Private Information Retrieval from Coded Databases

Karim Banawan Sennur Ulukus
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
kbanawan@umd.edu *ulukus@umd.edu*

***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 $C = \left(1 + \frac{K}{N} + \frac{K^2}{N^2} + \cdots + \frac{K^{M-1}}{N^{M-1}}\right)^{-1} = (1 + R_c + R_c^2 + \cdots + R_c^{M-1})^{-1} = \frac{1-R_c}{1-R_c^M}$, where R_c is the rate of the (N,K) 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. The result generalizes the achievability and converse results for the classical PIR with replicating databases to the case of coded databases.**

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 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 repetition coded. A trivial solution for this task is to download all of the contents of the databases. However, this solution is highly impractical, in particular 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, 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]. [10] extends this setting to the case of T colluding databases with and without node failures. Another interesting extension of the problem is the symmetric PIR [11], in which the privacy of the undesired messages need to be preserved against the user.

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 the 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 PIR problem, little research exists for the coded PIR with a few exceptions: [13] which 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 in the expense of having storage nodes that grow with the message size. [6] considers a general formulation for the coded PIR problem, and obtains a tradeoff between storage and retrieval costs based on certain sufficient conditions. [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 proposes new linear codes called the k-server PIR codes.

This work was supported by NSF Grants CNS 13-14733, CCF 14-22111, CCF 14-22129, and CNS 15-26608.

In this paper, we consider the PIR problem for non-colluding and coded databases. We use the information-theoretic formulation. We do not assume any specific structure on the generator matrix of the distributed storage code other than linear independence of every K columns. This condition is equivalent to restricting the storage code structure to MDS codes. Note 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 or the number of nodes. Surprisingly, the result implies the optimality of separation between the design of the PIR scheme and the 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 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 generalize the inductive relation in [7] to account for coding. We present two new lemmas that capture the essence of the converse proof, namely: interference lower bound for $M = 2$, and interference conditioning for general M.

II. System Model

Consider an (N, K) MDS 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 $\mathbf{h}_i \in \mathbb{F}_q^K$, $i \in \{1, \cdots, N\}$. For an MDS code, any set $\mathcal{K}$ of columns of $\mathbf{H}$ such that $|\mathcal{K}| \leq K$ are linearly independent. The storage code $f_n : \mathbf{w}_j^{[i]} \rightarrow y_{n,j}^{[i]}$ 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

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

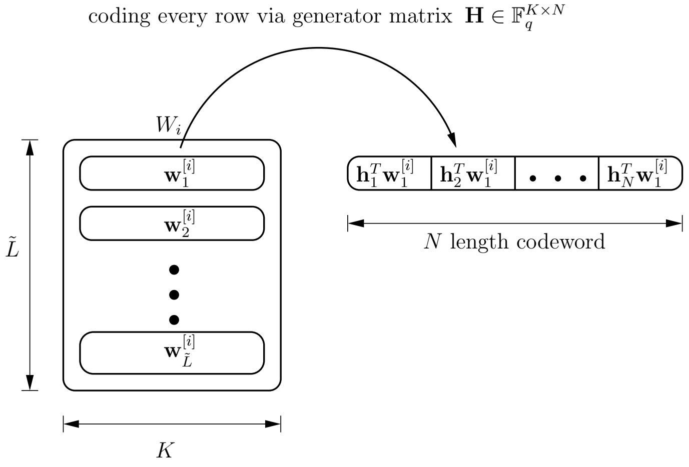
coding every row via generator matrix $\mathbf{H} \in \mathbb{F}_q^{K \times N}$
W_i
$\mathbf{w}_1^{[i]}$
$\mathbf{w}_2^{[i]}$
$\mathbf{w}_{\tilde{L}}^{[i]}$
$\tilde{L}$
K
$\mathbf{h}_1^T \mathbf{w}_1^{[i]}$
$\mathbf{h}_2^T \mathbf{w}_1^{[i]}$
$\mathbf{h}_N^T \mathbf{w}_1^{[i]}$
N length codeword

Fig. 1. Coding process for message W_i.

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 $|\mathcal{K}| \geq K$,

$$H(\mathbf{y}_{\bar{\mathcal{K}}} | \mathbf{y}_{\mathcal{K}}) = 0 \tag{6}$$

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 is $R_c = \frac{K}{N}$. 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,

$$I(Q_1^{[i]}, \cdots, Q_N^{[i]}; W_1, \cdots, W_M) = 0 \tag{7}$$

In order to ensure privacy, the queries should be independent of the desired message index i, i.e., the privacy constraint is,

$$I(Q_n^{[i]}; i) = 0, \quad n \in \{1, \cdots, N\} \tag{8}$$

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

$$H(A_n^{[i]} | Q_n^{[i]}, \mathbf{y}_n) = H(A_n^{[i]} | Q_n^{[i]}, W_1, \cdots, W_M) = 0 \tag{9}$$

In addition, the user should be able to decode W_i reliably from all the answer strings collected from the N databases. Consequently, we have the following reliability constraint,

$$H(W_i | A_1^{[i]}, \cdots, A_N^{[i]}, Q_1^{[i]}, \cdots, Q_N^{[i]}) = 0 \tag{10}$$

The retrieval rate R for the PIR problem is the ratio of the size of the desired message to the total download cost,

$$R = \frac{H(W_i)}{\sum_{n=1}^N H(A_n^{[i]})} \tag{11}$$

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	$\mathbf{h}_1^T\mathbf{w}_1^{[M]}$ $\mathbf{h}_1^T\mathbf{w}_2^{[M]}$ $\vdots$ $\mathbf{h}_1^T\mathbf{w}_{\tilde{L}}^{[M]}$	$\mathbf{h}_2^T\mathbf{w}_1^{[M]}$ $\mathbf{h}_2^T\mathbf{w}_2^{[M]}$ $\vdots$ $\mathbf{h}_2^T\mathbf{w}_{\tilde{L}}^{[M]}$	$\cdots$ $\cdots$ $\cdots$ $\cdots$	$\mathbf{h}_N^T\mathbf{w}_1^{[M]}$ $\mathbf{h}_N^T\mathbf{w}_2^{[M]}$ $\vdots$ $\mathbf{h}_N^T\mathbf{w}_{\tilde{L}}^{[M]}$

The PIR capacity C is the supremum of R over all retrieval schemes.

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 problem 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*

$$C = \frac{1 - R_c}{1 - R_c^M} \tag{12}$$

$$= \frac{1}{1 + R_c + \cdots + R_c^{M-1}} \tag{13}$$

$$= \left(1 + \frac{K}{N} + \frac{K^2}{N^2} + \cdots + \frac{K^{M-1}}{N^{M-1}}\right)^{-1} \tag{14}$$

We have the following remarks about the main result. We first note that the PIR capacity in (12) 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 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. This observation conforms with the result of [6], in which a fundamental trade off exists between storage cost and the retrieval download cost. The capacity expression in Theorem 1 is plotted in Fig. 2 as a function of the code rate R_c for various numbers of messages M.

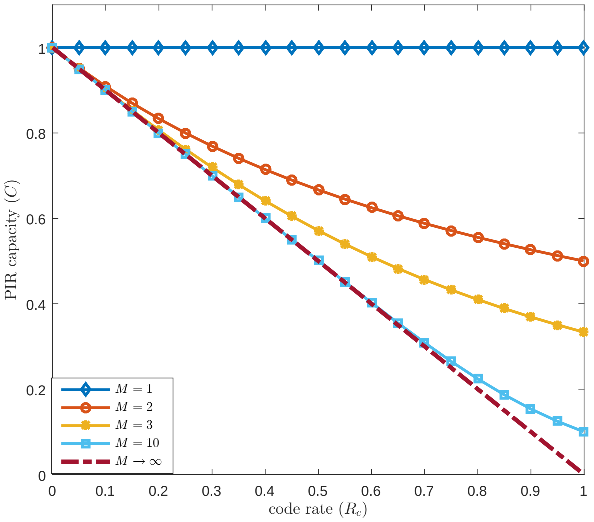

Fig. 2. PIR capacity versus R_c.

The capacity in (12) 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 (12) 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 user should download all messages as $M \to \infty$. 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 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 the extended version of this paper [15]. 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.

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_m,
$$\mathbf{x}_i^{[m]} = \mathbf{w}_{\pi(i)}^{[m]}, \quad i \in \{1, \cdots, \tilde{L}\} \tag{15}$$
where $\pi(\cdot)$ is a random interleaver 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 PIR bits from different rows from database 1 (DB1) and sets round index $i = 1$.
3) *Symmetry across databases:* The user downloads K^{M-1} desired bits each from a different row from each database. 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. 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. 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 (6). 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.
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. Calculation of the Achievable Rate

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. The achievable rate is,

$$\frac{1}{R} = 1 + \frac{\text{total undesired equations}}{\text{total desired equations}} \tag{16}$$
$$= 1 + \frac{\sum_{i=1}^{M-1}\binom{M-1}{i}K^{M-i}(N-K)^{i-1}}{\sum_{i=0}^{M-1}\binom{M-1}{i}K^{M-1-i}(N-K)^{i}} \tag{17}$$
$$= 1 + \frac{\frac{K}{N-K}\sum_{i=1}^{M-1}\binom{M-1}{i}K^{M-1-i}(N-K)^{i}}{N^{M-1}} \tag{18}$$
$$= 1 + \frac{\frac{K}{N-K}\left(N^{M-1} - K^{M-1}\right)}{N^{M-1}} \tag{19}$$
$$= 1 + \frac{K}{N-K}\left(1 - R_c^{M-1}\right) \tag{20}$$
$$= \frac{1 - R_c^M}{1 - R_c} \tag{21}$$

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. Converse Proof

A. Notations and Simplifications

We follow the notations and simplifications of [7]. We define,

$$\mathcal{Q} \triangleq \{Q_n^{[m]} : m \in \{1, \cdots, M\},\ n \in \{1, \cdots, N\}\} \tag{22}$$

$$A_{n_1:n_2}^{[m]} \triangleq \{A_{n_1}^{[m]}, \cdots, A_{n_2}^{[m]}\},\ n_1 \leq n_2,\ n_1, n_2 \in \{1, \cdots, N\} \quad (23)$$

We use $(n_1 : n_2) \mod N$ to denote the circular indices from n_1 to n_2, i.e., if $n \geq N$, then n is replaced by $(n \bmod N)$. Without loss of generality, we can make the following simplifications [7]:

1) We can assume that the PIR scheme is symmetric. This can be assumed without loss of generality, since for any asymmetric PIR scheme, one can construct an equivalent symmetric retrieval scheme that has the same retrieval rate by replicating all permutations of databases and messages with appropriate time sharing.
2) We can invoke the non-colluding privacy constraint by fixing the query to one database to be the same irrespective of the desired message, i.e., $Q_n^{[m]} = Q_n, m \in \{1, \cdots, M\}$ for some $n \in \{1, \cdots, N\}$. This implies that $A_n^{[m]} = A_n, m \in \{1, \cdots, M\}$. This simplification is without loss of generality, since the queries are independent of the desired message index. Note that the index of this database can be chosen arbitrarily, and hence without loss of generality, we choose it to be the first database, i.e., $A_1^{[m]} = A_1, \forall m$.

We first state the following lemma whose proof can be found in [7, Lemma 1].

Lemma 1 (Symmetry [7]) *Without loss of generality, we have*

$$H(A_n^{[1]}|W_2, \cdots, W_M, \mathcal{Q}) = H(A_1^{[1]}|W_2, \cdots, W_M, \mathcal{Q}), \quad n \in \{1, \cdots, N\} \quad (24)$$

$$H(A_1^{[1]}|W_2, \cdots, W_M, \mathcal{Q}) \geq \frac{H(W_1)}{N} = \frac{L}{N} \quad (25)$$

$$H(A_1|\mathcal{Q}) = H(A_n^{[m]}|\mathcal{Q}), m \in \{1, \cdots, M\}, n \in \{1, \cdots, N\} \quad (26)$$

We note that the equality in (26) remains true if the answer strings are conditioned on any subset $W_{\mathcal{S}} = \{W_i : i \in \mathcal{S}\}$ of messages, i.e.,

$$H(A_1|W_{\mathcal{S}}, \mathcal{Q}) = H(A_n^{[m]}|W_{\mathcal{S}}, \mathcal{Q}), \quad \forall m, n \quad (27)$$

because otherwise the nth database can break the privacy requirement by conditioning the answer strings on $W_{\mathcal{S}}$ before responding to the user, and from the difference in lengths, the database can infer some information about the desired message index. Next, we present Lemma 2, which implies that the answers from any K distinct databases are statistically independent. The proof can be found in [15].

Lemma 2 (Independence of answers of any K databases) *For any set $\mathcal{K}$ of databases such that $|\mathcal{K}| = K$,*

$$H(A_{\mathcal{K}}^{[1]}|\mathcal{Q}) = KH(A_1^{[1]}|\mathcal{Q}) \quad (28)$$

Furthermore, (28) is true if conditioned on any subset of messages $W_{\mathcal{S}}$, i.e.,

$$H(A_{\mathcal{K}}^{[1]}|W_{\mathcal{S}}, \mathcal{Q}) = KH(A_1^{[1]}|W_{\mathcal{S}}, \mathcal{Q}) \quad (29)$$

B. Converse Proof of the Case $M = 2$

The converse proof is constructed by mathematical induction over messages M. The base induction step is the case $M = 2$. First, we need the following lemma, which obtains a lower bound on the interference generated by the undesired message. The proof can be found in [15].

Lemma 3 (Interference lower bound) *For the case $M = 2$, the uncertainty on the interference from W_2 in the answers $A_{1:N}^{[1]}$ is lower bounded as,*

$$H(A_{1:N}^{[1]}|W_1, \mathcal{Q}) \geq \frac{K}{N} H(W_1) = \frac{KL}{N} \quad (30)$$

Now, we are ready to derive the converse proof for the case $M = 2$,

$$\begin{aligned} L &= H(W_1) && (31)\\ &= H(W_1|\mathcal{Q}) && (32)\\ &= H(W_1|\mathcal{Q}) - H(W_1|A_{1:N}^{[1]}, \mathcal{Q}) && (33)\\ &= I(W_1; A_{1:N}^{[1]}|\mathcal{Q}) && (34)\\ &= H(A_{1:N}^{[1]}|\mathcal{Q}) - H(A_{1:N}^{[1]}|W_1, \mathcal{Q}) && (35)\\ &\leq H(A_{1:N}^{[1]}|\mathcal{Q}) - \frac{KL}{N} && (36)\\ &\leq \sum_{n=1}^{N} H(A_n^{[1]}|\mathcal{Q}) - \frac{KL}{N} && (37) \end{aligned}$$

where (32) follows from the independence of the queries and the messages, (33) follows from the reliability constraint (10) for W_1, (36) follows from Lemma 3, and (37) follows from the fact that conditioning does not increase entropy. Hence, using Lemma 1,

$$NH(A_1|\mathcal{Q}) \geq L\left(1 + \frac{K}{N}\right) \quad (38)$$

Then, using (11), the retrieval rate is upper bounded by,

$$R = \frac{L}{\sum_{n=1}^{N} H(A_n^{[1]})} \leq \frac{L}{NH(A_1|\mathcal{Q})} \leq \frac{1}{1 + \frac{K}{N}} \quad (39)$$

C. Converse Proof for $M \geq 3$

We use a technique similar to that in [7]. In the sequel, we derive an inductive relation that can be used in addition to the base induction step of $M = 2$ to obtain a matching upper bound for the achievable rate in (12). We need the following lemma which upper bounds the uncertainty on the answer strings after knowing one of the interference messages. The proof can be found in [15]. Lemma 4 captures the main aspects of the problem, namely, coding, privacy, and alignment.

Lemma 4 (Interference conditioning lemma) *The remaining uncertainty on the answer strings after conditioning on one of the interfering messages is upper bounded by,*

$$H(A_{1:N}^{[2]}|W_1, \mathcal{Q}) \leq \frac{N}{K}\left(NH(A_1|\mathcal{Q}) - L\right) \quad (40)$$

Now, we derive the inductive relation for general M,

$$L = H(W_2) \tag{41}$$
$$= H(W_2|W_1, \mathcal{Q}) \tag{42}$$
$$= H(W_2|W_1, \mathcal{Q}) - H(W_2|A^{[2]}_{1:N}, W_1, \mathcal{Q}) \tag{43}$$
$$= I(A^{[2]}_{1:N}; W_2|W_1, \mathcal{Q}) \tag{44}$$
$$= H(A^{[2]}_{1:N}|W_1, \mathcal{Q}) - H(A^{[2]}_{1:N}|W_1, W_2, \mathcal{Q}) \tag{45}$$
$$\leq \frac{N}{K}\left(NH(A_1|\mathcal{Q}) - L\right) - H(A^{[2]}_{1:K}|W_1, W_2, \mathcal{Q}) \tag{46}$$
$$= \frac{N}{K}\left(NH(A_1|\mathcal{Q}) - L\right) - KH(A_1|W_1, W_2, \mathcal{Q}) \tag{47}$$

where (42) follows from the independence of $(W_1, W_2, \mathcal{Q})$, (43) follows from the reliability constraint on W_2, (46) follows from Lemma 4 and the non-negativity of the entropy function, and (47) follows from the independence of any K answer strings as proved in Lemma 2. The result in (47) gives,

$$\left(1 + \frac{N}{K}\right) L \leq \frac{N}{K} NH(A_1|\mathcal{Q}) - KH(A_1|W_1, W_2, \mathcal{Q}) \tag{48}$$

Leading to the following induction relation,

$$NH(A_1|\mathcal{Q}) \geq \left(1 + \frac{K}{N}\right) L + \frac{K^2}{N} H(A_1|W_1, W_2, \mathcal{Q}) \tag{49}$$

The relation (49) is the desired induction step as it forms a relationship between the original problem and a reduced PIR problem with $(M-2)$ messages. We note that this relation includes the induction relation in [7] as a special case with $K = 1$.

We state the induction hypothesis for M messages as follows,

$$NH(A_1|\mathcal{Q}) \geq L \sum_{i=0}^{M-1} \left(\frac{K}{N}\right)^i \tag{50}$$

We proved this relation for $M = 2$ in (38) as the base induction step. Now, assuming that this is true for M messages, we will prove it for $(M+1)$ messages based on (49) and (50). Since $H(A_1|W_1, W_2, \mathcal{Q})$ represents $H(A_1|\mathcal{Q})$ for a reduced PIR problem with $(M-1)$ messages, from the induction hypothesis, we have,

$$NH(A_1|W_1, W_2, \mathcal{Q}) \geq L \sum_{i=0}^{M-2} \left(\frac{K}{N}\right)^i \tag{51}$$

Substituting this in (49),

$$NH(A_1|\mathcal{Q}) \geq \left(1 + \frac{K}{N}\right) L + L \cdot \frac{K^2}{N^2} \sum_{i=0}^{M-2} \left(\frac{K}{N}\right)^i \tag{52}$$
$$\geq L \sum_{i=0}^{M} \left(\frac{K}{N}\right)^i \tag{53}$$

which concludes the induction argument. Consequently, the upper bound for the coded PIR problem starting from (11) is,

$$R = \frac{L}{\sum_{n=1}^{N} H(A^{[1]}_n)} \tag{54}$$
$$\leq \frac{L}{NH(A_1|\mathcal{Q})} \tag{55}$$
$$= \frac{1}{\sum_{i=0}^{M-1} \left(\frac{K}{N}\right)^i} \tag{56}$$
$$= \frac{1}{\sum_{i=0}^{M-1} R_c^i} = \frac{1 - R_c}{1 - R_c^M} \tag{57}$$

where (56) follows from (50).

VI. Conclusions

In this paper, we considered the private information retrieval (PIR) problem over 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) MDS 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 storage code and the retrieval procedure. The result also establishes a fundamental tradeoff between the code rate and the PIR capacity.

References

[1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan. Private information retrieval. *Journal of the ACM*, 45(6):965–981, 1998.

[2] W. Gasarch. A survey on private information retrieval. In *Bulletin of the EATCS*, 2004.

[3] R. Ostrovsky and W. Skeith III. A survey of single-database private information retrieval: Techniques and applications. In *International Workshop on Public Key Cryptography*, pages 393–411. Springer, 2007.

[4] S. Yekhanin. Private information retrieval. *Communications of the ACM*, 53(4):68–73, 2010.

[5] R. Tajeddine and S. El Rouayheb. Private information retrieval from MDS coded data in distributed storage systems. In *IEEE ISIT*, July 2016.

[6] T. Chan, S. Ho, and H. Yamamoto. Private information retrieval for coded storage. In *IEEE ISIT*, June 2015.

[7] H. Sun and S. Jafar. The capacity of private information retrieval. 2016. Available at arXiv:1602.09134.

[8] S. Jafar. Blind interference alignment. *IEEE Journal of Selected Topics in Signal Processing*, 6(3):216–227, June 2012.

[9] H. Sun and S. Jafar. Blind interference alignment for private information retrieval. 2016. Available at arXiv:1601.07885.

[10] H. Sun and S. Jafar. The capacity of robust private information retrieval with colluding databases. 2016. Available at arXiv:1605.00635.

[11] H. Sun and S. Jafar. The capacity of symmetric private information retrieval. 2016. Available at arXiv:1606.08828.

[12] A. Dimakis, K. Ramchandran, Y. Wu, and C. Suh. A survey on network codes for distributed storage. *Proceedings of the IEEE*, 99(3):476–489, 2011.

[13] N. B. Shah, K. V. Rashmi, and K. Ramchandran. One extra bit of download ensures perfectly private information retrieval. In *IEEE ISIT*, June 2014.

[14] A. Fazeli, A. Vardy, and E. Yaakobi. PIR with low storage overhead: coding instead of replication. 2015. Available at arXiv:1505.06241.

[15] K. Banawan and S. Ulukus. The capacity of private information retrieval from coded databases. *IEEE Trans. on Info. Theory*. Submitted Sepetember 2016. Also available at arXiv:1609.08138.