when querying for any message in a certain set of messages, we first introduce the definition of *privacy level*.

Definition 1 (Privacy level)**.** *Let the messages in the system be $W_1, W_2, \ldots, W_K$. The queries of a scheme have privacy level T for a subset of messages $W_{\mathcal{S}}$, where $\mathcal{S} \subseteq 1 : K$, if for any $\mathcal{T} \subseteq 1 : N$ with $|\mathcal{T}| = T$, for retrieving any message in $W_{\mathcal{S}}$, the queries to the servers in $\mathcal{T}$ have the same joint distribution, i.e.,*

$$Q_{\mathcal{T}}^{[k]} \sim Q_{\mathcal{T}}^{[k']}, \qquad \forall k, k' \in \mathcal{S}. \tag{6}$$

The notion of privacy level has the following operational meaning: if $W_{\mathcal{S}}$ has privacy level T, then when one of the messages in $W_{\mathcal{S}}$ is retrieved, even if any T of the N servers collude, the identity of the requested message in $W_{\mathcal{S}}$ remains private, however these colluding servers may be able to infer that the requested message is in the set $W_{\mathcal{S}}$. It is straightforward to verify that the set of messages with higher privacy level automatically has lower privacy levels. In addition, when the set $\mathcal{S}$ is a singleton, if T servers can infer the desired message is in $W_{\mathcal{S}}$, the identity of the desired message is known. Thus it is not meaningful to study the privacy level of $W_{\mathcal{S}}$ for singleton $\mathcal{S}$, though we will still allow it for notational convenience.

In this work, we consider the two-level PIR system. The system parameters in such a system are $(N, T_1 : K_1, T_2 : K_2)$ with $T_1 \geq T_2 \geq 1$ and $1 \leq K_1 \leq K_2$. All the messages $W_{1:K_2}$ have the default weaker privacy level T_2, but the first K_1 messages $W_{1:K_1}$ have an enhanced privacy level T_1. We are interested in the retrieval rate (or simply rate) which is the number of useful message symbols retrieved per unit download

$$R \triangleq \frac{L}{\sum_{n=1}^{N} \mathbb{E}[\ell_n^{[k^*]}]}. \tag{7}$$

The download cost D is defined as the inverse of R, i.e., $D \triangleq R^{-1}$. Schemes with higher achievable rates are preferred, and the supremum of the achievable rates among all possible schemes is called the capacity of the system, denoted as C.

III. Main Result

We first provide some new notation. Define the function $D_N^*(K, T)$ as follows

$$D_N^*(K, T) \triangleq 1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-1},\ \forall T, K, N \in \mathbb{N}, \tag{8}$$

whose inverse is the capacity of the T-colluding PIR system with N servers and K messages (sometimes simply referred to as a T-private system). The main result of this work is summarized in the theorem below.

Theorem 1. *The capacity C of the $(N, T_1 : K_1, T_2 : K_2)$ two-level PIR system satisfies*

$$\max\left(R_{\mathrm{NS}}, R_{\mathrm{NB}}\right) \leq C \leq \overline{R}, \tag{9}$$

where

$$\overline{R} = \left(D_N^*\left(K_1, T_1\right) + \frac{T_2}{N}\left(\frac{T_1}{N}\right)^{K_1 - 1} D_N^*(K_2 - K_1, T_2)\right)^{-1}, \tag{10}$$

$$R_{\mathrm{NS}} = \left(D_N^*\left(K_1, T_1\right) + \left(\frac{T_1}{N}\right)^{K_1} D_N^*(K_2 - K_1, T_2)\right)^{-1}, \tag{11}$$

$$R_{\mathrm{NB}} = \left(\max\left(D_N^*(K_1, T_1) + \frac{T_2}{N} D_N^*(K_2 - K_1, T_2),\right.\right.$$
$$\left.\left. D_N^*(K_2 - K_1, T_2) + \frac{T_2}{N} D_N^*(K_1, T_1)\right)\right)^{-1}. \tag{12}$$

The lower bound to the capacity in this theorem has two components: R_{NS} is obtained by the Non-uniform Successive-cancellation (NS) coding scheme given in Section V, and R_{NB} is obtained by the Non-uniform Block-cancellation (NB) coding scheme given in Section VI. The proof for the upper bound $\overline{R}$ is given in the supplementary material. The upper bound $\overline{R}$ in Theorem 1 is in general not tight. Specifically, the following proposition tightens the upper bound for the $(3, 2 : 2, 1 : 3)$ two-level PIR system, for which Theorem 1 gives an upper bound of $\frac{9}{17}$.

Proposition 1. *The capacity C of the $(3, 2 : 2, 1 : 3)$ two-level PIR system satisfies*

$$C \leq \frac{11}{21}. \tag{13}$$

The proof of this proposition is given in the supplementary material, which is obtained using the computer-aided approach discussed in [23]–[25].

To further understand these bounds in Theorem 1, define

$$\underline{D} = \overline{R}^{-1}, \quad D_{\mathrm{NS}} = R_{\mathrm{NS}}^{-1}, \quad D_{\mathrm{NB}} = R_{\mathrm{NB}}^{-1}.$$

Three observations are in order:

1) Theorem 1 gives that

$$\underline{D} \leq \min D \leq \min\left(D_{\mathrm{NS}}, D_{\mathrm{NB}}\right). \tag{14}$$

The difference between $\underline{D}$ and $D_{\rm NS}$ is

$$D_{\rm NS} - \underline{D} = \frac{T_1 - T_2}{N} \left(\frac{T_1}{N}\right)^{K_1 - 1} D^*(K_2 - K_1, T_2).$$

It is seen that this gap diminishes geometrically as K_1 grows, and also vanishes when $T_1 = T_2$ as expected.

2) Any $(N, T_1 : K_2, T_2 : K_2)$ code, i.e., a T_1-private code with N servers and K_2 messages, is valid for the $(N, T_1 : K_1, T_2 : K_2)$ PIR system. The optimal download cost of the former is exactly given by $D_{\text{T-PIR}} = D_N^*(K_2, T_1)$. Comparing with this naive approach, the coding gain of the proposed NS scheme is thus

$$D_{\text{T-PIR}} - D_{\rm NS} = \left(\frac{T_1}{N}\right)^{K_1} \left(D_N^*(K_2 - K_1, T_1) - D_N^*(K_2 - K_1, T_2)\right),$$

which is non-negative, and strictly positive if and only if $K_2 - K_1 \geq 2$. Note that the strategy of using an $(N, T_1 : K_2, T_2 : K_2)$ code when a message in $W_{\mathcal{S}}$ is requested, and using an $(N, T_1 : 1, T_2 : K_2)$ code for the other messages is not valid, since this would lead to privacy leakage in the latter case, i.e., leaking the information that the requested message is not in the set $\mathcal{S}$.

3) The relation between $R_{\rm NS}$ and $R_{\rm NB}$ is as follows.

- For the cases when

$$D_N^*(K_1, T_1) \geq D_N^*(K_2 - K_1, T_2) \text{ and } \frac{T_2}{N} < \left(\frac{T_1}{N}\right)^{K_1},$$

the lower bound $R_{\rm NB}$ is better

$$R_{\rm NS} < R_{\rm NB} = \left(D_N^*(K_1, T_1) + \frac{T_2}{N} D_N^*(K_2 - K_1, T_2)\right)^{-1};$$

- For the cases when

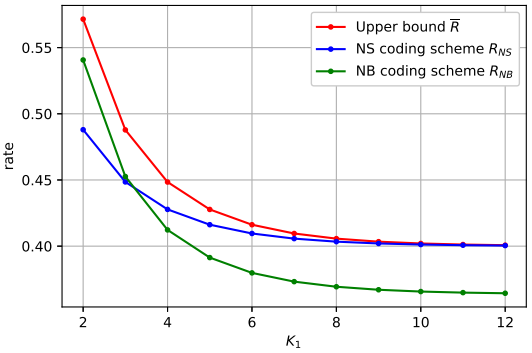
$$D_N^*(K_1, T_1) < D_N^*(K_2 - K_1, T_2) \quad \text{and} \quad \frac{D_N^*(K_1, T_1)}{1 - \left(\frac{T_1}{N}\right)^{K_1}} > \frac{D_N^*(K_2 - K_1, T_2)}{1 - \frac{T_2}{N}},$$

the lower bound $R_{\rm NB}$ is also better

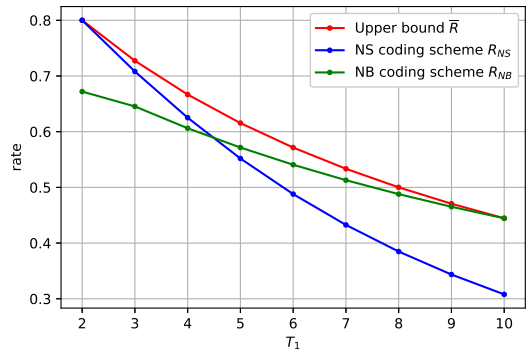
$$R_{\rm NS} < R_{\rm NB} = \left(D_N^*(K_2 - K_1, T_2) + \frac{T_2}{N} D_N^*(K_1, T_1)\right)^{-1};$$

- For all the other cases, the lower bound $R_{\rm NS}$ is better, i.e., $R_{\rm NB} \leq R_{\rm NS}$.

The upper bound and lower bounds are shown in Figure 1. In Figure 1(a), the gap between the upper bound $\overline{R}$ and the rate of NS coding scheme $R_{\rm NS}$ deminishes geometrically as K_1 grows. It can be seen in Figure 1(b), that when T_1 is close to T_2, the NS scheme performs better, and matches the upper bound if $T_1 = T_2$; when T_1 is close to N, the NB scheme performs better, and in this case matches the upper bound if $T_1 = N$.



(a) $(10, 6 : K_1, 2 : K_1 + 4)$ two-level PIR



(b) $(10, T_1 : 2, 2 : 6)$ two-level PIR

Fig. 1. Upper and lower bounds on the capacity of two-level PIR system

TABLE I
A 2-PRIVATE CODE WITH $(N, K) = (4, 2)$

Server-1	Server-2	Server-3	Server-4
a_1, b_1	a_2, b_2	a_3, b_3	a_4, b_4
$a_5 + b_5$	$a_6 + b_6$	$a_7 + b_7$	$a_8 + b_8$

IV. A Gentle Start

In this section, we first provide a brief review of the T-colluding PIR code using two example cases, and partly based on insights obtained from these example cases, we provide two example codes to illustrate the proposed coding schemes.

A. Two T-colluding PIR examples

As mentioned earlier, an $(N, T_1 : K_1, T_2 : K_2)$ two-level PIR system degrades to a T_1-private system when $K_1 = K_2$, and thus it is expected that there is a connection between the code construction for the T-private systems and that for the 2-level PIR systems. The capacity of the T-private system was identified in [2]. We next consider two special cases of the codes proposed there, in order to provide the necessary intuition for the proposed codes.

1) First set $(N, T_1 : K_1, T_2 : K_2) = (4, 2 : 2, 1 : 2)$, which is essentially a 2-private system with $N = 4$ servers and $K = K_1 = K_2 = 2$ messages. In the code given in [2], the message length is 8. The messages are first

TABLE II
A 1-PRIVATE CODE WITH $(N, K) = (4, 2)$

Server-1	Server-2	Server-3	Server-4
c_1, d_1	c_2, d_2	c_3, d_3	c_4, d_4
$c_5 + d_5$	$c_6 + d_6$	$c_7 + d_7$	$c_8 + d_8$
$c_9 + d_9$	$c_{10} + d_{10}$	$c_{11} + d_{11}$	$c_{12} + d_{12}$
$c_{13} + d_{13}$	$c_{14} + d_{14}$	$c_{15} + d_{15}$	$c_{16} + d_{16}$

precoded as $W_1^* = S_1 W_1$ and $W_2^* = S_2 W_2$, where S_1 and S_2 are random matrices drawn uniformly from the set of all 8×8 full-rank matrices over $\mathbb{F}_q$. Let $a_{1:8}$ and $b_{1:8}$ be MDS-coded symbols of messages W_1^* and W_2^*, respectively, using appropriate coding parameters. The coding structure is given in Table I. To retrieve W_1, we choose $a_{1:8} = W_1^*$ and $b_{1:8}$ to be $(8, 4)$-MDS coded symbols using any 4 symbols from W_2^*; the coding parameters for retrieving W_2 are obvious by symmetry. Since the symbols $b_{5:8}$ can be recovered from $b_{1:4}$, W_1 can be recovered correctly. It is not difficult to verify that the retrieval is private due to the precoding and MDS-coding steps.

2) Next let $(N, T_1 : K_1, T_2 : K_2) = (4, 1 : 2, 1 : 2)$, which is essentially the canonical PIR system with $N = 4$ servers and $K = K_1 = K_2 = 2$ messages. The coding structure is given in Table II, where $c_{1:16}$ and $d_{1:16}$ are MDS-coded symbols of two messages, respectively, with appropriate coding parameters. To retrieve W_1, we can use $c_{1:16} = W_1^*$ and let $d_{1:16}$ be $(16, 4)$-MDS coded of any 4 symbols in W_2^*; the coding parameters for retrieving W_2 are obvious by symmetry. Since the symbols $d_{5:16}$ can be recovered from $d_{1:4}$, W_1 can be recovered correctly and privately.

Comparing the two cases, a few observations are in order:

1) The codes in Table I and Table II have two layers: the first layer has single symbols, i.e., a, b, c, or d, and the second layer has summations of two symbols, i.e., $a + b$ or $c + d$.
2) Although the 2-private code meets the privacy requirement of the 1-private system, the coding structure in Table II is more efficient. Particularly, the ratio between the first layer transmissions and second layer transmissions changes from $8 : 4$ to $8 : 12$. Placing more symbols in the second layer is preferable, because one desired symbol essentially takes two symbol transmissions in the first layer, yet it takes only one in the second layer.
3) The improved transmission ratio between the two layers is a consequence of the chosen MDS coding parameters for the non-requested message (i.e., the interference): in Table I, it is $(8, 4)$ while in Table II it is $(16, 4)$. These parameters, which are chosen to satisfy the decoding and privacy requirements, determine the number of symbols in different layers.

These observations suggest that in a two-level PIR system, we will need to adjust the MDS coding parameters for different messages according to their privacy levels, but maintain the code structure consistent between the two cases when retrieving two types of messages. This is a considerable generalization of the T-private setting, since in the T-private setting the MDS coding parameters can be chosen uniformly for all the messages, except the requested message, while in our setting, the privacy levels create heterogeneity among the messages.

In the code construction given in [2], the following lemma plays an instrumental role in formally showing the privacy condition to hold, which we shall also utilize in this work.

Lemma 1 (Statistical effect of full rank matrices [2])**.** *Let $S_1, S_2, \ldots, S_K \in \mathbb{F}_q^{\alpha \times \alpha}$ be K random matrices, drawn independently and uniformly from all $\alpha \times \alpha$ full-rank matrices over $\mathbb{F}_q$. Let $G_1, G_2, \ldots, G_K \in \mathbb{F}_q^{\beta \times \beta}$ be K invertible square matrices of dimension $\beta \times \beta$ over $\mathbb{F}_q$. Let $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K \in \mathbb{N}^{\beta \times 1}$ be K index vectors, each containing β distinct indices from $[1 : \alpha]$. Then*

$$\begin{aligned}&(G_1 S_1[\mathcal{I}_1, :], G_2 S_2[\mathcal{I}_2, :], \ldots, G_K S_K[\mathcal{I}_K, :]) \\ &\quad \sim (S_1[1 : \beta, :], S_2[1 : \beta, :], \ldots, S_K[1 : \beta, :]),\end{aligned} \tag{15}$$

where the notation $S[\mathcal{I}, :]$ is used to indicate the submatrix of S by taking its rows in $\mathcal{I}$.

B. An example of the NS scheme

We next provide an example to illustrate the proposed NS coding scheme. In this example, the two-level PIR system is specified by the parameters $(N, T_1 : K_1, T_2 : K_2) = (4, 2 : 2, 1 : 4)$, i.e., there are 4 servers and 4 messages $W_{1:4}$, and messages $W_{1:2}$ have privacy level $T_1 = 2$, while all messages $W_{1:4}$ have privacy level $T_2 = 1$. The length of each message is $L = 64$ here.

Encoding: To retrieve a message, the answers are formed in three steps, and the queries are simply the encoding matrix for these answers. Assume for each (n, k) pair where $n \geq k$, an MDS code in $\mathbb{F}_q$ is given and fixed, and we refer to it as the (n, k) MDS code. The coding structure is illustrated in Table III and Table IV, for the retrieval of W_1 and W_4, respectively. The coding steps can be understood as follows:

1) *Precoding:* Let S_1, S_2, S_3, and S_4 be four random matrices, which are independently and uniformly drawn from the set of all 64×64 full rank matrices over $\mathbb{F}_q$; these matrices are known only to the user. The precoded messages $W_{1:4}^*$ are

$$\begin{aligned}W_1^* &= S_1 W_1; \quad W_2^* = S_2 W_2; \\ W_3^* &= S_3 W_3; \quad W_4^* = S_4 W_4.\end{aligned} \tag{16}$$

2) *Group-wise MDS coding:* The precoded messages are partitioned into non-overlapping segments, and each segment is MDS-coded under certain (n, k) parameters, the result of which is referred to as a coding group. These MDS-coded symbols for the four messages are denoted as $a_{1:64}, b_{1:64}, c_{1:64}, d_{1:64}$, respectively. In the tables, these coding groups are distinguished using different colors, with the corresponding MDS parameters given in the first column. For example, the red coding groups in Table III for both $b_{25:28,33:36}$ and $c_{9:12,25:28}$ are obtained by encoding 4 pre-coded symbols in W_2^* and W_4^*, respectively. In each coding group, the coded symbols are ordered

TABLE III
NS SCHEME IN $(N, T_1 : K_1, T_2 : K_2) = (4, 2 : 2, 1 : 4)$ FOR RETRIEVING W_1

Coding group	Server-1	Server-2	Server-3	Server-4
a: (64, 64)	a_1, a_2, a_3	a_4, a_5, a_6	a_7, a_8, a_9	a_{10}, a_{11}, a_{12}
b: (24, 12)	b_1, b_2, b_3	b_4, b_5, b_6	b_7, b_8, b_9	b_{10}, b_{11}, b_{12}
c: (8, 4)	c_1	c_2	c_3	c_4
d: (8, 4)	d_1	d_2	d_3	d_4
	$a_{13} + b_{13}$	$a_{14} + b_{14}$	$a_{15} + b_{15}$	$a_{16} + b_{16}$
	$a_{17} + b_{17}$	$a_{18} + b_{18}$	$a_{19} + b_{19}$	$a_{20} + b_{20}$
	$a_{21} + b_{21}$	$a_{22} + b_{22}$	$a_{23} + b_{23}$	$a_{24} + b_{24}$
	$a_{25} + c_5$	$a_{26} + c_6$	$a_{27} + c_7$	$a_{28} + c_8$
	$a_{29} + d_5$	$a_{30} + d_6$	$a_{31} + d_7$	$a_{32} + d_8$
b, c: (8,4)	$b_{25} + c_9$	$b_{26} + c_{10}$	$b_{27} + c_{11}$	$b_{28} + c_{12}$
b, d: (8,4)	$b_{29} + d_9$	$b_{30} + d_{10}$	$b_{31} + d_{11}$	$b_{32} + d_{12}$
c, d: (24, 12)	$c_{13} + d_{13}$	$c_{16} + d_{16}$	$c_{19} + d_{19}$	$c_{22} + d_{22}$
	$c_{14} + d_{14}$	$c_{17} + d_{17}$	$c_{20} + d_{20}$	$c_{23} + d_{23}$
	$c_{15} + d_{15}$	$c_{18} + d_{18}$	$c_{21} + d_{21}$	$c_{24} + d_{24}$
	$a_{33} + b_{33} + c_{25}$	$a_{34} + b_{34} + c_{26}$	$a_{35} + b_{35} + c_{27}$	$a_{36} + b_{36} + c_{28}$
	$a_{37} + b_{37} + d_{25}$	$a_{38} + b_{38} + d_{26}$	$a_{39} + b_{39} + d_{27}$	$a_{40} + b_{40} + d_{28}$
	$a_{41} + c_{29} + d_{29}$	$a_{42} + c_{30} + d_{30}$	$a_{43} + c_{31} + d_{31}$	$a_{44} + c_{32} + d_{32}$
	$a_{45} + c_{33} + d_{33}$	$a_{46} + c_{34} + d_{34}$	$a_{47} + c_{35} + d_{35}$	$a_{48} + c_{36} + d_{36}$
	$a_{49} + c_{37} + d_{38}$	$a_{50} + c_{38} + d_{38}$	$a_{51} + c_{39} + d_{39}$	$a_{52} + c_{40} + d_{40}$
b, c, d: (24, 12)	$b_{41} + c_{41} + d_{41}$	$b_{42} + c_{42} + d_{42}$	$b_{43} + c_{43} + d_{43}$	$b_{44} + c_{44} + d_{44}$
	$b_{45} + c_{45} + d_{45}$	$b_{46} + c_{46} + d_{46}$	$b_{47} + c_{47} + d_{47}$	$b_{48} + c_{48} + d_{48}$
	$b_{49} + c_{49} + d_{49}$	$b_{50} + c_{50} + d_{50}$	$b_{51} + c_{51} + d_{51}$	$b_{52} + c_{52} + d_{52}$
	$a_{53} + b_{53} + c_{53} + d_{53}$	$a_{54} + b_{54} + c_{54} + d_{54}$	$a_{55} + b_{55} + c_{55} + d_{55}$	$a_{56} + b_{56} + c_{56} + d_{56}$
	$a_{57} + b_{57} + c_{57} + d_{57}$	$a_{58} + b_{58} + c_{58} + d_{58}$	$a_{59} + b_{59} + c_{59} + d_{59}$	$a_{60} + b_{60} + c_{60} + d_{60}$
	$a_{61} + b_{61} + c_{61} + d_{61}$	$a_{62} + b_{62} + c_{62} + d_{62}$	$a_{63} + b_{63} + c_{63} + d_{63}$	$a_{64} + b_{64} + c_{64} + d_{64}$

and sequentially placed in the tables, indicated by their subscripts.

3) *Forming pre-coded message sums:* The summations of the MDS-coded messages are formed accordingly, which can be seen clearly from Table III and Table IV.

Decoding and correctness: The coding structure is layered, where in each layer the number of summands in each downloaded symbol is the same. From top to bottom, the number of summands increases from 1 to 4. The symbols of interference messages in each coding group are placed in two adjacent layers, where the signals (i.e., the summation of the symbols of interference messages) in the top layer can decode the interference signals in lower layer due to the common linear MDS code.

In Table III, for each coding group, the total number of interference signals placed in two adjacent layers and the top layer follow the ratio $(2:1) = (8:4) = (24:12)$. For example, 8 interference signals in the red coding group are placed in the second and third layers, where 4 downloaded symbols $b_{25:28} + c_{9:12}$ in the second layer can decode $b_{33:36} + c_{25:28}$ in the third layer, because b, c are encoded by the same linear $(8, 4)$ MDS code. Consequently, $a_{33:36}$ can be recovered. It can be verified that $a_{1:64}$ can all be recovered either directly or in this fashion. By symmetry, W_2 can be retrieved similarly.

In Table IV, for each coding group, the numbers of interference signals of each coding group placed in two adjacent layers and the top layer have the ratio at most $4:1$. For example, 16 interference signals in red coding groups are placed in the second and third layers, where any 4 of the 12 downloaded symbols $a_{13:24} + b_{13:24}$ in the second layer can decode $a_{37:40} + b_{37:40}$ in the third layer because a, b are encoded by the same linear $(16, 4)$ MDS code. Consequently, $d_{25:28}$ can be recovered. It can be verified that $d_{1:64}$ can all be recovered either directly, or in this fashion. By symmetry, W_3 can be retrieved similarly.

Privacy: The coding pattern, i.e., the manner of forming pre-coded message sums, is the same for the retrieval of any message in $W_{1:4}$. Since it is a linear code, the coded symbols can be generated by the corresponding coding matrices. From Table III, it is seen that the coding matrix of the coded symbols of any message from any two servers has full row-rank. For examples, the coded symbols a's in server-1 and server-2 can be generated by a full row rank coding matrix using the message W_1, due to the pre-coding and the group-wise MDS coding. By applying Lemma 1, the messages $W_{1:2}$ thus have privacy level 2. The 1-privacy for all the messages can be seen in a similar manner.

Performance: The total number of downloaded symbols is 116 and the message length is 64. Thus the rate is $R_{\text{NS}} = \frac{64}{116} = \frac{16}{29}$. The scheme for 2-private systems has rate $\frac{8}{15} < R_{\text{NS}}$.

Remark: The construction resembles the scheme in [2] (also discussed in Section IV-A), but it allows non-uniform coding structure to leverage the requirements of two levels of privacy.

TABLE IV
NS SCHEME IN $(N, T_1 : K_1, T_2 : K_2) = (4, 2 : 2, 1 : 4)$ FOR RETRIEVING W_4

Coding group	Server-1	Server-2	Server-3	Server-4
d: (64, 64)	d_1	d_2	d_3	d_4
a: (16, 4) b: (16, 4) c: (16, 4)	a_1, a_2, a_3 b_1, b_2, b_3 c_1	a_4, a_5, a_6 b_4, b_5, b_6 c_2	a_7, a_8, a_9 b_7, b_8, b_9 c_3	a_{10}, a_{11}, a_{12} b_{10}, b_{11}, b_{12} c_4
	$a_{29} + d_5$ $b_{29} + d_9$ $c_{13} + d_{13}$ $c_{14} + d_{14}$ $c_{15} + d_{15}$	$a_{30} + d_6$ $b_{30} + d_{10}$ $c_{16} + d_{16}$ $c_{17} + d_{17}$ $c_{18} + d_{18}$	$a_{31} + d_7$ $b_{31} + d_{11}$ $c_{19} + d_{19}$ $c_{20} + d_{20}$ $c_{21} + d_{21}$	$a_{32} + d_8$ $b_{32} + d_{12}$ $c_{22} + d_{22}$ $c_{23} + d_{23}$ $c_{24} + d_{24}$
a, b : (16, 4) a, c: (16,4) b, c: (16,4)	$a_{13} + b_{13}$ $a_{17} + b_{17}$ $a_{21} + b_{21}$ $a_{25} + c_5$ $b_{25} + c_9$	$a_{14} + b_{14}$ $a_{18} + b_{18}$ $a_{22} + b_{22}$ $a_{26} + c_6$ $b_{26} + c_{10}$	$a_{15} + b_{15}$ $a_{19} + b_{19}$ $a_{23} + b_{23}$ $a_{27} + c_7$ $b_{27} + c_{11}$	$a_{16} + b_{16}$ $a_{20} + b_{20}$ $a_{24} + b_{24}$ $a_{28} + c_8$ $b_{28} + c_{12}$
	$a_{37} + b_{37} + d_{25}$ $a_{41} + c_{29} + d_{29}$ $a_{45} + c_{33} + d_{33}$ $a_{49} + c_{37} + d_{38}$ $b_{41} + c_{41} + d_{41}$ $b_{45} + c_{45} + d_{45}$ $b_{49} + c_{49} + d_{49}$	$a_{38} + b_{38} + d_{26}$ $a_{42} + c_{30} + d_{30}$ $a_{46} + c_{34} + d_{34}$ $a_{50} + c_{38} + d_{38}$ $b_{42} + c_{42} + d_{42}$ $b_{46} + c_{46} + d_{46}$ $b_{50} + c_{50} + d_{50}$	$a_{39} + b_{39} + d_{27}$ $a_{43} + c_{31} + d_{31}$ $a_{47} + c_{35} + d_{35}$ $a_{51} + c_{39} + d_{39}$ $b_{43} + c_{43} + d_{43}$ $b_{47} + c_{47} + d_{47}$ $b_{51} + c_{51} + d_{51}$	$a_{40} + b_{40} + d_{28}$ $a_{44} + c_{32} + d_{32}$ $a_{48} + c_{36} + d_{36}$ $a_{52} + c_{40} + d_{40}$ $b_{44} + c_{44} + d_{44}$ $b_{48} + c_{48} + d_{48}$ $b_{52} + c_{52} + d_{52}$
a, b, c: (16,4)	$a_{33} + b_{33} + c_{25}$	$a_{34} + b_{34} + c_{26}$	$a_{35} + b_{35} + c_{27}$	$a_{36} + b_{36} + c_{28}$
	$a_{53} + b_{53} + c_{53} + d_{53}$ $a_{57} + b_{57} + c_{57} + d_{57}$ $a_{61} + b_{61} + c_{61} + d_{61}$	$a_{54} + b_{54} + c_{54} + d_{54}$ $a_{58} + b_{58} + c_{58} + d_{58}$ $a_{62} + b_{62} + c_{62} + d_{62}$	$a_{55} + b_{55} + c_{55} + d_{55}$ $a_{59} + b_{59} + c_{59} + d_{59}$ $a_{63} + b_{63} + c_{63} + d_{63}$	$a_{56} + b_{56} + c_{56} + d_{56}$ $a_{60} + b_{60} + c_{60} + d_{60}$ $a_{64} + b_{64} + c_{64} + d_{64}$

Due to the homogeneity of the privacy requirements for all the messages in T-private systems, the MDS coding parameters for each coding group are chosen to be (N, T). In the proposed scheme for the $(N, T_1 : K_1, T_2 : K_2)$ system, there is symmetry among servers, and also symmetries among $W_{1:K_1}$ and among $W_{K_1+1:K_2}$ but not across all the messages. Thus when retrieving message W_{k^*} with $k^* \in 1 : K_1$, the ratio of the MDS parameters (n, k) in each coding group of the undesired messages need to be chosen as (N, T_1), while as for message W_{k^*} with $k^* \in K_1 + 1 : K_2$, the MDS coding parameters in each coding group would be (N, T_2). However, since $N/T_1 < N/T_2$, with the same retrieval pattern, there exists certain slack in the placement pattern when retrieving W_{k^*} with $k^* \in K_1 + 1 : K_2$. For example, the red coding group in Table IV only needs 4 symbols in layer-2 to decode the remaining symbols in both layer-2 and layer-3, yet 12 symbols are retrieved and available directly in layer-2.

C. An example of the NB scheme

We provide an example to illustrate the proposed NB coding scheme for the same two-level PIR system specified by paramters $(N, T_1 : K_1, T_2 : K_2) = (4, 2 : 2, 1 : 4)$. The length of each message is again $L = 64$.

Encoding: The coding structure is illustrated in Table V and Table VI, for the retrieval of W_1 and W_4, respectively. The coding procedure also consists of three steps, as in the NS code, however the patterns are different, which is evident from the tables.

Decoding and correctness: There are three blocks in Table V and Table VI. In Table V, the symbols $c_{1:4}$, $d_{1:4}$, and $c_{5:16} + d_{5:16}$ in the second block can be used to reconstruct the interference signals in the third block, i.e., $c_{17:28}$, $d_{17:28}$, and $c_{29:64} + d_{29:64}$, by the property of the MDS code in each coding group. Canceling these interference signals generated by $W_{3:4}$, i.e., eliminating the coded symbols c and d, Table V essentially reduces to the scheme discussed in Section IV-A for the 2-private system: here 32 interference signals $b_{1:8,17:40}$ can be used to reconstruct $b_{9:16,41:64}$. The desired message W_1 can thus be recovered. By symmetry, W_2 can be retrieved similarly.

In Table VI, the symbols $a_{1:8}$, $b_{1:8}$, and $a_{9:16} + b_{9:16}$ in the first block can be used to reconstruct the interference signals in the third block, i.e., $a_{17:40}$, $b_{17:40}$, and $a_{41:64} + b_{41:64}$. Canceling the interference signals generated by $W_{1:2}$, i.e., eliminating the coded symbols a and b, Table VI reduces to the scheme discussed in Section IV-A for the 1-private system, and the desired message W_4 can be recovered. By symmetry, W_3 can be retrieved similarly.

Privacy: When message W_1 or W_2 is requested, the coded symbols of message $W_{3:4}$ are downloaded as interference signals, and the interference signals such as c, d, or $c + d$ are mixed to a, b, $a + b$. With the symbols c, d eliminated in Table V, we have the retrieval pattern of the 2-private system, which

TABLE V
NB SCHEME IN $(N, T_1 : K_1, T_2 : K_2) = (4, 2 : 2, 1 : 4)$ FOR RETRIEVING W_1

Coding group	Server-1	Server-2	Server-3	Server-4
$a : (64, 64)$ $b : (64, 32)$	a_1, b_1 a_2, b_2	a_3, b_3 a_4, b_4	a_5, b_5 a_6, b_6	a_7, b_7 a_8, b_8
	$a_9 + b_9$ $a_{10} + b_{10}$	$a_{11} + b_{11}$ $a_{12} + b_{12}$	$a_{13} + b_{13}$ $a_{14} + b_{14}$	$a_{15} + b_{15}$ $a_{16} + b_{16}$
$c : (16, 4)$; $d : (16, 4)$	c_1, d_1	c_2, d_2	c_3, d_3	c_4, d_4
$c + d : (48, 12)$	$c_5 + d_5$ $c_6 + d_6$ $c_7 + d_7$	$c_8 + d_8$ $c_9 + d_9$ $c_{10} + d_{10}$	$c_{11} + d_{11}$ $c_{12} + d_{12}$ $c_{13} + d_{13}$	$c_{14} + d_{14}$ $c_{15} + d_{15}$ $c_{16} + d_{16}$
	$a_{17} + c_{17}$, $b_{17} + d_{17}$ $a_{21} + c_{21}$, $b_{21} + d_{21}$ $a_{25} + c_{25}$, $b_{25} + d_{25}$ $a_{29} + c_{29} + d_{29}$ $a_{30} + c_{30} + d_{30}$ $a_{31} + c_{31} + d_{31}$ $b_{29} + c_{41} + d_{41}$ $b_{30} + c_{42} + d_{42}$ $b_{31} + c_{43} + d_{43}$ $a_{41} + b_{41} + c_{53} + d_{53}$ $a_{42} + b_{42} + c_{54} + d_{54}$ $a_{43} + b_{43} + c_{55} + d_{55}$ $a_{53} + b_{53}$ $a_{54} + b_{54}$ $a_{55} + b_{55}$	$a_{18} + c_{18}$, $b_{18} + d_{18}$ $a_{22} + c_{22}$, $b_{22} + d_{22}$ $a_{26} + c_{26}$, $b_{26} + d_{26}$ $a_{32} + c_{32} + d_{32}$ $a_{33} + c_{33} + d_{33}$ $a_{34} + c_{34} + d_{34}$ $b_{32} + c_{44} + d_{44}$ $b_{33} + c_{45} + d_{45}$ $b_{34} + c_{46} + d_{46}$ $a_{44} + b_{44} + c_{56} + d_{56}$ $a_{45} + b_{45} + c_{57} + d_{57}$ $a_{46} + b_{46} + c_{58} + d_{58}$ $a_{56} + b_{56}$ $a_{57} + b_{57}$ $a_{58} + b_{58}$	$a_{19} + c_{19}$, $b_{19} + d_{19}$ $a_{23} + c_{23}$, $b_{23} + d_{23}$ $a_{27} + c_{27}$, $b_{27} + d_{27}$ $a_{35} + c_{35} + d_{35}$ $a_{36} + c_{36} + d_{36}$ $a_{37} + c_{37} + d_{37}$ $b_{35} + c_{47} + d_{47}$ $b_{36} + c_{48} + d_{48}$ $b_{37} + c_{49} + d_{49}$ $a_{47} + b_{47} + c_{59} + d_{59}$ $a_{48} + b_{48} + c_{60} + d_{60}$ $a_{49} + b_{49} + c_{61} + d_{61}$ $a_{59} + b_{59}$ $a_{60} + b_{60}$ $a_{61} + b_{61}$	$a_{20} + c_{20}$, $b_{20} + d_{20}$ $a_{24} + c_{24}$, $b_{24} + d_{24}$ $a_{28} + c_{28}$, $b_{28} + d_{28}$ $a_{38} + c_{38} + d_{38}$ $a_{39} + c_{39} + d_{39}$ $a_{40} + c_{40} + d_{40}$ $b_{38} + c_{50} + d_{50}$ $b_{39} + c_{51} + d_{51}$ $b_{40} + c_{52} + d_{52}$ $a_{50} + b_{50} + c_{62} + d_{62}$ $a_{51} + b_{51} + c_{63} + d_{63}$ $a_{52} + b_{52} + c_{64} + d_{64}$ $a_{62} + b_{62}$ $a_{63} + b_{63}$ $a_{64} + b_{64}$

is clearly 2-private. To see all the messages have privacy level 1, observe that the coding pattern is the same for the retrieval of any message. In both Table V and Table VI, the coding matrix of the coded symbols for any single message from any single server has full row rank. Thus by Lemma 1, messages $W_{1:4}$ have privacy level 1.

Performance: The total number of downloaded symbols is 116 and the message length is 64. Thus the rate is $R_{\text{NB}} = \frac{64}{116} = \frac{16}{29}$, which coincides with R_{NS} in this example.

Remark: The coding structure has the following feature: eliminating the coded symbols c and d in Table V or Table VI, the remaining part has the same coding structure as the 2-private code discussed in Section IV-A; eliminating the coded symbols a and b in Table V or Table VI, the remaining part has the same coding structure as the 1-private code discussed in Section IV-A. The NB coding structure can be interpreted as a mixture of the T_1-private code of message $W_{1:K_1}$ and T_2-private code for messages $W_{K_1+1:K_2}$ discussed in Section IV-A, which is constructed in three blocks. Since the retrieval needs to follow the same pattern, the underlying T_1-private code and the underlying T_2-private code are required to have the same message length. A $\frac{T_2}{N}$ fraction of the T_1-private code forms the first block, a $\frac{T_2}{N}$ fraction of the T_2-private code forms the second block. The remaining $\frac{N-T_2}{N}$ fractions of both codes are mixed together to form the third block by simple pairwise summations in an arbitrary order; in case they have different numbers of remaining coded symbols, the remaining summands are included directly.

V. THE NON-UNIFORM SUCCESSIVE CANCELLATION SCHEME

In this section, we provide the general code construction for the non-uniform successive cancellation scheme.

A. Specifying coding group parameters

It is clear from the example in Section IV-B that the proposed code can be viewed as consisting of K_2 layers and multiple coding groups. We next first specify the appropriate parameters for each coding group. We identify each coding group by its composition. For example, in Table III, the red coding group is of form $b + c$, and thus we can use the set of message indices involved to identify it as $\mathcal{K} = \{2, 3\}$, i.e., it involves the messages (W_2, W_3). Clearly this coding group will be placed in the 2^{nd} and 3^{rd} layers.

More generally, for each coding group, there are a total of five parameters to specify: the total number of coded symbols $n_1(\mathcal{K})$ and $n_2(\mathcal{K})$, and the number of MDS code message symbols $k_1(\mathcal{K})$ and $k_2(\mathcal{K})$ when retrieving a message of privacy level T_1 and that of privacy level T_2, respectively; and the number of symbols to be placed in the top layer $m(\mathcal{K})$. In other words, during the retrieval of a message W_{k^*}, when $k^* \in 1 : K_1$, an $(n_1(\mathcal{K}), k_1(\mathcal{K}))$ MDS code is used for this coding group, while during the retrieval of a message W_{k^*}, when $k^* \in K_1+1 : K_2$, an $(n_2(\mathcal{K}), k_2(\mathcal{K}))$ MDS code is used for this coding group. After MDS encoding, $m(\mathcal{K})$ symbols will be placed in the $|\mathcal{K}|$-th layer, while the remaining will be placed in the $(|\mathcal{K}| + 1)$-th layer as interference, and the symbols are uniformly distributed across all servers.

TABLE VI
NB SCHEME IN $(N, T_1 : K_1, T_2 : K_2) = (4, 2 : 2, 1 : 4)$ FOR RETRIEVING W_4

Coding group	Server-1	Server-2	Server-3	Server-4
$a : (32, 8)$	a_1, b_1	a_3, b_3	a_5, b_5	a_7, b_7
$b : (32, 8)$	a_2, b_2	a_4, b_4	a_6, b_6	a_8, b_8
$a + b : (32, 8)$	$a_9 + b_9$	$a_{11} + b_{11}$	$a_{13} + b_{13}$	$a_{15} + b_{15}$
	$a_{10} + b_{10}$	$a_{12} + b_{12}$	$a_{14} + b_{14}$	$a_{16} + b_{16}$
d: (64, 64)	c_1, d_1	c_2, d_2	c_3, d_3	c_4, d_4
$c : (64, 16)$	$c_5 + d_5$	$c_8 + d_8$	$c_{11} + d_{11}$	$c_{14} + d_{14}$
	$c_6 + d_6$	$c_9 + d_9$	$c_{12} + d_{12}$	$c_{15} + d_{15}$
	$c_7 + d_7$	$c_{10} + d_{10}$	$c_{13} + d_{13}$	$c_{16} + d_{16}$
	$a_{17} + c_{17}$, $b_{17} + d_{17}$	$a_{18} + c_{18}$, $b_{18} + d_{18}$	$a_{19} + c_{19}$, $b_{19} + d_{19}$	$a_{20} + c_{20}$, $b_{20} + d_{20}$
	$a_{21} + c_{21}$, $b_{21} + d_{21}$	$a_{22} + c_{22}$, $b_{22} + d_{22}$	$a_{23} + c_{23}$, $b_{23} + d_{23}$	$a_{24} + c_{24}$, $b_{24} + d_{24}$
	$a_{25} + c_{25}$, $b_{25} + d_{25}$	$a_{26} + c_{26}$, $b_{26} + d_{26}$	$a_{27} + c_{27}$, $b_{27} + d_{27}$	$a_{28} + c_{28}$, $b_{28} + d_{28}$
	$a_{29} + c_{29} + d_{29}$	$a_{32} + c_{32} + d_{32}$	$a_{35} + c_{35} + d_{35}$	$a_{38} + c_{38} + d_{38}$
	$a_{30} + c_{30} + d_{30}$	$a_{33} + c_{33} + d_{33}$	$a_{36} + c_{36} + d_{36}$	$a_{39} + c_{39} + d_{39}$
	$a_{31} + c_{31} + d_{31}$	$a_{34} + c_{34} + d_{34}$	$a_{37} + c_{37} + d_{37}$	$a_{40} + c_{40} + d_{40}$
	$b_{29} + c_{41} + d_{41}$	$b_{32} + c_{44} + d_{44}$	$b_{35} + c_{47} + d_{47}$	$b_{38} + c_{50} + d_{50}$
	$b_{30} + c_{42} + d_{42}$	$b_{33} + c_{45} + d_{45}$	$b_{36} + c_{48} + d_{48}$	$b_{39} + c_{51} + d_{51}$
	$b_{31} + c_{43} + d_{43}$	$b_{34} + c_{46} + d_{46}$	$b_{37} + c_{49} + d_{49}$	$b_{40} + c_{52} + d_{52}$
	$a_{41} + b_{41} + c_{53} + d_{53}$	$a_{44} + b_{44} + c_{56} + d_{56}$	$a_{47} + b_{47} + c_{59} + d_{59}$	$a_{50} + b_{50} + c_{62} + d_{62}$
	$a_{42} + b_{42} + c_{54} + d_{54}$	$a_{45} + b_{45} + c_{57} + d_{57}$	$a_{48} + b_{48} + c_{60} + d_{60}$	$a_{51} + b_{51} + c_{63} + d_{63}$
	$a_{43} + b_{43} + c_{55} + d_{55}$	$a_{46} + b_{46} + c_{58} + d_{58}$	$a_{49} + b_{49} + c_{61} + d_{61}$	$a_{52} + b_{52} + c_{64} + d_{64}$
	$a_{53} + b_{53}$	$a_{56} + b_{56}$	$a_{59} + b_{59}$	$a_{62} + b_{62}$
	$a_{54} + b_{54}$	$a_{57} + b_{57}$	$a_{60} + b_{60}$	$a_{63} + b_{63}$
	$a_{55} + b_{55}$	$a_{58} + b_{58}$	$a_{61} + b_{61}$	$a_{64} + b_{64}$

The message length for the NS coding scheme is $L = N^{K_2}$ in the proposed scheme; note that the length may be reduced in some cases, however we choose this value to simplify the presentation of the code construction without any loss in terms of the download cost. To introduce $(n_1(\mathcal{K}), k_1(\mathcal{K}), n_2(\mathcal{K}), k_2(\mathcal{K}), m(\mathcal{K}))$, we first define

$$M \triangleq T_2^{K_2-K_1} + \frac{T_1 - T_2}{N - T_2}\left(N^{K_2-K_1} - T_2^{K_2-K_1}\right)$$
$$= N^{K_2-K_1} - \frac{N - T_1}{N - T_2}\left(N^{K_2-K_1} - T_2^{K_2-K_1}\right), \quad (17)$$

which is an integer. For any $(i, j) \in 0 : K_1 \times 0 : K_2 - K_1$, define $d_{0,0} \triangleq 0$ and for $i + j \geq 1$, define

$$d_{i,j} \triangleq \begin{cases} MT_1^{K_1-i}(N - T_1)^{i-1}, & \text{if } j = 0 \\ T_1^{K_1-i}(N - T_1)^i T_2^{K_2-K_1-j}(N - T_2)^{j-1}, & \text{o.w.} \end{cases}$$

Then we specify

$$m(\mathcal{K}) \triangleq N d_{|\mathcal{K}\cap 1:K_1|, |\mathcal{K}\cap K_1+1:K_2|}, \quad (18)$$

and

$$n_1(\mathcal{K}) \triangleq m(\mathcal{K}) + N d_{|\mathcal{K}\cap 1:K_1|+1, |\mathcal{K}\cap K_1+1:K_2|}, \quad (19)$$
$$n_2(\mathcal{K}) \triangleq m(\mathcal{K}) + N d_{|\mathcal{K}\cap 1:K_1|, |\mathcal{K}\cap K_1+1:K_2|+1}, \quad (20)$$
$$k_1(\mathcal{K}) \triangleq \frac{T_1}{N} n_1(\mathcal{K}), \quad k_2(\mathcal{K}) \triangleq \frac{T_2}{N} n_2(\mathcal{K}). \quad (21)$$

The properties of the functions used for encoding, correctness and privacy of the NS coding scheme, are summarized as Lemma 2 below, which is proved in the supplementary material.

Lemma 2. *The tuple $(n_1(\cdot), k_1(\cdot), n_2(\cdot), k_2(\cdot), m(\cdot))$ has the following properties:*

1) For any non-empty $\mathcal{K} \subset 1 : K_2$,

$$k_1(\mathcal{K}) = m(\mathcal{K}), \quad k_2(\mathcal{K}) \leq m(\mathcal{K}) \quad (22)$$

2) The following equality holds:

$$\sum_{\mathcal{K}\subset 1:K_2,\ k^*\in\mathcal{K}} m(\mathcal{K}) = L \quad (23)$$

3) When $k^* \in 1 : K_1$, for any $k \neq k^*$ the following inequality holds:

$$\sum_{\mathcal{K}\subset 1:K_2/\{k^*\},\ k\in\mathcal{K}} k_1(\mathcal{K}) < L \quad (24)$$

When $k^* \in K_1 + 1 : K_2$, for any $k \neq k^*$, the following inequality holds:

$$\sum_{\mathcal{K}\subset 1:K_2/\{k^*\},\ k\in\mathcal{K}} k_2(\mathcal{K}) < L \quad (25)$$

B. Encoding, decoding, privacy, and performance

Encoding: The queries and answers are formed in three steps, and the queries are simply the encoding matrix for these answers. Assume for each (n, k) pair where $n \geq k$, an MDS code in $\mathbb{F}_q$ is given and fixed, and we refer to it as the (n, k) MDS code.

1) *Precoding:* Let $S_{1:K_2}$ be K_2 independent random matrices, which are uniformly drawn from the set of all

$N^{K_2} \times N^{K_2}$ full rank matrices over $\mathbb{F}_q$; these matrices are known only to the user. The precoded messages $W^*_{1:K_2}$ are

$$W^*_k = S_k W_k, \quad \forall k \in 1:K_2. \tag{26}$$

2) *Group-wise MDS coding:*
The precoded messages are partitioned into non-overlapping segments, and each segment is MDS-coded under certain parameters. We use $W_k(\mathcal{K})$ to denote a segment of message W_k indexed by $\mathcal{K} \subset 1:K_2$. One special coding group corresponds to the precoded desired message $W^*_{k^*}$, where the precoded message $W^*_{k^*}$ is (N^{K_2}, N^{K_2}) MDS-coded into $\tilde{W}_{k^*}$, which is then partitioned into non-overlapping segments $\tilde{W}_{k^*}(\mathcal{K}\cup\{k^*\})$ for each $\mathcal{K} \subset 1:K_2$, where $\tilde{W}_{k^*}(\mathcal{K}\cup\{k^*\})$ has length $m(\mathcal{K}\cup\{k^*\})$. The non-overlapping segments of $W^*_{k^*}$ exist because of item 2 in Lemma 2. Other coding groups are indexed by non-empty sets $\mathcal{K} \subset 1:K_2/\{k^*\}$. For each $\mathcal{K} \subset 1:K_2/\{k^*\}$, the coding group indexed by $\mathcal{K}$ is specified as follows.
   - If $k^* \in 1:K_1$, for each $k \in \mathcal{K}$, a segment of W^*_k with length $k_1(\mathcal{K})$ is $(n_1(\mathcal{K}), k_1(\mathcal{K}))$ MDS-coded into $(\tilde{W}_k(\mathcal{K}), \tilde{W}_k(\mathcal{K}\cup\{k^*\}))$, which have lengths $m(\mathcal{K})$ and $m(\mathcal{K}\cup\{k^*\})$, respectively.
   - If $k^* \in K_1+1:K_2$, for each $k \in \mathcal{K}$, a segment of W^*_k with length $k_2(\mathcal{K})$ is $(n_2(\mathcal{K}), k_2(\mathcal{K}))$ MDS-coded into $(\tilde{W}_k(\mathcal{K}), \tilde{W}_k(\mathcal{K}\cup\{k^*\}))$, which have lengths $m(\mathcal{K})$ and $m(\mathcal{K}\cup\{k^*\})$, respectively.

   The non-overlapping segments of W^*_k for any $k \neq k^*$ exist because of item 3 in Lemma 2.
3) *Forming pre-coded message sums:* There are K_2 layers in the coding structure. The summation of the MDS-coded messages are placed in the layered structure from top to bottom as follows. For $i = 1, 2, \ldots, K_2$, in the i-th layer, the summations (which are vectors) are

$$A^{[k^*]}_{1:N}(\mathcal{K}) = \sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K}), \quad \forall \mathcal{K} \subset 1:K_2 \text{ with } |\mathcal{K}| = i, \tag{27}$$

   and each vector is partitioned and distributed to N servers uniformly. The MDS coded symbols of coding group indexed by $\mathcal{K}$ are shown in Table VII.

Decoding and correctness: For any non-empty $\mathcal{K} \subset 1:K_2/\{k^*\}$, the MDS-coded interference symbols $(\tilde{W}_k(\mathcal{K}), \tilde{W}_k(\mathcal{K}\cup\{k^*\}))_{k\in\mathcal{K}}$ in the coding group indexed by $\mathcal{K}$ are placed in two adjacent layers. Specifically, $(\tilde{W}_k(\mathcal{K}))_{k\in\mathcal{K}}$ are placed in the $|\mathcal{K}|$-th layer in the form of a signal

$$A^{[k^*]}_{1:N}(\mathcal{K}) = \sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K}), \tag{28}$$

and $(\tilde{W}_k(\mathcal{K}\cup\{k^*\}))_{k\in\mathcal{K}}$ are placed in the $(|\mathcal{K}|+1)$-th layer in the form of

$$A^{[k^*]}_{1:N}(\mathcal{K}\cup\{k^*\}) = \tilde{W}_{k^*}(\mathcal{K}\cup\{k^*\}) + \sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K}\cup\{k^*\}). \tag{29}$$

The interference signal in the top layer can cancel the interference signal in the bottom layer. The interference signal $\sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K})$ in the $|\mathcal{K}|$-th layer (top layer) has length $m(\mathcal{K})$.

- When $k^* \in 1:K_1$, since $(\tilde{W}_k(\mathcal{K}), \tilde{W}_k(\mathcal{K}\cup\{k^*\})$ are encoded by the same linear $(n_1(\mathcal{K}), k_1(\mathcal{K}))$ MDS code for each $k \in \mathcal{K}$, by item 1 in Lemma 2, that $m(\mathcal{K}) = k_1(\mathcal{K})$, the interference signal $\sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K}\cup\{k^*\})$ in the $(|\mathcal{K}|+1)$-th layer can indeed be recovered.
- When $k^* \in K_1+1:K_2$, since $(\tilde{W}_k(\mathcal{K}), \tilde{W}_k(\mathcal{K}\cup\{k^*\})$ are encoded by the same linear $(n_2(\mathcal{K}), k_2(\mathcal{K}))$ MDS code for each $k \in \mathcal{K}$, by item 1 in Lemma 2, that $m(\mathcal{K}) \geq k_2(\mathcal{K})$, the interference signal $\sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K}\cup\{k^*\})$ in the $(|\mathcal{K}|+1)$-th layer can be recovered.

Thus we have $\tilde{W}_{k^*}(\mathcal{K}\cup\{k^*\})$ for all $\mathcal{K} \subset 1:K$, and the desired message W_{k^*} can be recovered.

Privacy: The coding pattern, i.e., the manner of forming pre-coded message sums, is the same for the retrieval of any message W_{k^*}. Specifically, when the identity of the desired message $k^* \in 1:K_1$,

$$n_1(\mathcal{K}) = m(\mathcal{K}) + m(\mathcal{K}\cup\{k^*\}), \tag{30}$$

and when $k^* \in K_1:K_2$,

$$n_2(\mathcal{K}) = m(\mathcal{K}) + m(\mathcal{K}\cup\{k^*\}). \tag{31}$$

Moreover, there are $m(\mathcal{K})$ summations of form $\mathcal{K}$ placed in the $|\mathcal{K}|$-th layer. Thus the placements of the pre-coded message sums are the same for retrieving any message W_{k^*}. For example, there are 4 sums of form $b+c$ in the 2^{nd} layers of both Table III and Table IV. Similarly, the pre-coded sums can be indicated by the set of messages involved, e.g., summations of form $b+c$ are indicated by $\mathcal{K} = \{2,3\}$.

Since it is a linear code, the coded symbols can be generated by the corresponding coding matrices. When $k^* \in 1:K_1$, the desired precoded message $W^*_{k^*}$ is (N^{K_2}, N^{K_2}) MDS coded into $\tilde{W}_{k^*}$; and for each $k \neq k^*$, in the coding group $\mathcal{K} \subset 1:K_2/\{k^*\}$ with $k \in \mathcal{K}$, a non-overlapping segment of W^*_k is the $(n_1(\mathcal{K}), k_1(\mathcal{K}))$ MDS coded where $n_1(\mathcal{K}) : k_1(\mathcal{K}) = N : T_1$. Thus for any $k \in 1:K_2$ the coding matrix of MDS coded symbols $\tilde{W}_k$ in any T_1 servers from the segments of the precoded W^*_k is a $T_1N^{K_2-1} \times T_1N^{K_2-1}$ full rank matrix. By applying Lemma 1, the messages $W_{1:K_1}$ thus have privacy level T_1.

The statement above also implies that for any $k \in 1:K_2$ the coding matrix of MDS coded symbols $\tilde{W}_k$ in any T_2 servers from the segments of the precoded W^*_k is a $T_2N^{K_2-1} \times T_2N^{K_2-1}$ full rank matrix. In addition, when $k^* \in K_1+1:K_2$, the desired precoded message $W^*_{k^*}$ is (N^{K_2}, N^{K_2}) MDS coded into $\tilde{W}_{k^*}$; and for each $k \neq k^*$, in the coding group $\mathcal{K} \subset 1:K_2/\{k^*\}$ with $k \in \mathcal{K}$, a non-overlapping segment of W^*_k is the $(n_2(\mathcal{K}), k_2(\mathcal{K}))$ MDS coded where $n_2(\mathcal{K}) : k_2(\mathcal{K}) = N : T_2$. Thus for any $k^* \in 1:K_2$, the coding matrix of MDS coded symbols $\tilde{W}_k$ in any T_2 servers from the segments of the precoded W^*_k is a $T_2N^{K_2-1} \times T_2N^{K_2-1}$ full rank matrix. By applying Lemma 1, the messages $W_{1:K_2}$ thus have privacy level T_2.

Performance: The message length is $L = N^{K_2}$. The total

TABLE VII
PLACEMENT OF CODING GROUP INDEXED BY $\mathcal{K}$ IN THE $|\mathcal{K}|$-TH AND $(|\mathcal{K}|+1)$-TH LAYERS

Layer	Coding group	Servers $1:N$
$\vdots$	$\vdots$	$\vdots$
$\|\mathcal{K}\|$-th	$\cdots$ coding group $\mathcal{K}$: $(n_i(\mathcal{K}), k_i(\mathcal{K}))$, where $i=1$ if $k^* \in 1:K_1$ otherwise $i=2$	$\cdots$ $m(\mathcal{K})$ symbols: $\sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K})$ $\cdots$
$(\|\mathcal{K}\|+1)$-th	$\cdots$ $\cdots$	$\cdots$ $m(\mathcal{K}\cup\{k^*\})$ symbols: $\tilde{W}_{k^*}(\mathcal{K}) + \sum_{k\in\mathcal{K}} \tilde{W}_k(\mathcal{K}\cup\{k^*\})$ $\cdots$
$\vdots$	$\vdots$	$\vdots$

length of answers is

$$\sum_{n=1}^{N} \ell_n^{[k^*]} = \sum_{\mathcal{K}\subset 1:K_2} m(\mathcal{K}). \tag{32}$$

The rate can thus be computed as

$$R_{\text{NS}} = \frac{L}{\sum_{n=1}^{N} \mathbb{E}[\ell_n^{[k^*]}]} \tag{33}$$

$$= \frac{N}{N} \frac{N^{K_2-1}}{\sum_{i=1}^{K_1} \binom{K_1}{i} d_{i,0} + \sum_{i=0}^{K_1} \sum_{j=1}^{K_2-K_1} \binom{K_1}{i}\binom{K_2-K_1}{j} d_{i,j}} \tag{34}$$

$$= \left(1 + \frac{T_1}{N} + \cdots + \left(\frac{T_1}{N}\right)^{K_1-1} + \left(\frac{T_1}{N}\right)^{K_1} \cdot \left(1 + \frac{T_2}{N} + \cdots + \left(\frac{T_2}{N}\right)^{K_2-K_1-1}\right)\right)^{-1}. \tag{35}$$

Remark: In the general code construction, the message length is N^{K_2}. The message length can be further reduced as long as the length of each non-overlapping segments in Group-wise MDS coding step share a maximum common divisor greater than 1. For the example of the NS scheme for $(N, T_1 : K_1, T_2 : K_2) = (4, 2:2, 1:4)$ two-level PIR we discussed in Section IV-B, the message length $L = 64 = N^{K_2}/4$. It is the same for the NB general scheme we will present in the next section and the example of the NB scheme illustrated in Section IV-C.

VI. THE NON-UNIFORM BLOCK CANCELLATION SCHEME

From the example in Section IV-C, the proposed NB coding scheme uses the T-private code discussed in Section IV-A as base codes, and consists of three blocks. The NS coding scheme studied in the previous section naturally degrades to the T-private code when $K_1 = K$ and $T_1 = T_2 = T$, thus it is leveraged directly in the NB coding scheme. We first construct two precoded tables, which correspond to the NS codes for messages $W_{1:K_1}$ with privacy level T_1 and messages $W_{K_1+1:K_2}$ with privacy level T_2, respectively. Then a portion of the precoded Table-A is placed in the first block of NB code, a portion of the precoded Table-B is placed in the second block, and the rest of both precoded tables are mixed and form the third block.

A. Precoded tables

The message length for the NB coding scheme is $L = N^{K_2}$ for the $(N, T_1 : K_1, T_2 : K_2)$ two-level PIR system. The NS code proposed in Section V for the $(N, T_1 : K_1, T_1 : K_1)$ two-level PIR consists of K_1 layers and has a message length N^{K_1}. Since the message length here is $L = N^{K_2} = N^{K_1} N^{K_2-K_1}$, the NS code can be applied here by stacking the parameters $(m(\cdot), n_1(\cdot), k_1(\cdot), n_2(\cdot), k_2(\cdot))$ by a factor of $N^{K_2-K_1}$. We shall view this coding structure as precoded Table-A. Similarly, define the NS code with message length N^{K_2} for the $(N, T_2 : K_2 - K_1, T_2 : K_2 - K_1)$ two-level PIR with messages $W_{K_1+1:K_2}$ as precoded Table-B.

In the precoded Table-A, there are K_1 layers of precoded sums. The precoded sums can be indicated by the set of messages involved. Here $\tilde{m}(\mathcal{K}_1)$ summations of composition $\mathcal{K}_1$ are placed in the $|\mathcal{K}_1|$-th layer for any non-empty subset $\mathcal{K}_1 \subset 1:K_1$, where

$$\tilde{m}_1(\mathcal{K}_1) = N^{K_2-K_1+1}(N-T_1)^{|\mathcal{K}_1|-1} T_1^{K_1-|\mathcal{K}_1|}. \tag{36}$$

Similarly, there are $K_2 - K_1$ layers in the precoded Table-B, and $\tilde{m}(\mathcal{K}_2)$ summations of compositions $\mathcal{K}_2$ placed in the $|\mathcal{K}_2|$-th layer for any non-empty subset $\mathcal{K}_2 \subset K_1+1:K_2$, where

$$\tilde{m}_2(\mathcal{K}_2) = N^{K_1+1}(N-T_2)^{|\mathcal{K}_2|-1} T_2^{K_2-K_1-|\mathcal{K}_2|}. \tag{37}$$

When the identity of the desired message k^* satisfies $k^* \in K_1+1:K_2$, the precoded Table-B is well-defined, and the precoded Table-A is a *pure-interference table* specified as follows.

1) *Precoding:* Let $S_{1:K_1}$ be K_1 independent random matrices, uniformly drawn from the set of all $N^{K_2} \times N^{K_2}$ full rank matrices over $\mathbb{F}_q$; these matrices are known only to the user. The precoded messages $W^*_{1:K_1}$ are
$$W_k^* = S_k W_k, \quad \forall k \in 1:K_1. \tag{38}$$
2) *Group-wise MDS coding:* The precoded messages are partitioned into non-overlapping segments, and each segment is MDS-coded under certain appropriate parameters. The coding groups are indexed by non-empty sets $\mathcal{K}_1 \subset 1:K_1$. For any non-empty set $\mathcal{K}_1 \subset 1:K_1$, for each $k \in \mathcal{K}_1$, a segment of W_k^* with length $\frac{T_2}{N}\tilde{m}_1(\mathcal{K}_1)$ is $(\tilde{m}_1(\mathcal{K}_1), \frac{T_2}{N}\tilde{m}_1(\mathcal{K}_1))$ MDS-coded into $\tilde{W}_k(\mathcal{K}_1)$.
3) *Forming pre-coded message sums:* From the 1^{st} layer to the K_1-th layer, the summations (vectors) placed in the

i-th layer are formed as

$$\sum_{k\in\mathcal{K}_1} \tilde{W}_k(\mathcal{K}_1), \quad \forall \mathcal{K}_1 \subset 1:K_1 \text{ with } |\mathcal{K}_1| = i, \quad (39)$$

for $i = 1, 2, \ldots, K_1$.

We can similarly define the pure-interference precoded Table-B, with the MDS parameters $(\tilde{m}_2(\mathcal{K}_2), \frac{T_2}{N}\tilde{m}_2(\mathcal{K}_2))$ for any coding group indexed by a nonempty set $\mathcal{K}_2 \subset K_1+1:K_2$.

B. Encoding, decoding, privacy, and performance

Encoding: When the identity of the desired message $k^* \in 1:K_1$, the precoded Table-A is an $N^{K_2-K_1}$-stacked NS code and the precoded Table-B is a pure-interference table. When $k^* \in K_1+1:K_2$, the precoded Table-A is a pure-interference table while the precoded Table-B is an N^{K_1}-stacked NS code. The three blocks of NB code are specified as follows.

In precoded Table-A, there are $\tilde{m}_1(\mathcal{K}_1)$ precoded summations indexed by $\mathcal{K}_1$ for any non-empty set $\mathcal{K}_1 \subset 1:K_1$. For each non-empty set $\mathcal{K}_1 \subset 1:K_1$, $\frac{T_2}{N}$ fractions of the summations indexed by $\mathcal{K}_1$ are placed in the $|\mathcal{K}_1|$-th layer of the first block. Thus a $\frac{T_2}{N}$ fraction of the precoded Table-A forms the first block. Similarly, a $\frac{T_2}{N}$ fraction of the precoded Table-B forms the second block. The remaining $\frac{N-T_2}{N}$ fractions of both tables are mixed together to form the third block by simple pairwise summations in an arbitrary order; in case they have different numbers of remaining coded symbols, these remaining summands are included directly. The summations of each form are partitioned and distributed to N servers uniformly.

Decoding and correctness: When $k^* \in 1:K_1$, the precoded Table-B is a pure-interference table. Since the coding group indexed by non-empty set $\mathcal{K}_2 \subset K_1+1:K_2$ are $(\tilde{m}_2(\mathcal{K}_2), \frac{T_2}{N}\tilde{m}_2(\mathcal{K}_2))$ MDS coded, the $\frac{T_2}{N}$ fraction of the precoded summations placed in the second block can cancel the $\frac{N-T_2}{N}$ fraction of the precoded summations placed in the third block. After canceling all the interference signals involving messages $W_{K_1+1:K_2}$, the NB code becomes precoded Table-A, which can recover the desired message W_{k^*}. Similarly, when $k^* \in K_1+1:K_2$, the precoded Table-A is a pure-interference table, and the MDS parameters (n,k) again satisfy $n:k = N:T_2$. Thus the interference signals in the first block can cancel the interference signals in the third block, and the the desired message W_{k^*} can be recovered by the remaining precoded Table-B.

Privacy: When $k^* \in 1:K_1$, any T_1 of N servers collude may be able to infer the desired message is in $W_{1:K_1}$. However, since the pure-interference precoded Table-B is mixed to the precoded Table-A arbitrarily in the third block, and precoded Table-A has privacy level T_1 for retrieving any message in $W_{1:K_1}$, i.e., even if any T_1 of N servers collude, the identity of the request message W_{k^*} in $W_{1:K_1}$ remains private. It is straightforward to verify that $W_{1:K_2}$ have privacy level T_2 since both precoded tables are T_2-private.

Performance: The message length is $L = N^{K_2}$. The size of precoded Table-A is

$$t_1 = \sum_{\mathcal{K}_1\subset 1:K_1} \tilde{m}_1(\mathcal{K}_1) = \frac{N^{K_1}-T_1^{K_1}}{N-T_1} N^{K_2-K_1+1}; \quad (40)$$

the size of precoded Table-B is

$$t_2 = \sum_{\mathcal{K}_2\subset K_1+1:K_2} \tilde{m}_2(\mathcal{K}_2) = \frac{N^{K_2-K_1}-T_2^{K_2-K_1}}{N-T_2} N^{K_1+1}; \quad (41)$$

and the size of the third block is

$$m = \left(1-\frac{T_2}{N}\right)\max\left(t_1, t_2\right). \quad (42)$$

Since the sizes of the first block and second block are $\frac{T_2}{N}t_1$ and $\frac{T_2}{N}t_2$ separately, the rate is thus

$$R_{\mathrm{NB}} = \frac{L}{\sum_{n=1}^{N}\mathbb{E}[\ell_n^{[k^*]}]} = \frac{L}{\frac{T_2}{N}t_1 + \frac{T_2}{N}t_2 + m}, \quad (43)$$

which is indeed (12) after elementary simplification.

VII. Conclusion

We considered two-level private information retrieval systems, and provided a capacity lower bound by proposing two novel code constructions and a capacity upper bound. It is further shown that the upper bound can be improved in a special case, however the improved bound also does not match the proposed lower bound. We suspect the proposed code constructions can also be improved to yield better lower bounds, which we leave as a future work. Some of the techniques given in this work can be adopted to multilevel PIR with more than two privacy levels, and when storage constraint is introduced. The two-level model can be viewed as natural generalization of the canonical PIR model. In addition to the extensions and generalizations we discussed in the introduction section, there have been other PIR models in the literature, such as private computation [27], PIR with side information [28], and weakly private information retrieval [26]. The multilevel privacy model we proposed here can also be further extended to such scenarios.

References

[1] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4075–4088, 2017.

[2] ——, "The capacity of robust private information retrieval with colluding databases," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2361–2370, 2017.

[3] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb, "Private information retrieval schemes for coded data with arbitrary collusion patterns," in *Proc. 2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1908–1912, 2017.

[4] X. Yao, N. Liu, and W. Kang, "The capacity of private information retrieval under arbitrary collusion patterns," *IEEE Transactions on Information Theory*, vol. 67, no. 10, pp. 6841–6855, 2021.

[5] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 7613–7627, 2019.

[6] H. Sun and 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.

[7] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from mds-coded databases with minimum message size," *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 4904–4916, 2020.

[8] K. Banawan and S. Ulukus, "Asymmetry hurts: Private information retrieval under asymmetric traffic constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 7628–7645, 2019.

[9] H.-Y. Lin, S. Kumar, E. Rosnes, and A. G. i Amat, "Asymmetry helps: Improved private information retrieval protocols for distributed storage," in *Proc. 2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2018.

[10] S. Vithana, K. Banawan, and S. Ulukus, "Semantic private information retrieval," *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2635–2652, 2022.

[11] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1945–1956, 2018.

[12] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM Journal on Applied Algebra and Geometry*, vol. 1, no. 1, pp. 647–664, 2017.

[13] S. Kumar, H.-Y. Lin, E. Rosnes, and A. G. i Amat, "Achieving maximum distance separable private information retrieval capacity with linear codes," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4243–4273, 2019.

[14] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from uncoded storage constrained databases," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 6617–6634, 2020.

[15] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus, "The capacity of private information retrieval from heterogeneous uncoded caching databases," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3407–3416, 2020.

[16] N. Woolsey, R.-R. Chen, and M. Ji, "Uncoded placement with linear sub-messages for private information retrieval from storage constrained databases," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6039–6053, 2020.

[17] C. Tian, H. Sun, and J. Chen, "A Shannon-theoretic approach to the storage-retrieval tradeoff in pir systems," in *Proc. 2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1904–1908, 2018.

[18] H. Sun and C. Tian, "Breaking the MDS-PIR capacity barrier via joint storage coding," *Information*, vol. 10, no. 9, p. 265, 2019.

[19] C. Tian, "On the storage cost of private information retrieval," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7539–7549, 2020.

[20] T. Guo, R. Zhou, and C. Tian, "New results on the storage-retrieval tradeoff in private information retrieval systems," *IEEE Transactions on Information Theory*, vol. 2, no. 1, pp. 403–414, 2021.

[21] H. Sun and S. A. Jafar, "Private information retrieval from mds coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al." *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1000–1022, 2017.

[22] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Transactions on Information Theory*, vol. 64, no. 10, pp. 6842–6862, 2018.

[23] C. Tian, "Symmetry, outer bounds, and code constructions: A computer-aided investigation on the fundamental limits of caching," *Entropy*, vol. 20, no. 8, 603.1–35, 2018.

[24] C. Tian, J. S. Plank, B. Hurst, and R. Zhou, "User manual CAI version-1.0: An open-source toolbox for computer-aided investigation on the fundamental limits of information systems," *arXiv preprint arXiv:1910.08567v2*, 2019.

[25] C. Tian, J. S. Plank, B. Hurst, and R. Zhou, "Computational techniques for investigating information theoretic limits of information systems," *Information*, vol. 12, no. 1, 82.1–16, 2020.

[26] I. Samy, M. Attia, R. Tandon, Ravi and L. Lazos, "Asymmetric leaky private information retrieval," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5352–5369, 2021.

[27] H. Sun, S.A. Jafar, "The capacity of private computation," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp.3880-3897, 2018.

[28] A. Heidarzadeh, F. Kazemi and A. Sprintson, "The Role of Coded Side Information in Single-Server Private Information Retrieval," *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 25–44, 2021.