The Capacity of Private Information Retrieval From Coded Databases

Karim Banawan, *Student Member, IEEE*, and Sennur Ulukus, *Fellow, IEEE*

***Abstract*— We consider the problem of private information retrieval (PIR) over a distributed storage system. The storage system consists of N non-colluding databases, each storing an MDS-coded version of M messages. In the PIR problem, the user wishes to retrieve one of the available messages without revealing the message identity to any individual database. We derive the information-theoretic capacity of this problem, which is defined as the maximum number of bits of the desired message that can be privately retrieved per one bit of downloaded information. We show that the PIR capacity in this case is $C = (1 + K/N + K^2/N^2 + \cdots + K^{M-1}/N^{M-1})^{-1} = (1 + R_c + R_c^2 + \cdots + R_c^{M-1})^{-1} = (1 - R_c)/(1 - R_c^M)$, where R_c is the rate of the (N, K) MDS code used. The capacity is a function of the code rate and the number of messages only regardless of the explicit structure of the storage code. The result implies a fundamental tradeoff between the optimal retrieval cost and the storage cost when the storage code is restricted to the class of MDS codes. The result generalizes the achievability and converse results for the classical PIR with replicated databases to the case of MDS-coded databases.**

***Index Terms*— Private information retrieval, distributed storage, MDS code, capacity, alignment.**

I. Introduction

PROTECTING the privacy of downloaded information from curious publicly accessible databases has been the focus of considerable research within the computer science community [1]–[4]. Practical examples for this problem include: ensuring privacy of investors upon downloading records in a stock market, and ensuring the privacy of activists against authoritarian regimes while browsing restricted contents from the internet, see [1], [5]. In the seminal paper of Chor *et al.* [1], the classical problem of private information retrieval (PIR) is introduced. In the classical PIR setting, a user requests to download a certain message (or file) from N non-communicating databases without leaking the identity of the message to any individual database. The contents of these databases are identical, i.e., they are coded using a repetition code. A trivial solution for this seemingly challenging task is to download all of the contents of the databases. However, this solution is highly impractical, especially for large number of messages, which is the case in modern storage systems. The aim of the PIR problem is to design efficient retrieval schemes that maximize the ratio of the desired information bits to the total downloaded bits under the privacy constraint.

In the classical PIR problem, the user prepares N queries each directed to a specific database. The queries are designed such that they do not reveal any information about the identity of the desired message. Upon receiving these queries, the databases respond truthfully with answering strings. Based on the collected answer strings, the user reconstructs the desired message. In the original formulation of the problem in the computer science literature [1], the messages are assumed to have a size of one bit. In this formulation, the performance metric was the sum of lengths of the answer strings (download cost) and the size of the queries (upload cost). The information-theoretic reformulation of the problem assumes that the messages are of arbitrarily large size and hence the upload cost can be neglected with respect to the download cost [6]. The pioneering work [7] derives the exact capacity of the classical PIR problem. The capacity is defined as the maximum number of bits of the desired message per bit of total download. The achievable scheme is based on an interesting relationship between PIR and blind interference alignment introduced for wireless networks in [8] as observed in [9]. Reference [10] extends this setting to the case of T colluding databases with and without node failures. The main difference from the non-colluding case is that the user asks for MDS-coded versions of the contents of the databases. Another interesting extension of the problem is symmetric PIR [11], in which the privacy of the undesired messages needs to be preserved against the user as well.

Due to node failures and erasures that arise naturally in any storage system, redundancy should be introduced [12]. The simplest form of redundancy is repetition coding. Although repetition coding across databases offers the highest immunity against erasures and simplicity in designing PIR schemes, it results in extremely large storage cost. This motivates the use of erasure coding techniques that achieve the same level of reliability with less storage cost. A common erasure coding technique is the MDS code that achieves the optimal redundancy-reliability tradeoff. An (N, K) MDS code maps K sub-packets of data into N sub-packets of coded data. This code tolerates upto $N - K$ node failures (or erasures). By connecting to any K storage nodes, the node failure can be repaired. Despite the ubiquity of work on the classical

Manuscript received September 23, 2016; revised November 8, 2017; accepted December 22, 2017. Date of publication January 11, 2018; date of current version February 15, 2018. This work was supported by NSF under Grant CNS 13-14733, Grant CCF 14-22111, Grant CCF 14-22129, and Grant CNS 15-26608. This paper was presented in part at the 2017 IEEE International Conference on Communications.

The authors are with the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD 20742 USA (e-mail: kbanawan@umd.edu; ulukus@umd.edu).

Communicated by M. Bloch, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2791994

PIR problem, little research exists for coded PIR with a few exceptions: [13] which has initiated the work on coded databases and has designed an explicit erasure code and PIR algorithm that requires only one extra bit of download to provide perfect privacy. The result is achieved at the expense of having the number of storage nodes N grow with the message size. Reference [6] considers a general formulation for the coded PIR problem, and obtains a tradeoff between storage and retrieval costs based on certain sufficient conditions. Reference [5] presents the best-known achievable scheme for the MDS-coded PIR problem, which achieves a retrieval rate of $R = 1-R_c$, where R_c is the code rate of the storage system. The scheme is universal in that it depends only on the code rate. Finally, [14] investigates the problem from the storage overhead perspective and shows that information-theoretic PIR can be achieved with storage overhead arbitrarily close to the optimal value of 1 by proposing new binary linear codes called the k-server PIR codes.

In this paper, we consider the PIR problem for non-colluding and coded databases. We use the information-theoretic formulation. We formulate the problem such that each message is a matrix that consists of K columns, while the number of rows can grow arbitrarily large to conform with the conventional information-theoretic arguments. We assume that the contents of the databases are coded using a linear (N, K) storage code with a generator matrix $\mathbf{H}$. We do not assume any specific structure on the generator matrix of the distributed storage code other than the linear independence of every K columns[1] and that the encoding is performed independently over the rows, i.e., the rows/messages are not mixed.[2,3] This condition is equivalent to restricting the storage code structure to MDS codes. Note also that the dimensions of the generator matrix (N, K) are not design parameters that can grow with the message size as in [13]. This formulation includes the models of [7] and [5] as special cases. We show that the exact PIR capacity in this case is given by $C = \left(1 + \frac{K}{N} + \frac{K^2}{N^2} + \cdots + \frac{K^{M-1}}{N^{M-1}}\right)^{-1} = (1 + R_c + R_c^2 + \cdots + R_c^{M-1})^{-1} = \frac{1-R_c}{1-R_c^M}$. The PIR capacity depends only on the code rate R_c and the number of messages M irrespective of the generator matrix structure or the number of nodes. Surprisingly, the result implies the optimality of separation between the design of the PIR scheme and the MDS storage code for a fixed code rate. The result outperforms the best-known lower bound in [5]. The result reduces to the repetition-coded case (which is a special case of MDS codes) in [7] by observing that $R_c = \frac{1}{N}$ in that case. The achievable scheme is similar to the scheme in [7] with extra steps that entail decoding of the interference and the desired message by solving K linearly independent equations. The converse proof hinges on the fact that the contents of any K storage nodes are independent and hence the answer strings in turn are independent. We present two lemmas that capture the essence of the converse proof, namely, interference lower bound lemma and induction lemma. The proof of the induction lemma uses Han's inequality to lower bound the entropy of any K answer strings. These lemmas generalize the converse technique in [7, Lemmas 5 and 6] to account for MDS coding. By applying the two lemmas successively for $M-1$ times, we derive an explicit upper bound on the retrieval rate for the PIR problem from MDS-coded databases. A different converse proof that uses the assumption that the answer strings are symmetric without loss of generality can be found in the conference version of this paper [15].

II. System Model

Consider an (N, K) MDS-coded distributed storage system storing M messages (or files). The messages are independent and identically distributed with

$$H(W_i) = L, \quad i \in \{1, \cdots, M\} \tag{1}$$

$$H(W_1, W_2, \cdots, W_M) = ML \tag{2}$$

The message W_i, $i \in \{1, \cdots, M\}$ is a $\mathbb{F}_q^{\tilde{L}\times K}$ matrix with sufficiently large field $\mathbb{F}_q$, such that $\tilde{L} \times K = L$. The elements of W_i are picked uniformly and independently from $\mathbb{F}_q$. We denote the jth row of message W_i by $\mathbf{w}_j^{[i]} \in \mathbb{F}_q^K$. The generator matrix of the (N, K) storage code $\mathbf{H}$ is a $\mathbb{F}_q^{K\times N}$ matrix such that

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_N \end{bmatrix}_{K\times N} \tag{3}$$

where $\mathbf{h}_i \in \mathbb{F}_q^K$, $i \in \{1, \cdots, N\}$.[4] For an MDS code, any set $\mathcal{K}$ of columns of $\mathbf{H}$ such that $|\mathcal{K}| \leq K$ are linearly independent.

[1] For the converse proof, the linear independence requirement of every K columns in $\mathbf{H}$ is not strictly needed. In fact, from the converse point of view, any storage code that enforces the contents of every K databases to be statistically independent leads to the same upper bound even if the code is not linear. In this paper, the linear independence assumption, which is equivalent to having an MDS code, is important for the construction of the achievable scheme (see Section IV) that relies on solving K linear equations, in addition to creating an instance of statistical independence that is needed in the converse proof.

[2] By *non-mixing* MDS code, we mean that each message is encoded separately. Furthermore, we assume that each row within each message is encoded separately as well. This assumption is made to enable the MDS code to be flexible enough so that the code structure makes sense for every message size L, which is needed to characterize the capacity in the Shannon sense (i.e., as $L \to \infty$). Here we give a concrete example: if $W_1 = (a_1, \cdots, a_4)$, and $W_2 = (b_1, \cdots, b_4)$ and they are encoded via a $(3, 2)$ non-mixing MDS code, then each message is arranged in 2 rows. Each row is encoded separately, for example, row 1 is encoded as $(a_1, a_2, a_1 + a_2)$, and row 2 is encoded as $(a_3, a_4, a_3 + a_4)$, and similarly for W_2. Note that this example MDS code neither mixes messages, nor the rows of each message. The results of this paper are restricted to such non-mixing code structures and hence the qualifier "non-mixing" is dropped.

[3] We note that in [6, Example 2], an example for a mixing $(3, 2)$ MDS code for $M = 2$ is presented. In this case, letting $W_1 = (a_1, a_2)$, $W_2 = (b_1, b_2)$, database 1 stores (a_1, a_2), database 2 stores (b_1, b_2) and database 3 stores $(a_1 + b_1, a_2 + b_2)$. This code mixes W_1, W_2 in database 3. Reference [6] provides a retrieval scheme for this specific code that achieves a retrieval rate of $\frac{2}{3}$, which is higher than the capacity of non-mixing $(3, 2)$ MDS codes ($C = \frac{3}{5}$). The characterization of the capacity of mixing MDS codes is an interesting open problem.

[4] We note that the assumption of encoding each row with the same generator matrix is indeed without loss of generality and is made to simplify the presentation. If each row is encoded via a different MDS generator matrix, i.e., the jth row of message i is encoded via $\mathbf{H}_j^{[i]}$, the capacity is still given by Theorem 1. For the achievable scheme, we note that the scheme downloads K coded symbols directly from the databases with no further processing. This suffices to decode the entire row because the MDS property is still valid for each row. The converse proof still holds since the contents of every K databases are statistically independent and hence Lemma 1 is still valid.

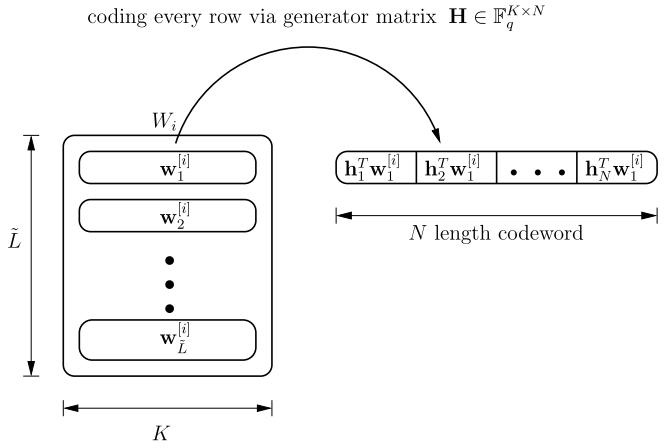

Fig. 1. Coding process for message W_i.

The storage code $f_n : \mathbf{w}_j^{[i]} \to y_{n,j}^{[i]}$ on the nth database maps each row of W_i separately into coded bit $y_{n,j}^{[i]}$, see Fig. 1,

$$y_{n,j}^{[i]} = \mathbf{h}_n^T \mathbf{w}_j^{[i]} \tag{4}$$

Consequently, the stored bits $\mathbf{y}_n \in \mathbb{F}_q^{M\tilde{L}}$ on the nth database, $n \in \{1, \cdots, N\}$ are concatenated projections of all messages $\{W_1, \cdots, W_M\}$ and are given by

$$\mathbf{y}_n = \begin{bmatrix} W_1 \\ \vdots \\ W_M \end{bmatrix} \mathbf{h}_n \tag{5}$$

$$= \Big[\mathbf{h}_n^T\mathbf{w}_1^{[1]} \ \cdots \ \mathbf{h}_n^T\mathbf{w}_{\tilde{L}}^{[1]} \ \mathbf{h}_n^T\mathbf{w}_1^{[2]} \ \cdots \ \mathbf{h}_n^T\mathbf{w}_{\tilde{L}}^{[2]} \ \cdots \ \mathbf{h}_n^T\mathbf{w}_1^{[M-1]} \ \cdots \ \mathbf{h}_n^T\mathbf{w}_{\tilde{L}}^{[M-1]} \ \mathbf{h}_n^T\mathbf{w}_1^{[M]} \ \cdots \ \mathbf{h}_n^T\mathbf{w}_{\tilde{L}}^{[M]}\Big]^T \tag{6}$$

The explicit structure of the coded storage system is illustrated in Table I.

The described storage code can tolerate up to $N-K$ errors by connecting to any K databases. Thus, we have for any set $\mathcal{K}$ such that $|\mathcal{K}| \geq K$,

$$H(\mathbf{y}_{\bar{\mathcal{K}}}|\mathbf{y}_{\mathcal{K}}) = 0 \tag{7}$$

where $\mathbf{y}_{\mathcal{K}}$ are the stored bits on databases indexed by $\mathcal{K}$, and $\bar{\mathcal{K}}$ is the complement of the set $\mathcal{K}$. The code rate of this distributed storage system R_c is given by

$$R_c = \frac{K}{N} \tag{8}$$

The retrieval process over MDS-coded databases is illustrated in Fig. 2. To retrieve W_i, the user generates a query $Q_n^{[i]}$ and sends it to the nth database. Since the user does not have knowledge about the messages in advance, the queries are independent of the messages,

$$I(Q_1^{[i]}, \cdots, Q_N^{[i]}; W_1, \cdots, W_M) = 0 \tag{9}$$

In order to ensure privacy, the retrieval strategy for the ith message should be indistinguishable from the retrieval strategy of W_1, hence, for $i \in \{1, \cdots, M\}$, $n \in \{1, \cdots N\}$

$$(Q_n^{[i]}, A_n^{[i]}, W_1, \cdots, W_M) \sim (Q_n^{[1]}, A_n^{[1]}, W_1, \cdots, W_M) \tag{10}$$

TABLE I

EXPLICIT STRUCTURE OF (N, K) CODE FOR DISTRIBUTED DATABASES WITH M MESSAGES

	DB1 ($\mathbf{y}_1$)	DB2 ($\mathbf{y}_2$)	$\cdots$	DBN ($\mathbf{y}_N$)
message 1	$\mathbf{h}_1^T\mathbf{w}_1^{[1]}$ $\mathbf{h}_1^T\mathbf{w}_2^{[1]}$ $\vdots$ $\mathbf{h}_1^T\mathbf{w}_{\tilde{L}}^{[1]}$	$\mathbf{h}_2^T\mathbf{w}_1^{[1]}$ $\mathbf{h}_2^T\mathbf{w}_2^{[1]}$ $\vdots$ $\mathbf{h}_2^T\mathbf{w}_{\tilde{L}}^{[1]}$	$\cdots$ $\cdots$ $\cdots$ $\cdots$	$\mathbf{h}_N^T\mathbf{w}_1^{[1]}$ $\mathbf{h}_N^T\mathbf{w}_2^{[1]}$ $\vdots$ $\mathbf{h}_N^T\mathbf{w}_{\tilde{L}}^{[1]}$
message 2	$\mathbf{h}_1^T\mathbf{w}_1^{[2]}$ $\mathbf{h}_1^T\mathbf{w}_2^{[2]}$ $\vdots$ $\mathbf{h}_1^T\mathbf{w}_{\tilde{L}}^{[2]}$	$\mathbf{h}_2^T\mathbf{w}_1^{[2]}$ $\mathbf{h}_2^T\mathbf{w}_2^{[2]}$ $\vdots$ $\mathbf{h}_2^T\mathbf{w}_{\tilde{L}}^{[2]}$	$\cdots$ $\cdots$ $\cdots$ $\cdots$	$\mathbf{h}_N^T\mathbf{w}_1^{[2]}$ $\mathbf{h}_N^T\mathbf{w}_2^{[2]}$ $\vdots$ $\mathbf{h}_N^T\mathbf{w}_{\tilde{L}}^{[2]}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
message M	$\mathbf{h}_1^T\mathbf{w}_1^{[M]}$ $\mathbf{h}_1^T\mathbf{w}_2^{[M]}$ $\vdots$ $\mathbf{h}_1^T\mathbf{w}_{\tilde{L}}^{[M]}$	$\mathbf{h}_2^T\mathbf{w}_1^{[M]}$ $\mathbf{h}_2^T\mathbf{w}_2^{[M]}$ $\vdots$ $\mathbf{h}_2^T\mathbf{w}_{\tilde{L}}^{[M]}$	$\cdots$ $\cdots$ $\cdots$ $\cdots$	$\mathbf{h}_N^T\mathbf{w}_1^{[M]}$ $\mathbf{h}_N^T\mathbf{w}_2^{[M]}$ $\vdots$ $\mathbf{h}_N^T\mathbf{w}_{\tilde{L}}^{[M]}$

which implies that the queries and answers should be independent of the desired message index i, i.e., the privacy constraint is,

$$I(Q_n^{[i]}, A_n^{[i]}, W_1, \cdots, W_M; i) = 0, \quad n \in \{1, \cdots, N\} \tag{11}$$

Each database responds with an answer string $A_n^{[i]}$, which is a deterministic function[5] of the received query and the stored coded bits in the nth database. Hence, by the data processing inequality,

$$H(A_n^{[i]}|Q_n^{[i]}, \mathbf{y}_n) = H(A_n^{[i]}|Q_n^{[i]}, W_1, \cdots, W_M) = 0 \tag{12}$$

In addition, the user should be able to decode W_i reliably from all the answer strings collected from the N databases with a small probability of error. Consequently, from Fano's inequality, we have the following reliability constraint,

$$H(W_i|A_1^{[i]}, \cdots, A_N^{[i]}, Q_1^{[i]}, \cdots, Q_N^{[i]}) = o(L) \tag{13}$$

where $\frac{o(L)}{L} \to 0$ as $L \to \infty$. The retrieval rate R for the PIR problem is the ratio of the size of the desired message to the total download cost under the reliability constraint (13) and the privacy constraint (10) for some $L \in \mathbb{N}$, i.e.,

$$R = \frac{H(W_i)}{\sum_{n=1}^{N} H(A_n^{[i]})}, \quad \text{subject to (10), (13)} \tag{14}$$

[5]We note that the assumption that the answer strings are deterministic functions of the queries and the stored information is indeed without loss of generality and is kept for the simplicity of presentation. The converse proof can be extended to the case of allowing the databases to use randomized strategies. In this case, a common randomness should be shared between the user and the databases. More specifically, we can assume that there exists a random variable $\mathbb{G}$ that is shared between the user and the databases such that $\mathbb{G}$ is independent of $(i, W_{1:M})$, and $H(A_n^{[i]}|Q_n^{[i]}, \mathbf{y}_n, \mathbb{G}) = 0$. This does not change the converse lemmas except for conditioning all inequalities on $\mathbb{G}$. A similar formulation of this idea can be found in [16].

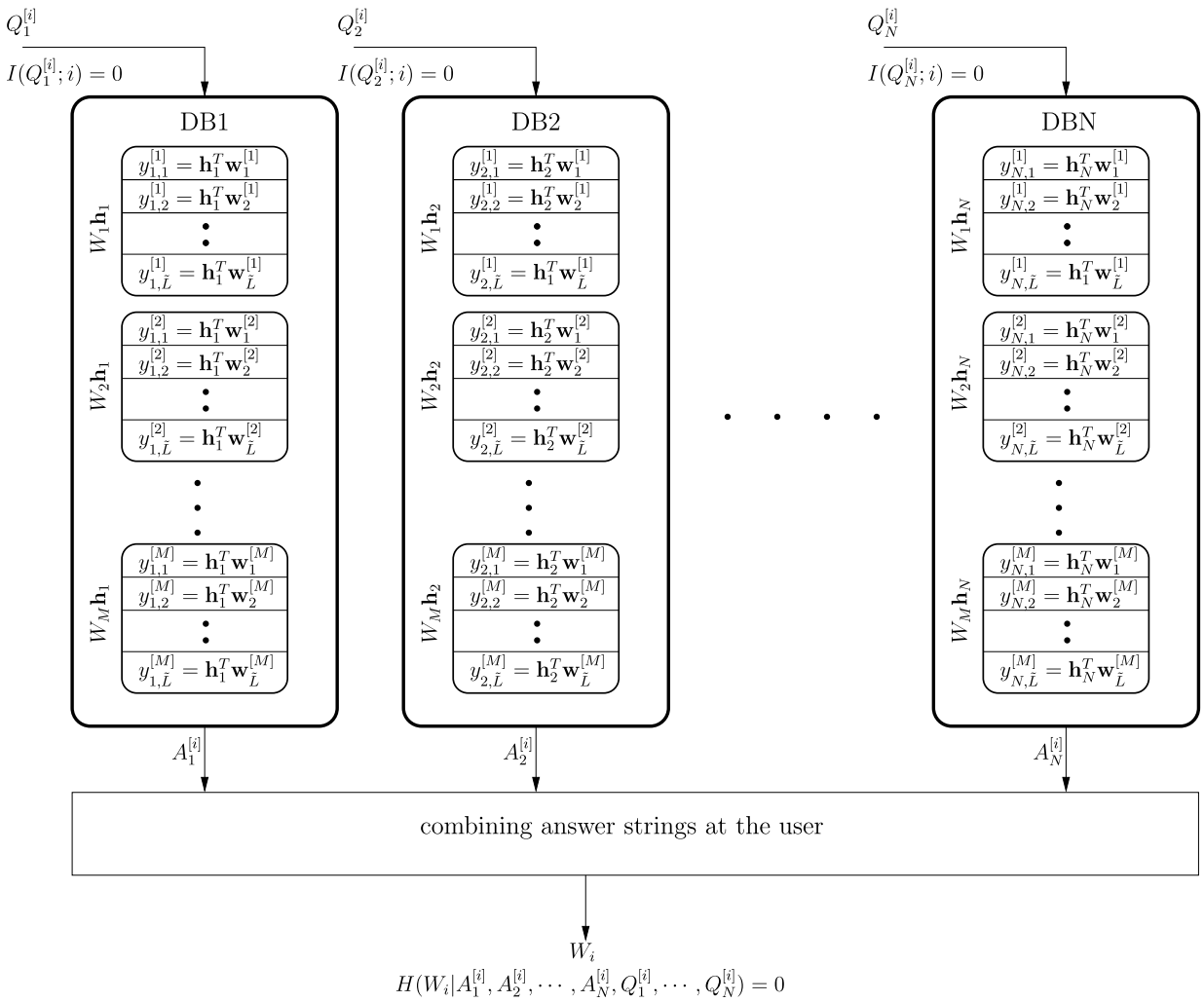

Fig. 2. System model of the coded PIR problem.

The PIR capacity C is the supremum of R over all retrieval schemes as $L \to \infty$.

In this paper, as in [7], we follow a Shannon theoretic formulation by assuming that the message size can be arbitrarily large. Also, we neglect the upload cost with respect to the download cost as in [7].

We note that the described storage code is a generalization of the repetition-coded problem in [7]. If $K = 1$ and $h_n = 1$, $n \in \{1, \cdots, N\}$, then the problem reduces to the classical PIR in [7]. In addition, the systematic MDS-coded instance[6] presented in [5] is a special case of this setting with $\mathbf{h}_n = \mathbf{e}_n$, $n \in \{1, \cdots, K\}$, where $\mathbf{e}_n$ is the nth standard basis vector.

III. Main Result

Theorem 1: For an (N, K) MDS-coded distributed database system with coding rate $R_c = \frac{K}{N}$ and M messages, the PIR capacity is given by

$$C = \frac{1 - R_c}{1 - R_c^M} \tag{15}$$

$$= \frac{1}{1 + R_c + \cdots + R_c^{M-1}} \tag{16}$$

$$= \left(1 + \frac{K}{N} + \frac{K^2}{N^2} + \cdots + \frac{K^{M-1}}{N^{M-1}}\right)^{-1} \tag{17}$$

We have the following remarks about the main result. We first note that the PIR capacity in (15) is a function of the coding rate R_c and the number of messages M only, and does not depend on the explicit structure of the coding scheme (i.e., the generator matrix) or the number of databases. This observation implies the universality of the scheme over any MDS-coded database system with the same coding rate and number of messages. The result also entails the optimality of separation between distributed storage code design and PIR scheme design for a fixed R_c. We also note that the capacity C decreases as R_c increases. As $R_c \to 0$, the PIR capacity approaches $C = 1$. On the other hand, as $R_c \to 1$, the PIR capacity approaches $\frac{1}{M}$ which is the trivial retrieval rate obtained by downloading the contents of all databases.

[6]We note that although the code structure presented in [5] is assumed to be systematic, this assumption is indeed without loss of generality. The scheme in [5] is universal and can be applied for any (N, K) MDS code and was presented for systematic MDS codes for sake of simpler exposition of the retrieval scheme.

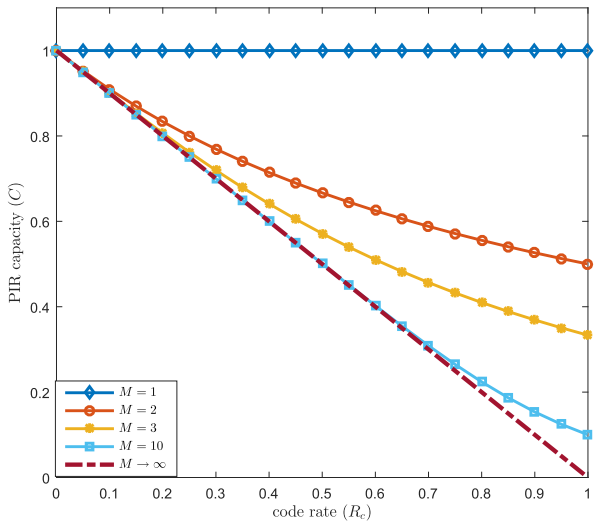

Fig. 3. PIR capacity versus R_c.

This observation implies that a fundamental tradeoff exists between storage cost and the retrieval download cost when the storage code is restricted to the class of MDS codes. This tradeoff conforms with the result of [6]. The capacity expression in Theorem 1 is plotted in Fig. 3 as a function of the code rate R_c for various numbers of messages M.

The capacity in (15) is strictly larger than the best-known achievable rate in [5], where $R = 1 - R_c$ for any finite number of messages. We observe also that the PIR capacity for a given fixed code rate R_c is monotonically decreasing in M. The rate in (15) converges to $1 - R_c$ as $M \to \infty$. Intuitively, as the number of messages increases, the undesired download rate must increase to hide the identity of the desired message; eventually, the gain from applying the greedy algorithm in Section IV over the scheme in [5] diminishes. This confirms that the achievable scheme in [5] is asymptotically optimal. Our capacity here generalizes the capacity in [7] where $R_c = \frac{1}{N}$. That is, the classical PIR problem may be viewed as a special case of the MDS-coded PIR problem with a specific code structure which is repetition coding.

IV. Achievability Proof

In this section, we present the general achievable scheme for Theorem 1. We give a few specific examples in Section V. Our achievable scheme generalizes the achievable scheme in [7] which induces symmetry across databases and symmetry across messages, and exploits the side information. The achievable scheme here includes two extra steps due to the presence of coding: decoding of the interference and decoding of the desired rows which are not present in [7].

A. Achievable Scheme

The scheme requires $\tilde{L} = N^M$, which implies that the size of message $H(W_i) = L = KN^M$. The scheme is completed in M rounds, each corresponding to the sum of i terms, $i \in \{1, \cdots, M\}$, and is repeated K times to decode the desired message; see Tables II and III for examples.

1) *Index preparation:* The user interleaves the indices of rows for all messages randomly and independently from each other, i.e., for any message W_ℓ,
$$\mathbf{x}_i^{[\ell]} = \mathbf{w}_{\pi_\ell(i)}^{[\ell]}, \quad i \in \{1, \cdots, \tilde{L}\} \tag{18}$$
where $\pi_\ell(\cdot)$ is a random interleaver used for message ℓ and known privately to the user only. In this case the rows chosen at any database appear to be chosen at random and independent from the desired message index.
2) *Initialization:* The user downloads K^{M-1} desired coded bits from different rows of the desired message W_m from database 1 (DB1) and sets round index $i = 1$, i.e., the user starts by downloading the symbols $\mathbf{h}_1^T \mathbf{x}_1^{[m]}, \cdots, \mathbf{h}_1^T \mathbf{x}_{K^{M-1}}^{[m]}$ from database 1.
3) *Symmetry across databases:* The user downloads K^{M-1} desired bits each from a different row from each database, i.e., the user downloads from database 2 the symbols $\mathbf{h}_2^T \mathbf{x}_{K^{M-1}+1}^{[m]}, \cdots, \mathbf{h}_2^T \mathbf{x}_{2K^{M-1}}^{[m]}$, from database 3 the symbols $\mathbf{h}_3^T \mathbf{x}_{2K^{M-1}+1}^{[m]}, \cdots, \mathbf{h}_3^T \mathbf{x}_{3K^{M-1}}^{[m]}$, $\cdots$, similarly until the user downloads $\mathbf{h}_N^T \mathbf{x}_{(N-1)K^{M-1}+1}^{[m]}, \cdots, \mathbf{h}_N^T \mathbf{x}_{NK^{M-1}}^{[m]}$ from database N. Then, the total number of desired bits in the ith round is NK^{M-1}.
4) *Message symmetry:* To satisfy the privacy constraint, the user needs to download an equal amount of coded bits from all other messages. Consequently, the user downloads $\binom{M-1}{i} K^{M-i}(N-K)^{i-1}$ bits from each database. The undesired equation is a sum of i terms picked from the remaining undesired messages. To be more specific, the user downloads the sum $\mathbf{h}_n^T(\mathbf{x}_{j_1}^{[\ell_1]} + \mathbf{x}_{j_2}^{[\ell_2]} + \cdots + \mathbf{x}_{j_i}^{[\ell_i]})$ from the rows $j_1, \cdots, j_i \in \{1, \cdots, \tilde{L}\}$ of messages $\ell_1, \cdots, \ell_i \in \{1, \cdots, M\} \setminus m$ from the nth database. The specification of rows will become clear in step 5. Hence, the number of undesired equations downloaded in the ith round is $N\binom{M-1}{i} K^{M-i}(N-K)^{i-1}$.
5) *Decoding the interference:* The main difference of the coded problem from the uncoded PIR (i.e., repetition-coded counterpart) is that in order to exploit the undesired coded bits in the form of side information, the interference needs to be decoded first. Note that we are not interested in decoding the individual components of each term of the sum, but rather the components of the *aligned sum*. To perform this, we group each K undesired equations to be from the same rows, i.e., the user downloads the same sum from the rows $j_1, \cdots, j_i \in \{1, \cdots, \tilde{L}\}$ of messages $\ell_1, \cdots, \ell_i \in \{1, \cdots, M\} \setminus m$ as $\mathbf{h}_{n \bmod N}^T(\mathbf{x}_{j_1}^{[\ell_1]} + \mathbf{x}_{j_2}^{[\ell_2]} + \cdots + \mathbf{x}_{j_i}^{[\ell_i]}), \cdots, \mathbf{h}_{n+K-1 \bmod N}^T(\mathbf{x}_{j_1}^{[\ell_1]} + \mathbf{x}_{j_2}^{[\ell_2]} + \cdots + \mathbf{x}_{j_i}^{[\ell_i]})$. The rows are chosen in order starting from row 1, and the index of the row is incremented whenever K symbols from the same row is downloaded. For example: the user downloads $\mathbf{h}_1^T \mathbf{x}_1^{[\ell]}$ from the undesired message ℓ from database 1, then the user downloads $\mathbf{h}_2^T \mathbf{x}_1^{[\ell]}$ from database 2, $\cdots$, until the user downloads $\mathbf{h}_K^T \mathbf{x}_1^{[\ell]}$ from database K. Starting from this

point the user increments the index of the row to 2 and downloads $\mathbf{h}_{K+1}^T \mathbf{x}_2^{[\ell]}$ from database $K+1$, and so on. In this case, we have K linearly independent equations that can be uniquely solved, and hence the corresponding row of the interfering messages is decoded due to (7). Therefore, this generates $N\binom{M-1}{i}K^{M-(i+1)}(N-K)^{i-1}$ side information equations in the form of i term sums.

6) *Exploiting side information:* The side information generated in the previous step can be exploited in the $(i+1)$th round within the remaining $N-K$ databases that did not participate in generating them. The side information is used in $i+1$ term sum that includes the desired message as one of the terms. Since side information is successfully decoded, it can be canceled from these equations to leave desired coded bits. Hence, we can download $N\binom{M-1}{i}K^{M-(i+1)}(N-K)^{i}$ extra desired coded bits. More specifically, the user downloads the sums $\mathbf{h}_{n_1(n)}^T(\mathbf{x}_{\theta_1}^{[m]}+\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]})$, $\cdots$, $\mathbf{h}_{n_{N-K}(n)}^T(\mathbf{x}_{\theta_{N-K}}^{[m]}+\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]})$ from databases $n_1(n)=n+K \mod N$, $\cdots$, $n_{N-K}(n)=n+N-1 \mod N$ in the $(i+1)$th round, where $\mathbf{x}_{\theta_l}^{[m]}$ is the row θ_l from the desired message W_m, i.e., the user downloads the sum of the row from the desired message to the side information generated in the ith round.

7) Repeat steps 4, 5, 6 after setting $i=i+1$ until $i=M-1$.

8) *Decoding the desired message:* Till this point the scheme has downloaded one bit from each row of the desired message. To reliably decode the desired message, the scheme (precisely steps 2-7) is repeated K times. We repeat the scheme exactly except for shifting the order of databases circularly at each repetition for the desired coded bits. Note that the chosen indices for the desired message is the same up to circular shift at each repetition, however we download new undesired coded bits at each repetition. This creates K different equations for each row of the message and hence decodable.

9) *Shuffling the order of queries:* Since all databases know the retrieval scheme, every database can identify the desired message by observing the first query only. By shuffling the order of queries uniformly, all possible queries can be made equally likely regardless of the message index. This guarantees the privacy.

B. Decodability, Privacy, and Calculation of the Achievable Rate

1) Decodability: The decodability follows from the MDS property of the storage code, which states that in a $K \times N$ MDS generator matrix, any $K \times K$ submatrix is invertible. To show decodability formally, let W_m be the desired message without loss of generality. In each repetition, at the ith round, the user downloads $\binom{M-1}{i}K^{M-i}(N-K)^{i-1}$ symbols from the undesired messages from every database. These coded symbols are constructed as the sums of i coded symbols from some rows, i.e., the user downloads the sum $\mathbf{h}_n^T(\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]})$ from the rows $j_1,\cdots,j_i \in \{1,\cdots,\tilde{L}\}$ of messages $\ell_1,\cdots,\ell_i \in \{1,\cdots,M\}\setminus m$ from the nth database. The same sum is downloaded from K different databases, i.e., the user downloads the same sum from the rows $j_1,\cdots,j_i \in \{1,\cdots,\tilde{L}\}$ of messages $\ell_1,\cdots,\ell_i \in \{1,\cdots,M\}\setminus m$ as $\mathbf{h}_{n+1 \mod N}^T(\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]})$, $\cdots$, $\mathbf{h}_{n+K-1 \mod N}^T(\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]})$. Since the submatrix $[\mathbf{h}_n \ \mathbf{h}_{n+1 \mod N} \ \cdots \ \mathbf{h}_{n+K-1 \mod N}]$ is an invertible matrix by the MDS property, the sum of rows of $\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]}$ is decodable. Note that there are a total of $N\binom{M-1}{i}K^{M-i}(N-K)^{i-1}$ of such symbols in the ith round, therefore $N\binom{M-1}{i}K^{M-i-1}(N-K)^{i-1}$ rows can be decoded as every K sums must be derived from the same set of rows.

These rows are used as side information in the $(i+1)$th round at the remaining $N-K$ databases that do not contribute to the process of creating these side information. The user downloads from databases $n_1(n)=n+K \mod N, \cdots, n_{N-K}(n)=n+N-1 \mod N$ the sums $\mathbf{h}_{n_1(n)}^T(\mathbf{x}_{\theta_1}^{[m]}+\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]})$, $\cdots$, $\mathbf{h}_{n_{N-K}(n)}^T(\mathbf{x}_{\theta_{N-K}}^{[m]}+\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]})$ in the $(i+1)$th round, where $\mathbf{x}_{\theta_l}^{[m]}$ is the row θ_l from the desired message W_m, i.e., the user downloads the sum of rows from the desired message with the side information generated in the ith round. Since the user has decoded the sum $\mathbf{x}_{j_1}^{[\ell_1]}+\mathbf{x}_{j_2}^{[\ell_2]}+\cdots+\mathbf{x}_{j_i}^{[\ell_i]}$, all undesired symbols can be canceled, and the user is left with the desired symbols only.

Now, for the desired symbols, we note that the user downloads from different rows within each repetition. Since the scheme repeats itself K times with the starting database shifted circularly, the user is left with $\mathbf{h}_n^T\mathbf{x}_\theta^{[m]}$, $\mathbf{h}_{n+1 \mod N}^T\mathbf{x}_\theta^{[m]}, \cdots, \mathbf{h}_{n+K-1 \mod N}^T\mathbf{x}_\theta^{[m]}$ for $\theta \in \{1,\cdots,\tilde{L}\}$. This creates K linearly independent equations for each row from W_m by the MDS property. Therefore, all rows can be decoded reliably.

2) Privacy: The scheme downloads all combinations of the sums containing i terms in the ith round from each database. Therefore, the same number of symbols from each message is queried from each database (specifically, KN^{M-1} coded symbols). Note that due to the fact that the user downloads the symbols (desired/undesired) from K databases in a circular shift pattern, each row is queried once within the same database. Thus, the user downloads from KN^{M-1} distinct rows from each database from every message. Since the indices of these rows are chosen randomly and uniformly, and the order of the queries is shuffled randomly and uniformly, the privacy constraint $(Q_n^{[i]}, A_n^{[i]}, W_1, \cdots, W_M) \sim (Q_n^{[1]}, A_n^{[1]}, W_1, \cdots, W_M)$ is satisfied as all the query realizations are equally likely regardless of the message index i.

3) Achievable Rate Calculation: In each repetition, at the ith round, the user downloads the K coded symbols from $N\binom{M-1}{i}K^{M-i-1}(N-K)^{i-1}$ different rows of each message distributed among the N databases. From the described scheme, we note that other than the initial download of NK^{M-1} coded desired bits, at each round the scheme downloads $N\binom{M-1}{i}K^{M-(i+1)}(N-K)^{i}$ desired equations and $N\binom{M-1}{i}K^{M-i}(N-K)^{i-1}$ undesired equations. Hence, the total number of desired equations is $KN\sum_{i=0}^{M-1}\binom{M-1}{i}K^{M-1-i}(N-K)^{i}$, and the total number of undesired equations is $KN\sum_{i=1}^{M-1}\binom{M-1}{i}K^{M-i}(N-K)^{i-1}$

along the K repetitions of the scheme. Therefore, the achievable rate is,

$$\frac{1}{R} = 1 + \frac{\text{total undesired equations}}{\text{total desired equations}} \tag{19}$$

$$= 1 + \frac{\sum_{i=1}^{M-1} \binom{M-1}{i} K^{M-i} (N-K)^{i-1}}{\sum_{i=0}^{M-1} \binom{M-1}{i} K^{M-1-i} (N-K)^{i}} \tag{20}$$

$$= 1 + \frac{\frac{K}{N-K} \sum_{i=1}^{M-1} \binom{M-1}{i} K^{M-1-i} (N-K)^{i}}{N^{M-1}} \tag{21}$$

$$= 1 + \frac{\frac{K}{N-K} \left(\sum_{i=0}^{M-1} \binom{M-1}{i} K^{M-1-i} (N-K)^{i} - K^{M-1} \right)}{N^{M-1}} \tag{22}$$

$$= 1 + \frac{\frac{K}{N-K} \left(N^{M-1} - K^{M-1} \right)}{N^{M-1}} \tag{23}$$

$$= 1 + \frac{K}{N-K} \left(1 - R_c^{M-1} \right) \tag{24}$$

$$= \frac{N - K R_c^{M-1}}{N-K} \tag{25}$$

$$= \frac{1 - R_c^{M}}{1 - R_c} \tag{26}$$

Hence, $R = \frac{1-R_c}{1-R_c^M}$. Note that if $K = 1$, our achievable scheme reduces to the one presented in [7]. We note that our scheme inherits all the properties of the scheme in [7], in particular, its optimality over any subset of messages.

V. Examples

In this section, we give two explicit examples for our scheme. Without loss of generality, we assume that the desired message is W_1.

A. (5, 3) Code With $M = 2$

Initially, sub-indices of all messages are randomly and independently interleaved. For this case, we will have $M = 2$ rounds and then $K = 3$ repetitions; see Table II. We begin round one by downloading $K^{M-1} = 3$ coded bits for the desired message (message W_1) from every database, e.g., we download $\mathbf{h}_1^T \mathbf{x}_1^{[1]}, \mathbf{h}_1^T \mathbf{x}_2^{[1]}, \mathbf{h}_1^T \mathbf{x}_3^{[1]}$ from database 1, and similarly for databases 2-5 by database symmetry. By message symmetry, we download another 3 coded bits from W_2 from each database. Note that for the undesired message, we group every $K = 3$ databases to download from the same row, e.g., we download $\mathbf{h}_1^T \mathbf{x}_1^{[2]}, \mathbf{h}_2^T \mathbf{x}_1^{[2]}, \mathbf{h}_3^T \mathbf{x}_1^{[2]}$ from databases 1-3, $\mathbf{h}_4^T \mathbf{x}_2^{[2]}, \mathbf{h}_5^T \mathbf{x}_2^{[2]}, \mathbf{h}_1^T \mathbf{x}_2^{[2]}$ from databases 4,5,1, and similarly for the remaining databases. By downloading 3 linearly independent equations for every row, we solve for the interference generated by W_2 and create 5 useful side information rows for round two, which are rows $\mathbf{x}_1^{[2]}$ to $\mathbf{x}_5^{[2]}$ from W_2.

In round two, we download sums of the coded bits from W_1, W_2. Since each of the rows $\mathbf{x}_1^{[2]}$ to $\mathbf{x}_5^{[2]}$ is decoded from 3 databases, we can exploit these side information to download further coded bits from W_1 in the remaining $N - K = 2$ databases that do not participate in decoding this row. For example, we use $\mathbf{x}_1^{[2]}$ in databases 4,5 by downloading the sums $\mathbf{h}_4^T (\mathbf{x}_{19}^{[1]} + \mathbf{x}_1^{[2]})$, and $\mathbf{h}_5^T (\mathbf{x}_{20}^{[1]} + \mathbf{x}_1^{[2]})$ and similarly for the rows $\mathbf{x}_2^{[2]}$ to $\mathbf{x}_5^{[2]}$. This creates extra 10 decodable equations in round two in the form of a sum of the two messages. At this point symmetry exists across databases and within messages, and all the interference from the undesired message W_2 is decoded and exploited. However, until this point, we downloaded one equation from every row of W_1. To reliably decode W_1, we need to repeat the previous steps a total of $K = 3$ times by shifting the starting database in a circular pattern, e.g., in repetition 2, we download new equations for the rows $\mathbf{x}_1^{[1]}, \mathbf{x}_2^{[1]}, \mathbf{x}_3^{[1]}$ from database 2 instead of database 1 in repetition 1, and $\mathbf{x}_4^{[1]}, \mathbf{x}_5^{[1]}, \mathbf{x}_6^{[1]}$ from database 3 instead of database 2, etc. As a final step, we shuffle the order of the queries to preclude the databases from identifying the message index from the index of the first downloaded bit.

Since we download symmetric amount of W_1, W_2 from each database and their indices are randomly chosen, privacy constraint is satisfied. Since vectors $\mathbf{x}_i^{[2]}, i \in \{1, \cdots, 5\}$ are downloaded from K databases, their interference is completely decoded. Hence, they can be canceled from round two. Finally, we repeat the scheme 3 times with circular shifts, every desired row is received from K different databases and hence reliably decoded. The explicit query table is shown in Table II. The retrieval rate in this case is $R = \frac{75}{120} = \frac{5}{8} = \frac{1-\frac{3}{5}}{1-(\frac{3}{5})^2}$.

B. (3, 2) Code With $M = 3$

As in the previous example, the messages are randomly and independently interleaved. For this case, the scheme is completed in $M = 3$ rounds and then repeated for $K = 2$ repetitions, see Table III. In the first round, we download $K^{M-1} = 4$ coded bits for W_1 from each database, e.g., $\mathbf{h}_1^T \mathbf{x}_i^{[1]}, i \in \{1, \cdots, 4\}$ from the first database. Similarly, we download one equation from the rows $\mathbf{x}_1^{[1]}$ to $\mathbf{x}_{12}^{[1]}$ by applying the database symmetry. We apply message symmetry to download $N\binom{M-1}{1} K^{M-1} = 24$ undesired coded bits from W_2, W_3. Every 2 coded bits from the undesired bits are grouped together to generate single solved side information vector, e.g., we download as $\mathbf{h}_1^T \mathbf{x}_1^{[2]}, \mathbf{h}_2^T \mathbf{x}_2^{[2]}$ from databases 1,2, $\mathbf{h}_3^T \mathbf{x}_2^{[2]}, \mathbf{h}_1^T \mathbf{x}_1^{[2]}$ from databases 3,1, and similarly for rows $\mathbf{x}_1^{[m]}$ to $\mathbf{x}_6^{[m]}$ where $m = 2, 3$. Hence, we have $N\binom{M-1}{1} K^{M-2} = 12$ side information rows to be used in round two.

In round two, we download sums of every two messages. We exploit the generated side information within the $N - K = 1$ remaining database that does not participate in generating them. For example, we decoded $\mathbf{x}_1^{[2]}$ by downloading equations from databases 1,2, then we use $\mathbf{x}_1^{[2]}$ in database 3 by downloading the sum $\mathbf{h}_3(\mathbf{x}_{15}^{[1]} + \mathbf{x}_1^{[2]})$. Hence, we can download $N\binom{M-1}{1} K^{M-2}(N-K) = 12$ new coded bits of W_1 by using every decoded side information in a sum of W_1 with one of W_2 or W_3. These bits are reliably decoded, since the generated side information can be canceled from the downloaded equation. It remains to add sums of W_2 and W_3 to ensure the privacy. Therefore, we download $N\binom{M-1}{2} K^{M-2}(N-K) = 6$ undesired equations, that will be grouped further to form $N\binom{M-1}{2} K^{M-3}(N-K) = 3$ solved side information equations in the form of sums of W_2 and W_3. As an example, we download $\mathbf{h}_1^T (\mathbf{x}_7^{[2]} + \mathbf{x}_7^{[3]}), \mathbf{h}_2^T (\mathbf{x}_7^{[2]} + \mathbf{x}_7^{[3]})$

TABLE II
PIR FOR CODE (5,3) AND $M = 2$

		DB1	DB2	DB3	DB4	DB5
repetition 1	round 1	$\mathbf{h}_1^T\mathbf{x}_1^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_2^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_3^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_1^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_2^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_4^{[2]}$	$\mathbf{h}_2^T\mathbf{x}_4^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_5^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_6^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_1^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_3^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_4^{[2]}$	$\mathbf{h}_3^T\mathbf{x}_7^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_8^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_9^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_1^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_3^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_5^{[2]}$	$\mathbf{h}_4^T\mathbf{x}_{10}^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_{11}^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_{12}^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_2^{[2]}$ $\mathbf{h}_4^T\mathbf{x}_3^{[2]}$ $\mathbf{h}_4^T\mathbf{x}_5^{[2]}$	$\mathbf{h}_5^T\mathbf{x}_{13}^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_{14}^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_{15}^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_2^{[2]}$ $\mathbf{h}_5^T\mathbf{x}_4^{[2]}$ $\mathbf{h}_5^T\mathbf{x}_5^{[2]}$
	round 2	$\mathbf{h}_1^T(\mathbf{x}_{16}^{[1]}+\mathbf{x}_3^{[2]})$ $\mathbf{h}_1^T(\mathbf{x}_{21}^{[1]}+\mathbf{x}_5^{[2]})$	$\mathbf{h}_2^T(\mathbf{x}_{17}^{[1]}+\mathbf{x}_2^{[2]})$ $\mathbf{h}_2^T(\mathbf{x}_{22}^{[1]}+\mathbf{x}_5^{[2]})$	$\mathbf{h}_3^T(\mathbf{x}_{18}^{[1]}+\mathbf{x}_2^{[2]})$ $\mathbf{h}_3^T(\mathbf{x}_{23}^{[1]}+\mathbf{x}_4^{[2]})$	$\mathbf{h}_4^T(\mathbf{x}_{19}^{[1]}+\mathbf{x}_1^{[2]})$ $\mathbf{h}_4^T(\mathbf{x}_{24}^{[1]}+\mathbf{x}_4^{[2]})$	$\mathbf{h}_5^T(\mathbf{x}_{20}^{[1]}+\mathbf{x}_1^{[2]})$ $\mathbf{h}_5^T(\mathbf{x}_{25}^{[1]}+\mathbf{x}_3^{[2]})$
repetition 2	round 1	$\mathbf{h}_1^T\mathbf{x}_{13}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{14}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{15}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_6^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_7^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_9^{[2]}$	$\mathbf{h}_2^T\mathbf{x}_1^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_2^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_3^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_6^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_8^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_9^{[2]}$	$\mathbf{h}_3^T\mathbf{x}_4^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_5^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_6^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_6^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_8^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_{10}^{[2]}$	$\mathbf{h}_4^T\mathbf{x}_7^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_8^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_9^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_7^{[2]}$ $\mathbf{h}_4^T\mathbf{x}_8^{[2]}$ $\mathbf{h}_4^T\mathbf{x}_{10}^{[2]}$	$\mathbf{h}_5^T\mathbf{x}_{10}^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_{11}^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_{12}^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_7^{[2]}$ $\mathbf{h}_5^T\mathbf{x}_9^{[2]}$ $\mathbf{h}_5^T\mathbf{x}_{10}^{[2]}$
	round 2	$\mathbf{h}_1^T(\mathbf{x}_{20}^{[1]}+\mathbf{x}_8^{[2]})$ $\mathbf{h}_1^T(\mathbf{x}_{25}^{[1]}+\mathbf{x}_{10}^{[2]})$	$\mathbf{h}_2^T(\mathbf{x}_{16}^{[1]}+\mathbf{x}_7^{[2]})$ $\mathbf{h}_2^T(\mathbf{x}_{21}^{[1]}+\mathbf{x}_{10}^{[2]})$	$\mathbf{h}_3^T(\mathbf{x}_{17}^{[1]}+\mathbf{x}_7^{[2]})$ $\mathbf{h}_3^T(\mathbf{x}_{22}^{[1]}+\mathbf{x}_9^{[2]})$	$\mathbf{h}_4^T(\mathbf{x}_{18}^{[1]}+\mathbf{x}_6^{[2]})$ $\mathbf{h}_4^T(\mathbf{x}_{23}^{[1]}+\mathbf{x}_9^{[2]})$	$\mathbf{h}_5^T(\mathbf{x}_{19}^{[1]}+\mathbf{x}_6^{[2]})$ $\mathbf{h}_5^T(\mathbf{x}_{24}^{[1]}+\mathbf{x}_8^{[2]})$
repetition 3	round 1	$\mathbf{h}_1^T\mathbf{x}_{10}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{11}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{12}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{11}^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_{12}^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_{14}^{[2]}$	$\mathbf{h}_2^T\mathbf{x}_{13}^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_{14}^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_{15}^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_{11}^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_{13}^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_{14}^{[2]}$	$\mathbf{h}_3^T\mathbf{x}_1^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_2^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_3^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_{11}^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_{13}^{[2]}$ $\mathbf{h}_{13}^T\mathbf{x}_{15}^{[2]}$	$\mathbf{h}_4^T\mathbf{x}_4^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_5^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_6^{[1]}$ $\mathbf{h}_4^T\mathbf{x}_{12}^{[2]}$ $\mathbf{h}_4^T\mathbf{x}_{13}^{[2]}$ $\mathbf{h}_4^T\mathbf{x}_{15}^{[2]}$	$\mathbf{h}_5^T\mathbf{x}_7^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_8^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_9^{[1]}$ $\mathbf{h}_5^T\mathbf{x}_{12}^{[2]}$ $\mathbf{h}_5^T\mathbf{x}_{14}^{[2]}$ $\mathbf{h}_5^T\mathbf{x}_{15}^{[2]}$
	round 2	$\mathbf{h}_1^T(\mathbf{x}_{19}^{[1]}+\mathbf{x}_{13}^{[2]})$ $\mathbf{h}_1^T(\mathbf{x}_{24}^{[1]}+\mathbf{x}_{15}^{[2]})$	$\mathbf{h}_2^T(\mathbf{x}_{20}^{[1]}+\mathbf{x}_{12}^{[2]})$ $\mathbf{h}_2^T(\mathbf{x}_{25}^{[1]}+\mathbf{x}_{15}^{[2]})$	$\mathbf{h}_3^T(\mathbf{x}_{16}^{[1]}+\mathbf{x}_{12}^{[2]})$ $\mathbf{h}_3^T(\mathbf{x}_{21}^{[1]}+\mathbf{x}_{14}^{[2]})$	$\mathbf{h}_4^T(\mathbf{x}_{17}^{[1]}+\mathbf{x}_{11}^{[2]})$ $\mathbf{h}_4^T(\mathbf{x}_{22}^{[1]}+\mathbf{x}_{14}^{[2]})$	$\mathbf{h}_5^T(\mathbf{x}_{18}^{[1]}+\mathbf{x}_{11}^{[2]})$ $\mathbf{h}_5^T(\mathbf{x}_{23}^{[1]}+\mathbf{x}_{13}^{[2]})$

from databases 1,2. In this case the interference from the rows $\mathbf{x}_7^{[2]}+\mathbf{x}_7^{[3]}$ is decoded. Note that we do not solve for the individual $\mathbf{x}_7^{[2]}$ or $\mathbf{x}_7^{[3]}$ but we *align* them in the same subspace, and solve for their sum.

In round three, we use the newly generated side information, e.g., $\mathbf{x}_7^{[2]}+\mathbf{x}_7^{[3]}$, to download extra $N\binom{M-1}{2}K^{M-3}(N-K)^2 = 3$ desired coded bits in the form of sum of three terms, e.g., $\mathbf{h}_3^T(\mathbf{x}_{27}^{[1]}+\mathbf{x}_7^{[2]}+\mathbf{x}_7^{[3]})$. Finally, the previous steps are repeated $K = 2$ times to reliably decode W_1 and the queries are shuffled for privacy. The retrieval rate in this case is $R = \frac{54}{114} = \frac{9}{19} = \frac{1-\frac{2}{3}}{1-(\frac{2}{3})^3}$. The explicit query structure is shown in Table III.

VI. CONVERSE PROOF

In this section, we prove the converse for PIR from MDS-coded databases. The proof extends the techniques in [7] to the case of MDS-coded databases. The proof presented here does not use symmetrization arguments or fixing of an individual query as in the conference version [15], which presents an alternative proof that provides an alternative perspective.

We need the following lemma which states that in the PIR problem from (N, K) MDS-coded databases, the answers from any K databases are statistically independent.

Lemma 1 (Independence of Answers of Any K Databases): In the PIR problem from (N, K) MDS-coded databases, for any set $\mathcal{K}$ of databases such that $|\mathcal{K}| = K$, i.e., for $m \in \{1, \cdots, M\}$

$$H(A_{\mathcal{K}}^{[m]}|Q_{\mathcal{K}}^{[m]}) = \sum_{n\in\mathcal{K}} H(A_n^{[m]}|Q_n^{[m]}) \tag{27}$$

Furthermore, (27) is true if conditioned on any subset of messages $W_{\mathcal{S}}$, i.e., for $m \in \{1, \cdots, M\}$

$$H(A_{\mathcal{K}}^{[m]}|Q_{\mathcal{K}}^{[m]}, W_{\mathcal{S}}) = \sum_{n\in\mathcal{K}} H(A_n^{[m]}|Q_n^{[m]}, W_{\mathcal{S}}) \tag{28}$$

Proof: Consider a set of databases $\mathcal{K}$ such that $|\mathcal{K}| = K$. We prove first the statistical independence between the vectors $\{\mathbf{y}_n, n \in \mathcal{K}\}$ where $\mathbf{y}_n$ represents the contents of the nth database. The contents of set $\mathcal{K}$ of databases can be written as

$$[\mathbf{y}_n,\ n \in \mathcal{K}] = \begin{bmatrix} W_1 \\ \vdots \\ W_M \end{bmatrix} [\mathbf{h}_n,\ n \in \mathcal{K}] = \begin{bmatrix} W_1 \\ \vdots \\ W_M \end{bmatrix} \mathbf{H}_{\mathcal{K}} \tag{29}$$

where $\mathbf{H}_{\mathcal{K}} = [\mathbf{h}_n,\ n \in \mathcal{K}]$ is a $\mathbb{F}_q^{K\times K}$ matrix. By construction of the distributed storage code, the matrix $\mathbf{H}_{\mathcal{K}}$ is an

TABLE III
PIR FOR CODE (3,2) AND $M = 3$

		DB1	DB2	DB3
repetition 1	round 1	$\mathbf{h}_1^T\mathbf{x}_1^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_2^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_3^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_4^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_1^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_2^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_4^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_5^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_1^{[3]}$ $\mathbf{h}_1^T\mathbf{x}_2^{[3]}$ $\mathbf{h}_1^T\mathbf{x}_4^{[3]}$ $\mathbf{h}_1^T\mathbf{x}_5^{[3]}$	$\mathbf{h}_2^T\mathbf{x}_5^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_6^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_7^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_8^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_1^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_3^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_4^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_6^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_1^{[3]}$ $\mathbf{h}_2^T\mathbf{x}_3^{[3]}$ $\mathbf{h}_2^T\mathbf{x}_4^{[3]}$ $\mathbf{h}_2^T\mathbf{x}_6^{[3]}$	$\mathbf{h}_3^T\mathbf{x}_9^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_{10}^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_{11}^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_{12}^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_2^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_3^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_5^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_6^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_2^{[3]}$ $\mathbf{h}_3^T\mathbf{x}_3^{[3]}$ $\mathbf{h}_3^T\mathbf{x}_5^{[3]}$ $\mathbf{h}_3^T\mathbf{x}_6^{[3]}$
	round 2	$\mathbf{h}_1^T(\mathbf{x}_{13}^{[1]}+\mathbf{x}_3^{[2]})$ $\mathbf{h}_1^T(\mathbf{x}_{16}^{[1]}+\mathbf{x}_3^{[3]})$ $\mathbf{h}_1^T(\mathbf{x}_7^{[2]}+\mathbf{x}_7^{[3]})$ $\mathbf{h}_1^T(\mathbf{x}_{19}^{[1]}+\mathbf{x}_6^{[2]})$ $\mathbf{h}_1^T(\mathbf{x}_{22}^{[1]}+\mathbf{x}_6^{[3]})$ $\mathbf{h}_1^T(\mathbf{x}_8^{[2]}+\mathbf{x}_8^{[3]})$	$\mathbf{h}_2^T(\mathbf{x}_{14}^{[1]}+\mathbf{x}_2^{[2]})$ $\mathbf{h}_2^T(\mathbf{x}_{17}^{[1]}+\mathbf{x}_2^{[3]})$ $\mathbf{h}_2^T(\mathbf{x}_7^{[2]}+\mathbf{x}_7^{[3]})$ $\mathbf{h}_2^T(\mathbf{x}_{20}^{[1]}+\mathbf{x}_5^{[2]})$ $\mathbf{h}_2^T(\mathbf{x}_{23}^{[1]}+\mathbf{x}_5^{[3]})$ $\mathbf{h}_2^T(\mathbf{x}_9^{[2]}+\mathbf{x}_9^{[3]})$	$\mathbf{h}_3^T(\mathbf{x}_{15}^{[1]}+\mathbf{x}_1^{[2]})$ $\mathbf{h}_3^T(\mathbf{x}_{18}^{[1]}+\mathbf{x}_1^{[3]})$ $\mathbf{h}_3^T(\mathbf{x}_8^{[2]}+\mathbf{x}_8^{[3]})$ $\mathbf{h}_3^T(\mathbf{x}_{21}^{[1]}+\mathbf{x}_4^{[2]})$ $\mathbf{h}_3^T(\mathbf{x}_{24}^{[1]}+\mathbf{x}_4^{[3]})$ $\mathbf{h}_3^T(\mathbf{x}_9^{[2]}+\mathbf{x}_9^{[3]})$
	rd. 3	$\mathbf{h}_1^T(\mathbf{x}_{25}^{[1]}+\mathbf{x}_9^{[2]}+\mathbf{x}_9^{[3]})$	$\mathbf{h}_2^T(\mathbf{x}_{26}^{[1]}+\mathbf{x}_8^{[2]}+\mathbf{x}_8^{[3]})$	$\mathbf{h}_3^T(\mathbf{x}_{27}^{[1]}+\mathbf{x}_7^{[2]}+\mathbf{x}_7^{[3]})$
repetition 2	round 1	$\mathbf{h}_1^T\mathbf{x}_9^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{10}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{11}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{12}^{[1]}$ $\mathbf{h}_1^T\mathbf{x}_{10}^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_{11}^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_{13}^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_{14}^{[2]}$ $\mathbf{h}_1^T\mathbf{x}_{10}^{[3]}$ $\mathbf{h}_1^T\mathbf{x}_{11}^{[3]}$ $\mathbf{h}_1^T\mathbf{x}_{13}^{[3]}$ $\mathbf{h}_1^T\mathbf{x}_{14}^{[3]}$	$\mathbf{h}_2^T\mathbf{x}_1^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_2^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_3^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_4^{[1]}$ $\mathbf{h}_2^T\mathbf{x}_{10}^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_{12}^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_{13}^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_{15}^{[2]}$ $\mathbf{h}_2^T\mathbf{x}_{10}^{[3]}$ $\mathbf{h}_2^T\mathbf{x}_{12}^{[3]}$ $\mathbf{h}_2^T\mathbf{x}_{13}^{[3]}$ $\mathbf{h}_2^T\mathbf{x}_{15}^{[3]}$	$\mathbf{h}_3^T\mathbf{x}_5^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_6^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_7^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_8^{[1]}$ $\mathbf{h}_3^T\mathbf{x}_{11}^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_{12}^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_{14}^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_{15}^{[2]}$ $\mathbf{h}_3^T\mathbf{x}_{11}^{[3]}$ $\mathbf{h}_3^T\mathbf{x}_{12}^{[3]}$ $\mathbf{h}_3^T\mathbf{x}_{14}^{[3]}$ $\mathbf{h}_3^T\mathbf{x}_{15}^{[3]}$
	round 2	$\mathbf{h}_1^T(\mathbf{x}_{15}^{[1]}+\mathbf{x}_{12}^{[2]})$ $\mathbf{h}_1^T(\mathbf{x}_{18}^{[1]}+\mathbf{x}_{12}^{[3]})$ $\mathbf{h}_1^T(\mathbf{x}_{16}^{[2]}+\mathbf{x}_{16}^{[3]})$ $\mathbf{h}_1^T(\mathbf{x}_{21}^{[1]}+\mathbf{x}_{15}^{[2]})$ $\mathbf{h}_1^T(\mathbf{x}_{24}^{[1]}+\mathbf{x}_{15}^{[3]})$ $\mathbf{h}_1^T(\mathbf{x}_{17}^{[2]}+\mathbf{x}_{17}^{[3]})$	$\mathbf{h}_2^T(\mathbf{x}_{13}^{[1]}+\mathbf{x}_{11}^{[2]})$ $\mathbf{h}_2^T(\mathbf{x}_{16}^{[1]}+\mathbf{x}_{11}^{[3]})$ $\mathbf{h}_2^T(\mathbf{x}_{16}^{[2]}+\mathbf{x}_{16}^{[3]})$ $\mathbf{h}_2^T(\mathbf{x}_{19}^{[1]}+\mathbf{x}_{14}^{[2]})$ $\mathbf{h}_2^T(\mathbf{x}_{22}^{[1]}+\mathbf{x}_{14}^{[3]})$ $\mathbf{h}_2^T(\mathbf{x}_{18}^{[2]}+\mathbf{x}_{18}^{[3]})$	$\mathbf{h}_3^T(\mathbf{x}_{14}^{[1]}+\mathbf{x}_{10}^{[2]})$ $\mathbf{h}_3^T(\mathbf{x}_{17}^{[1]}+\mathbf{x}_{10}^{[3]})$ $\mathbf{h}_3^T(\mathbf{x}_{17}^{[2]}+\mathbf{x}_{17}^{[3]})$ $\mathbf{h}_3^T(\mathbf{x}_{20}^{[1]}+\mathbf{x}_{13}^{[2]})$ $\mathbf{h}_3^T(\mathbf{x}_{23}^{[1]}+\mathbf{x}_{13}^{[3]})$ $\mathbf{h}_3^T(\mathbf{x}_{18}^{[2]}+\mathbf{x}_{18}^{[3]})$
	rd. 3	$\mathbf{h}_1^T(\mathbf{x}_{27}^{[1]}+\mathbf{x}_{18}^{[2]}+\mathbf{x}_{18}^{[3]})$	$\mathbf{h}_2^T(\mathbf{x}_{25}^{[1]}+\mathbf{x}_{17}^{[2]}+\mathbf{x}_{17}^{[3]})$	$\mathbf{h}_3^T(\mathbf{x}_{26}^{[1]}+\mathbf{x}_{16}^{[2]}+\mathbf{x}_{16}^{[3]})$

invertible matrix. Using [10, Lemma 1] and the fact that elements of the messages are chosen independently and uniformly over $\mathbb{F}_q^{\tilde{L}\times K}$, we conclude that

$$[\mathbf{y}_n,\ n\in\mathcal{K}] = \begin{bmatrix} W_1 \\ \vdots \\ W_M \end{bmatrix} \mathbf{H}_{\mathcal{K}} \sim \begin{bmatrix} W_1 \\ \vdots \\ W_M \end{bmatrix} \tag{30}$$

where $A \sim B$ denotes that random variables A and B are identically distributed. Therefore, the contents of the databases are statistically equivalent to the messages. Hence, the columns of $[\mathbf{y}_n,\ n\in\mathcal{K}]$ are statistically independent since the elements of the messages are independent.

Since $A_n^{[m]}, n\in\mathcal{K}$ are deterministic functions of $(\mathbf{y}_n, Q_n^{[m]})$, $\{A_n^{[m]} : n\in\mathcal{K}\}$ are statistically independent as they are

deterministic functions of independent random variables. Therefore, if $\mathcal{K} = \{n_1, n_2, \cdots, n_K\}$

$$H(A_{\mathcal{K}}^{[m]}|Q_{\mathcal{K}}^{[m]}) = \sum_{i=1}^{K} H(A_{n_i}^{[m]}|A_{1:n_{i-1}}^{[m]}, Q_{\mathcal{K}}^{[m]}) \tag{31}$$

$$= \sum_{i=1}^{K} H(A_{n_i}^{[m]}|Q_{\mathcal{K}}^{[m]}) \tag{32}$$

$$= \sum_{n\in\mathcal{K}} H(A_{n}^{[m]}|Q_{n}^{[m]}) \tag{33}$$

where (32) follows from the independence of any K answer strings, (33) follows from the fact that $Q_{\mathcal{K}}^{[m]} \to Q_{n}^{[m]} \to A_{n}^{[m]}$ is a Markov chain. We note that since coding is applied on individual messages, conditioning on any subset of messages $W_{\mathcal{S}}$ with $|W_{\mathcal{S}}| = S$ is equivalent to reducing the problem to storing $M - S$ independent messages instead of M messages. Hence, the statistical independence argument in (28) follows as before. ∎

We use Han's inequality [17, Th. 17.6.1] in a similar way to [10].

Lemma 2 (Han's Inequality): Let $\mathcal{K} \subseteq \{1, \cdots, N\}$, such that $|\mathcal{K}| = K$. Then, for any subset of messages $W_{\mathcal{S}}$,

$$\frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} H(A_{\mathcal{K}}^{[m]}|W_{\mathcal{S}}, Q_{1:N}^{[m]}) \geq \frac{K}{N} H(A_{1:N}^{[m]}|W_{\mathcal{S}}, Q_{1:N}^{[m]}) \tag{34}$$

The following lemma characterizes a lower bound on the interference components in $A_{1:N}^{[1]}$ that result from the interfering messages $W_{2:M}$ which is represented by $\frac{L}{R} - L$. The following lemma is exactly [7, Lemma 5]. The result does not change due to the distributed storage code introduced in our problem. We include the proof of this lemma here for completeness.

Lemma 3 (Interference Lower Bound): The interference from undesired messages within the answer strings, $\frac{L}{R} - L$, is lower bounded by,

$$L\left(\frac{1}{R} - 1 + \frac{o(L)}{L}\right) \geq I(W_{2:M}; Q_{1:N}^{[1]}, A_{1:N}^{[1]}|W_1) \tag{35}$$

Proof: We start with the right hand side of (35),

$$I(W_{2:M}; Q_{1:N}^{[1]}, A_{1:N}^{[1]}|W_1)$$

$$= I(W_{2:M}; Q_{1:N}^{[1]}, A_{1:N}^{[1]}, W_1) \tag{36}$$

$$= I(W_{2:M}; Q_{1:N}^{[1]}, A_{1:N}^{[1]}) + I(W_{2:M}; W_1|Q_{1:N}^{[1]}, A_{1:N}^{[1]}) \tag{37}$$

$$= I(W_{2:M}; Q_{1:N}^{[1]}) + I(W_{2:M}; A_{1:N}^{[1]}|Q_{1:N}^{[1]}) + o(L) \tag{38}$$

$$= I(W_{2:M}; A_{1:N}^{[1]}|Q_{1:N}^{[1]}) + o(L) \tag{39}$$

$$= H(A_{1:N}^{[1]}|Q_{1:N}^{[1]}) - H(A_{1:N}^{[1]}|Q_{1:N}^{[1]}, W_{2:M}) + o(L) \tag{40}$$

$$\leq \sum_{n=1}^{N} H(A_n^{[1]}) - H(W_1, A_{1:N}^{[1]}|Q_{1:N}^{[1]}, W_{2:M}) + H(W_1|Q_{1:N}^{[1]}, A_{1:N}^{[1]}, W_{2:M}) + o(L) \tag{41}$$

$$= \frac{L}{R} - H(W_1|Q_{1:N}^{[1]}, W_{2:M}) - H(A_{1:N}^{[1]}|Q_{1:N}^{[1]}, W_{1:M}) + o(L) \tag{42}$$

$$= \frac{L}{R} - L + o(L) \tag{43}$$

$$= L\left(\frac{1}{R} - 1 + \frac{o(L)}{L}\right) \tag{44}$$

where (36) follows from the independence of messages, (38) and (42) follow from the decodability of W_1 from $(Q_{1:N}^{[1]}, A_{1:N}^{[1]})$, (39) follows from the independence of the queries $Q_{1:N}^{[1]}$ and the messages $W_{2:M}$, (41) follows from the fact that conditioning reduces entropy, and (43) follows from the fact that the answers $A_{1:N}^{[1]}$ are deterministic functions of $(Q_{1:N}^{[1]}, W_{1:M})$ and the independence of $(W_1, Q_{1:N}^{[1]}, W_{2:M})$. ∎

In the following lemma, we prove an inductive relation for the mutual information term on the right hand side of (35).

Lemma 4 (Induction Lemma): We have the following inductive relationship,

$$I(W_{m:M}; Q_{1:N}^{[m-1]}, A_{1:N}^{[m-1]}|W_{1:m-1}) \geq \frac{K}{N} I(W_{m+1:M}; Q_{1:N}^{[m]}, A_{1:N}^{[m]}|W_{1:m}) + \frac{KL\left(1 - \frac{o(L)}{L}\right)}{N} \tag{45}$$

Proof: We start with the left hand side of (45),

$$I(W_{m:M}; Q_{1:N}^{[m-1]}, A_{1:N}^{[m-1]}|W_{1:m-1})$$

$$\geq \frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} I(W_{m:M}; Q_{\mathcal{K}}^{[m-1]}, A_{\mathcal{K}}^{[m-1]}|W_{1:m-1}) \tag{46}$$

$$= \frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} I(W_{m:M}; A_{\mathcal{K}}^{[m-1]}|W_{1:m-1}, Q_{\mathcal{K}}^{[m-1]}) \tag{47}$$

$$= \frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} H(A_{\mathcal{K}}^{[m-1]}|W_{1:m-1}, Q_{\mathcal{K}}^{[m-1]}) \tag{48}$$

$$= \frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} \sum_{n\in\mathcal{K}} H(A_{n}^{[m-1]}|W_{1:m-1}, Q_{n}^{[m-1]}) \tag{49}$$

$$= \frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} \sum_{n\in\mathcal{K}} H(A_{n}^{[m]}|W_{1:m-1}, Q_{n}^{[m]}) \tag{50}$$

$$= \frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} H(A_{\mathcal{K}}^{[m]}|W_{1:m-1}, Q_{\mathcal{K}}^{[m]}) \tag{51}$$

$$\geq \frac{1}{\binom{N}{K}} \sum_{\mathcal{K}:|\mathcal{K}|=K} H(A_{\mathcal{K}}^{[m]}|W_{1:m-1}, Q_{1:N}^{[m]}) \tag{52}$$

$$\geq \frac{K}{N} H(A_{1:N}^{[m]}|W_{1:m-1}, Q_{1:N}^{[m]}) \tag{53}$$

$$= \frac{K}{N} I(W_{m:M}; Q_{1:N}^{[m]}, A_{1:N}^{[m]}|W_{1:m-1}) \tag{54}$$

$$= \frac{K}{N}\left[I(W_{m:M}; W_m, Q_{1:N}^{[m]}, A_{1:N}^{[m]}|W_{1:m-1}) - o(L)\right] \tag{55}$$

$$= \frac{K}{N}\left[I(W_{m:M}; W_m|W_{1:m-1}) + I(W_{m:M}; Q_{1:N}^{[m]}, A_{1:N}^{[m]}|W_{1:m}) - o(L)\right] \tag{56}$$

$$= \frac{K}{N}\left[L + I(W_{m+1:M}; Q_{1:N}^{[m]}, A_{1:N}^{[m]}|W_{1:m}) - o(L)\right] \tag{57}$$

$$= \frac{K}{N} I(W_{m+1:M}; Q_{1:N}^{[m]}, A_{1:N}^{[m]}|W_{1:m}) + \frac{KL\left(1 - \frac{o(L)}{L}\right)}{N} \tag{58}$$

where (46) follows from the fact that for every subset $\mathcal{K}$ such that $|\mathcal{K}| = K$ we have $I(W_{m:M}; Q_{1:N}^{[m-1]}, A_{1:N}^{[m-1]}|W_{1:m-1}) \geq I(W_{m:M}; Q_{\mathcal{K}}^{[m-1]}, A_{\mathcal{K}}^{[m-1]}|W_{1:m-1})$ by the non-negativity of the mutual information, (47) follows from the independence of the messages and the queries, (48) follows from the fact that the answers $A_{\mathcal{K}}^{[m-1]}$ are deterministic functions of $(W_{1:M}, Q_{\mathcal{K}}^{[m-1]})$, (49) and (51) follow from the independence of any K answers as a consequence of Lemma 1, (50) follows from the privacy constraint, (52) follows from conditioning reduces entropy, (53) follows from Han's inequality in Lemma 2, (54) follows from the fact that $A_{1:N}^{[m]}$ is a deterministic function of $(W_{1:M}, Q_{1:N}^{[m]})$ and the independence of the messages and the queries, (55) follows from the decodability of W_m from $(Q_{1:N}^{[m]}, A_{1:N}^{[m]})$, and (57) follows from $I(W_{m:M}; W_m|W_{1:m-1}) = H(W_m) = L$ from the independence of the messages. ■

Now, we are ready to complete the converse proof by applying Lemma 3 and Lemma 4 successively. We have

$$
\begin{aligned}
&L\left(\frac{1}{R} - 1 + \frac{o(L)}{L}\right) \\
&\quad \geq I(W_{2:M}; Q_{1:N}^{[1]}, A_{1:N}^{[1]}|W_1) && (59)\\
&\quad \geq \frac{K}{N} I(W_{3:M}; Q_{1:N}^{[2]}, A_{1:N}^{[2]}|W_{1:2}) + \frac{KL\left(1 - \frac{o(L)}{L}\right)}{N} && (60)\\
&\quad \geq \cdots && (61)\\
&\quad \geq \frac{K^{M-2}}{N^{M-2}} I(W_{M:M}; Q_{1:N}^{[M-1]}, A_{1:N}^{[M-1]}|W_{1:M-1}) \\
&\qquad + \left(\frac{K}{N} + \frac{K^2}{N^2} + \cdots + \frac{K^{M-2}}{N^{M-2}}\right)\left(1 - \frac{o(L)}{L}\right) L && (62)\\
&\quad \geq \left(\frac{K}{N} + \frac{K^2}{N^2} + \cdots + \frac{K^{M-1}}{N^{M-1}}\right)\left(1 - \frac{o(L)}{L}\right) L && (63)
\end{aligned}
$$

where (59) follows from Lemma 3, and (60)-(63) follow from applying Lemma 4 successively for $M - 1$ times. Hence, we have

$$\frac{1}{R} \geq \left(1 + \frac{K}{N} + \frac{K^2}{N^2} + \cdots + \frac{K^{M-1}}{N^{M-1}}\right)\left(1 - \frac{o(L)}{L}\right) \quad (64)$$

By taking $L \to \infty$, and noting $\frac{o(L)}{L} \to 0$, we have

$$
\begin{aligned}
R &\leq \frac{1}{\sum_{i=0}^{M-1} \left(\frac{K}{N}\right)^i} && (65)\\
&= \frac{1}{\sum_{i=0}^{M-1} R_c^i} = \frac{1 - R_c}{1 - R_c^M} && (66)
\end{aligned}
$$

Remark 1: In the conference version of this work [15], we presented a different converse proof. In this remark, we briefly describe this alternative proof for a more complete and insightful exposition. The converse proof in [15] assumes without loss of generality that the answer strings are symmetric across messages and databases, and an individual answer string (e.g., A_1) can be the same no matter what the desired message is. The converse proof is obtained by induction over M. We start the proof by considering the case of $M = 2$ messages as a base induction step. In this case, we derive a lower bound on the interference from W_2 to be [15, Lemma 3],

$$H(A_{1:N}^{[1]}|W_1, \mathcal{Q}) \geq \frac{KL}{N} \quad (67)$$

where $\mathcal{Q} \triangleq \{Q_n^{[m]} : m \in \{1, \cdots, M\},\ n \in \{1, \cdots, N\}\}$. From [15, Lemma 3], we prove that $R \leq \frac{1}{1+\frac{K}{N}}$ for $M = 2$, which proves the base induction step. For any M, we prove that the remaining uncertainty on the answer strings after conditioning on one of the interfering messages is upper bounded by [15, Lemma 4],

$$H(A_{1:N}^{[2]}|W_1, \mathcal{Q}) \leq \frac{N}{K}\left(NH(A_1|\mathcal{Q}) - L\right) \quad (68)$$

Consequently, we obtain an inductive relation for any M as,

$$NH(A_1|\mathcal{Q}) \geq \left(1 + \frac{K}{N}\right) L + \frac{K^2}{N} H(A_1|W_1, W_2, \mathcal{Q}) \quad (69)$$

Using the induction hypothesis,

$$NH(A_1|\mathcal{Q}) \geq L \sum_{i=0}^{M-1} \left(\frac{K}{N}\right)^i \quad (70)$$

and plugging it to the inductive relation concludes the converse proof.

VII. Conclusions

In this paper, we considered the private information retrieval (PIR) problem over MDS-coded and non-colluding databases. We employed information-theoretic arguments to derive the optimal retrieval rate for the desired message for any given (N, K) storage code. We showed that the PIR capacity in this case is given by $C = \frac{1-R_c}{1-R_c^M}$. The optimal retrieval rate is strictly higher than the best-known achievable scheme in the literature for any finite number of messages. This result reduces to the capacity of the classical PIR problem, i.e., with repetition-coded databases, by observing that for repetition coding $R_c = \frac{1}{N}$. Our result shows that the optimal retrieval cost is independent of the explicit structure of the storage code, and the number of databases, but depends only on the code rate R_c and the number of messages M. Interestingly, the result implies that there is no gain of joint design of the MDS storage code and the retrieval procedure. The result also establishes a fundamental tradeoff between the code rate and the PIR capacity for the MDS codes.

References

[1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–981, 1998.

[2] W. Gasarch, "A survey on private information retrieval," *Bull. EATCS*, vol. 82, pp. 72–107, 2004.

[3] R. Ostrovsky and W. E. Skeith, III, "A survey of single-database private information retrieval: Techniques and applications," in *Proc. Int. Workshop Public Key Cryptograph.*, 2007, pp. 393–411.

[4] S. Yekhanin, "Private information retrieval," *Commun. ACM*, vol. 53, no. 4, pp. 68–73, 2010.

[5] R. Tajeddine and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," in *Proc. IEEE ISIT*, Jul. 2016, pp. 1411–1415.

[6] T. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *Proc. IEEE ISIT*, Jun. 2015, pp. 2842–2846.

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

[8] S. A. Jafar, "Blind interference alignment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 3, pp. 216–227, Jun. 2012.

[9] H. Sun and S. A. Jafar. (2016). "Blind interference alignment for private information retrieval." [Online]. Available: https://arxiv.org/abs/1601.07885

[10] H. Sun and S. A. Jafar. (2016). "The capacity of robust private information retrieval with colluding databases." [Online]. Available: https://arxiv.org/abs/1605.00635

[11] H. Sun and S. A. Jafar. (2016). "The capacity of symmetric private information retrieval." [Online]. Available: https://arxiv.org/abs/1606.08828
[12] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network codes for distributed storage," *Proc. IEEE*, vol. 99, no. 3, pp. 476–489, Mar. 2011.
[13] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *Proc. IEEE ISIT*, Jun./Jul. 2014, pp. 856–860.
[14] A. Fazeli, A. Vardy, and E. Yaakobi. (2015). "PIR with low storage overhead: Coding instead of replication." [Online]. Available: https://arxiv.org/abs/1505.06241
[15] K. Banawan and S. Ulukus, "Private information retrieval from coded databases," in *Proc. IEEE ICC*, May 2017, pp. 1–6.
[16] H. Sun and S. A. Jafar. (2016). "Multiround private information retrieval: Capacity and storage overhead." [Online]. Available: https://arxiv.org/abs/1611.02257
[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

Karim Banawan (S'13) received the B.Sc. and M.Sc. degrees (Highest Hons.) in electrical engineering from Alexandria University, Alexandria, Egypt, in 2008, 2012, respectively. He is currently pursuing the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. His research interests include information theory, wireless communications, physical layer security and private information retrieval.

Sennur Ulukus (S'90–M'98–SM'15–F'16) is a Professor of Electrical and Computer Engineering at the University of Maryland at College Park, where she also holds a joint appointment with the Institute for Systems Research (ISR). Prior to joining UMD, she was a Senior Technical Staff Member at AT&T Labs-Research. She received her Ph.D. degree in Electrical and Computer Engineering from Wireless Information Network Laboratory (WINLAB), Rutgers University, and B.S. and M.S. degrees in Electrical and Electronics Engineering from Bilkent University. Her research interests are in wireless communications, information theory, signal processing, and networks, with recent focus on information theoretic physical layer security, private information retrieval, energy harvesting communications, and wireless energy and information transfer.

Dr. Ulukus is a fellow of the IEEE, and a Distinguished Scholar-Teacher of the University of Maryland. She received the 2003 IEEE Marconi Prize Paper Award in Wireless Communications, an 2005 NSF CAREER Award, the 2010-2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Education Award. She is on the Editorial Board of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING (2016-). She was an Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS-Series on Green Communications and Networking (2015-2016), IEEE TRANSACTIONS ON INFORMATION THEORY (2007-2010), and IEEE TRANSACTIONS ON COMMUNICATIONS (2003-2007). She was a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (2015 and 2008), *Journal of Communications and Networks* (2012), and IEEE TRANSACTIONS ON INFORMATION THEORY (2011). She was a general TPC co-chair of 2017 IEEE ISIT, 2016 IEEE Globecom, 2014 IEEE PIMRC, and 2011 IEEE CTW.