# Semantic Private Information Retrieval From MDS-Coded Databases

Sajani Vithana[1], Karim Banawan[2], and Sennur Ulukus[1]

[1]Department of Electrical and Computer Engineering, University of Maryland
[2]Electrical Engineering Department, Faculty of Engineering, Alexandria University

***Abstract*—We investigate the problem of semantic private information retrieval (PIR) from coded databases, where a user requires to download a message out of $M$ independent messages, without revealing its identity to the databases. These messages are coded using an $(N, K)$ MDS code and stored in $N$ non-colluding databases. The $M$ messages are allowed to have different semantics, e.g., different sizes and different probabilities of retrieval. We characterize the exact capacity of semantic PIR with coded databases, and provide an achievable scheme with non-uniform subpacketization. We show that the retrieval rate of semantic PIR with coded databases outperforms that of classical PIR with coded databases when the effects of zero padding shorter messages are taken into account.**

## I. Introduction

The classical private information retrieval (PIR) problem refers to a setting, where a user downloads a required message from a system of non-colluding replicated databases containing a number of messages while not revealing the identity of the downloaded message to the databases. This problem was first introduced by Chor et al. in [1]. The information-theoretic characterization of the classical problem is presented in [2]. In [2], the performance metric is the retrieval rate, which is the ratio between the desired message bits and the total download. The supermum of all achievable retrieval rates is called the PIR capacity. The PIR capacity of many variants of the problem have been studied (see for instance [3]–[43]).

The most closely related works to ours are PIR from MDS-coded databases, which is studied extensively in [44]–[53], and the semantic PIR problem from replicated databases in [54]. In semantic PIR, the messages exhibit different semantics. These semantics include message sizes and prior probabilities. This is motivated in practice by the fact that files in storage systems have different sizes (e.g., databases may simultaneously store text files and video files with vastly different message sizes) and popularity levels (e.g., databases may simultaneously store trending and stale files). The work of [54] characterizes the capacity of semantic PIR from replicated databases. On the other hand, MDS-coded storage systems provide an increased level of reliability for information stored in systems of databases without incurring the excessive storage cost of direct replication. The capacity of the MDS-coded PIR problem is characterized in [44] to be a function of the MDS code parameters. Due to the relevance of MDS coding and semantic heterogeneity in practice, it is desirable to provide a viable PIR scheme that efficiently optimizes the PIR rate corresponding to the storage code structure and the given message semantics.

In this paper, we aim at characterizing the capacity of semantic PIR from $N$ non-colluding coded databases using an arbitrary $(N, K)$ MDS storage code. In this problem, there are $M$ messages with different semantics, i.e., each message has a different message size and a different popularity level. More specifically, the storage system possesses messages $(W_1, \ldots, W_M)$ in matrix form, with $K$ columns and $L_i, i \in \{1, \ldots, M\}$ rows. Each row is mapped to the content of the databases via an $(N, K)$ MDS storage code and then distributed to a system of $N$ non-colluding databases. Furthermore, the messages have arbitrary probabilities of retrieval $(p_i, i \in \{1, \ldots, M\})$ to reflect the popularity levels. We investigate the interplay between the storage code parameters, non-uniform subpacketization, and its effect on the PIR rate, i.e., for a given $(N, K)$ MDS code, how can we design the retrieval parameters to exploit the heterogeneity of the message semantics to maximize the retrieval rate?

We characterize the semantic PIR from $(N, K)$ MDS-coded databases as $C = \left(\frac{L_1}{\mathbb{E}[L]} + \left(\frac{K}{N}\right)\frac{L_2}{\mathbb{E}[L]} + \ldots + \left(\frac{K}{N}\right)^{M-1}\frac{L_M}{\mathbb{E}[L]}\right)^{-1}$, where the expected number of rows $\mathbb{E}[L]$ is with respect to the retrieval probability distribution. We provide an achievable scheme, which is an extension to the scheme introduced in [44]. The main difference of our scheme compared to its counterpart in [44] is that our achievable scheme uses non-uniform subpacketization that is parameterized by the message sizes and the storage code parameters. Specifically, a single application of the scheme results in different number of useful downloads (which correspond to the number of rows in the required message) for different message requirements of the user. The converse is extended from [44] with the incorporation of the heterogeneity of message sizes. Compared to the achievable MDS-coded PIR rate with zero padding, i.e., when all messages are zero-padded to match the size of the largest message, the semantic PIR capacity expression shows a strict rate gain when message semantics are not identical.

## II. Problem Formulation

We consider an $(N, K)$ MDS coded distributed storage system containing $M$ independent messages. The messages are allowed to have different lengths and different prior probabilities of retrieval. The prior probability distribution

$(p_i, i \in \{1,\dots,M\})$ is known by the databases and the user. Each message $W_i, i \in \{1,\dots,M\}$ is represented as a matrix in $\mathbb{F}_q^{L_i \times K}$, where the $L_i \times K$ elements of $W_i$ are chosen uniformly and independently from $\mathbb{F}_q$. Without loss of generality, we assume that the messages are ordered with respect to their sizes, such that $L_1 \geq L_2 \geq \cdots \geq L_M$. The message sizes can be expressed in $q$-ary symbols as,

$$H(W_i) = KL_i, \quad i = 1,\dots,M. \tag{1}$$

The generator matrix of the $(N,K)$ code $H$ is a $\mathbb{F}_q^{K\times N}$ matrix, which is represented as $H = [h_1, h_2, \dots, h_N]$, where $h_i \in \mathbb{F}_q^K, i \in [N]$. The MDS property implies that any combination of up to $K$ columns of $H$ are linearly independent. Let the $j$th row of $W_i$ be denoted by $x_j^{[i]}$, $j = 1,\dots,L_i$ and $i = 1,\dots,M$. The $n$th database stores a projection of this row as $x_j^{[i]}h_n$, $n = 1,\dots,N$.

In order to retrieve a message $W_i$, the user sends query $Q_n^{[i]}$ to the $n$th database, $n = 1,\dots,N$. The objective is to retrieve $W_i$ without revealing the index $i$ to any database. The queries and messages are independent of each other. Once the databases receive the queries from the user, they generate answer strings denoted by $A_n^{[i]}, n = 1\dots,N$, to send back to the user. These answer strings are deterministic functions of the stored coded messages and the received query.

An achievable PIR scheme satisfies the following constraints:

**Correctness:** The user should be able to perfectly decode the desired message using the received answer strings. Thus,

$$H(W_i|A_1^{[i]},\dots,A_N^{[i]},Q_1^{[i]},\dots,Q_N^{[i]}) = 0, \ \ i = 1,\dots,M \tag{2}$$

**Privacy:** The queries sent to the databases by the user should not leak any information on the required message index. Thus, the joint PMFs of queries, answers and messages need to satisfy the following for $n \in [N], j \in [M], \ i \neq j$

$$(Q_n^{[i]}, A_n^{[i]}, W_1,\dots,W_M) \sim (Q_n^{[j]}, A_n^{[j]}, W_1,\dots,W_M) \tag{3}$$

An achievable semantic PIR scheme for a coded distributed system is a scheme that satisfies the correctness and privacy conditions in (2) and (3). The achievable rate of the semantic PIR scheme with coded databases is defined as,

$$R = \frac{\mathbb{E}[H(W_i)]}{\mathbb{E}[D]} = \frac{K\mathbb{E}[L]}{\mathbb{E}[D]} \tag{4}$$

where $\mathbb{E}[L] = \sum_{i=1}^M p_i L_i$ is the expected number of rows in a message and $\mathbb{E}[D]$ is the expected number of total bits downloaded. The expectation $\mathbb{E}[\cdot]$ is with respect to the a priori probability distribution. The capacity of semantic PIR from coded distributed databases is the supremum of the expected retrieval rates over all achievable schemes.

## III. Main Results and Discussions

In this section, we present the capacity of semantic PIR with MDS-coded databases, which depends on the number of messages and their semantics as well as the MDS storage code.

**Theorem 1** *The capacity of semantic PIR with an $(N,K)$ MDS-coded distributed storage system containing $M$ messages ($W_i, i \in \{1,\dots,M\}$) with $L_i$ rows in each of their matrix representations (arranged in decreasing order as $L_1 \geq L_2 \geq \cdots \geq L_M$), and prior probabilities $p_i$, is*

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \left(\frac{K}{N}\right)\frac{L_2}{\mathbb{E}[L]} + \ldots + \left(\frac{K}{N}\right)^{M-1}\frac{L_M}{\mathbb{E}[L]}\right)^{-1} \tag{5}$$

*where $\mathbb{E}[L] = \sum_{i=1}^M p_i L_i$.*

The achievability proof of Theorem 1 is presented in Section IV and the converse proof is presented in Section V.

**Corollary 1** *For messages with arbitrary lengths and for any given $(N,K)$ MDS code, the semantic PIR capacity with coded databases outperforms the achievable PIR rate with zero-padding, when message lengths are not identical.*

**Proof:** Let $R_c$ and $C_c$ be the achievable rate and the capacity of classical coded PIR, respectively. For a setting with an $(N,K)$ MDS code and $M$ messages with $L_1 \geq L_2 \geq \ldots \geq L_M$, the classical coded PIR scheme zero pads messages $W_2,\dots,W_M$ such that all messages contain $L_1$ rows. Therefore, the download cost in classical coded PIR is $D = \frac{L_1}{C_c}$. However, since the number of rows of all messages are not the same, the effective retrieval rate of classical coded PIR is,

$$R_c = \left(\frac{L_1}{\mathbb{E}[L]} + \left(\frac{K}{N}\right)\frac{L_1}{\mathbb{E}[L]} + \ldots + \left(\frac{K}{N}\right)^{M-1}\frac{L_1}{\mathbb{E}[L]}\right)^{-1} \tag{6}$$

Since $L_1 \geq \ldots \geq L_M$, comparison of (6) and (5) shows that semantic coded PIR outperforms classical coded PIR in terms of the retrieval rate. ■

## IV. Achievability Proof

In this section, we present a scheme that achieves the capacity expression in Theorem 1.

### A. Achievable Scheme

This scheme is an extension to the scheme in [44]. All queries of this scheme are in blocks of $\ell$-sums of coded message bits for $\ell \in \{1,\dots,M\}$. Our proposed scheme optimizes the number of queries in each block of $\ell$-sums, based on the message lengths $\{L_i\}_{i=1}^M$. This scheme utilizes non-uniform subpacketization, as opposed to uniform subpacketization used in [44]. The general achievability scheme is as follows:

1) **Message index assignment, permutation of row indices and calculation of retrieval parameters**: Assign message indices in the descending order of the message sizes, i.e., $L_1 \geq L_2 \geq \cdots \geq L_M$. Permute the rows of all messages independently from each other using random permutations, privately from the databases. Calculate the retrieval parameters $\upsilon_1,\dots,\upsilon_M$ using (14). Assume that the desired message is $W_j$ in the sequel.
2) **Singletons:** Download $\upsilon_j$ different coded bits corresponding to message $W_j$ from the $n$th database, for

$n \in [N]$. From each database, download $\upsilon_i$ coded bits of $W_i$, $i \in [M], i \neq j$, such that there exists $K$ coded bits that are downloaded from $K$ different databases that correspond to the same row of $W_i$ (i.e., for some row $t$, $x_t^{[i]}h_{n_1}, \ldots, x_t^{[i]}h_{n_K}$ for $n_1, \ldots, n_K \in [N]$ and $n_i \neq n_j$ for $i, j \in \{1, \ldots, K\}$). This is required to decode the side information. Hence, for $W_i, i \neq j$, there are $N\upsilon_i$ coded bits corresponding to $\frac{N\upsilon_i}{K}$ different rows.

3) **Sums of two elements:** There are two types of blocks in this step. The first block is the sums involving bits of the desired message, $W_j$, and the other block is the sums that do not have any bits from $W_j$. In the first block, make use of the side information (singletons corresponding to $W_i, i \neq j$) downloaded in the previous step. Consider a 2-sum corresponding to coded bits of $W_j$, $W_i$, $i \neq j$. For a given database, download $(\frac{N}{K} - 1)\min\{\upsilon_i, \upsilon_j\}$ 2-sums utilizing all the side information bits of $W_i$ from other databases. Each 2-sum contains a coded bit corresponding to a new row of $W_j$ and a coded bit of $W_i$ which corresponds to a row that was already decoded from $K$ different databases (which does not include this database) in the previous step.
The second block of 2-sums contains coded bits corresponding to $W_{i_1}$, $W_{i_2}$, $i_1 \neq i_2 \neq j$. Download $\left(\frac{N}{K} - 1\right)\min\{\upsilon_{i_1}, \upsilon_{i_2}\}$ 2-sums each from each databases of the form $(x_{r_a}^{[i_1]} + x_{r_b}^{[i_2]})h_n, n \in [N]$, where $r_a$ and $r_b$ are new rows of $W_{i_1}$ and $W_{i_2}$. Each pair of rows $(r_a, r_b)$ must be downloaded from $K$ different databases for correct decoding of $x_{r_a}^{[i_1]} + x_{r_b}^{[i_2]}$. Thus, the second block contains $N\left(\frac{N}{K} - 1\right)\min\{\upsilon_{i_1}, \upsilon_{i_2}\}$ bits, which corresponds to $\frac{N\left(\frac{N}{K}-1\right)\min\{\upsilon_{i_1},\upsilon_{i_2}\}}{K}$ different rows.
4) **Sums of $\ell$ elements:** There are two types of blocks similar to sums of two. the first block contains queries of the form $(x_r^{[j]} + x_{r_1}^{[i_1]} + \ldots + x_{r_{\ell-1}}^{[i_{\ell-1}]})h_n$ for $i_1 \neq \ldots \neq i_{\ell-1} \neq j$ where $x_r^{[j]}$ is a new row of $W_j$ and $x_{r_1}^{[i_1]} + \ldots + x_{r_{\ell-1}}^{[i_{\ell-1}]}$ is an already decoded $(\ell - 1)$-sum from the second block in the previous step. For a given database and given $i_1, \ldots, i_{\ell-1}$, there are $\left(\frac{N}{K} - 1\right)^{\ell-1} \upsilon_{\min\{j, i_1, \ldots, i_{\ell-1}\}}$ such $\ell$-sums.
The second block of $\ell$-sums contains queries of the form $(x_{t_1}^{[i_1]} + \ldots + x_{t_\ell}^{[i_\ell]})h_n$ for $i_1 \neq \ldots \neq i_\ell \neq j$ where $(t_1, \ldots, t_\ell)$ are new rows of $W_{i_1}, \ldots, W_{i_\ell}$ which are repeated at $K$ different databases for the correct decoding of $x_{t_1}^{[i_1]}, \ldots, x_{t_\ell}^{[i_\ell]}$. There are $\left(\frac{N}{K} - 1\right)^{\ell-1} \upsilon_{\min\{i_1, \ldots, i_\ell\}}$ such $\ell$-sums in a given database for a given $\ell$-tuple $(i_1, \ldots, i_\ell)$. Therefore, there are $N\left(\frac{N}{K} - 1\right)^{\ell-1} \upsilon_{\min\{i_1, \ldots, i_\ell\}}$ $\ell$-sums in total, belonging to $\frac{N\left(\frac{N}{K}-1\right)^{\ell-1}\upsilon_{\min\{i_1,\ldots,i_\ell\}}}{K}$ different $\ell$-tuples of rows of $(W_{i_1}, \ldots, W_{i_\ell})$.
5) **Repeat the process up to sums of $M$ elements.**
6) **Query Repetition:** To decode each row of $W_j$, the user has to repeat the above process $K$ times, while shifting the queries that contain rows of $W_j$ to its neighboring database and by choosing new sets of rows of $W_i, i \in \{1, \ldots, M\}, i \neq j$ in each repetition. The $K$ different linear combinations of the elements of each row of $W_j$ resulting from this process makes it possible to recover all $L_j \times K$ elements of $W_j$.

### *B. Rate of Semantic PIR Scheme with Coded Databases*

$\mathbb{E}[D]$ needs to be constant for any desired message to guarantee user privacy. Therefore, it suffices to calculate $\mathbb{E}[D]$ when retrieving $W_j$ for some $j$. Within one round of queries, there are $\sum_{i=1}^{M} N\upsilon_i$ number of singletons and $\sum_{i=\ell}^{M} N\left(\frac{N}{K} - 1\right)^{\ell-1} \upsilon_i \binom{i-1}{\ell-1}$ number of sums of $\ell$ elements,

$$\frac{\mathbb{E}[D]}{K} = \sum_{i=1}^{M} N\upsilon_i + \sum_{\ell=2}^{M}\sum_{i=\ell}^{M} N\left(\frac{N}{K} - 1\right)^{\ell-1} \upsilon_i \binom{i-1}{\ell-1} \quad (7)$$

$$= N\left[\sum_{i=1}^{M} \upsilon_i + \sum_{\ell=2}^{M} \upsilon_\ell \sum_{i=2}^{\ell} \left(\frac{N}{K} - 1\right)^{i-1} \binom{\ell-1}{i-1}\right] \quad (8)$$

$$= K\left[\frac{N}{K}\upsilon_1 + \sum_{\ell=2}^{M} \upsilon_\ell \left(\frac{N}{K}\right)^{\ell}\right] = K\sum_{\ell=1}^{M} \left(\frac{N}{K}\right)^{\ell} \upsilon_\ell \quad (9)$$

For $\mathbb{E}[L]$, we calculate the total number of useful bits (bits of $W_j$). Considering a single round of queries, there are $N\upsilon_j$ rows of $W_j$ retrieved in terms of singletons. The number of rows retrieved when $W_j$ is the shortest message in an $\ell$-sum containing a row of $W_j$ is $N\left(\frac{N}{K} - 1\right)^{\ell-1} \binom{j-1}{\ell-1}\upsilon_j$. The number of rows retrieved when $W_i, i \neq j$ is the shortest message in an $\ell$-sum containing a row of $W_j$ is $N\left(\frac{N}{K} - 1\right)^{\ell-1} \binom{i-2}{\ell-2}\upsilon_i$. Let $U_j$ denote the total number of useful bits retrieved. Therefore, $\frac{U_j}{K}$ can be calculated by counting the number of different rows of $W_j$ that correspond to all coded bits as,

$$\begin{aligned}\frac{U_j}{K} &= N\upsilon_j + \sum_{\ell=2}^{j} N\left(\frac{N}{K} - 1\right)^{\ell-1} \binom{j-1}{\ell-1}\upsilon_j \\ &\quad + \sum_{\ell=2}^{j}\sum_{i=j+1}^{M} N\left(\frac{N}{K} - 1\right)^{\ell-1} \binom{i-2}{\ell-2}\upsilon_i \\ &\quad + \sum_{\ell=j+1}^{M}\sum_{i=\ell}^{M} N\left(\frac{N}{K} - 1\right)^{\ell-1} \binom{i-2}{\ell-2}\upsilon_i \qquad (10)\end{aligned}$$

$$\begin{aligned}&= N\upsilon_j \sum_{\ell=0}^{j-1} \gamma^\ell \binom{j-1}{\ell} + N\upsilon_{j+1}\gamma \sum_{\ell=0}^{j-1} \gamma^\ell \binom{j-1}{\ell} \\ &\quad + N\upsilon_{j+2}\gamma \sum_{\ell=0}^{j} \gamma^\ell \binom{j}{\ell} + \cdots + N\upsilon_M \gamma \sum_{\ell=0}^{M-2} \gamma^\ell \binom{M-2}{\ell} \qquad (11)\end{aligned}$$

$$= K\left[\left(\frac{N^j}{K^j}\right)\upsilon_j + \gamma \sum_{i=j+1}^{M} \left(\frac{N}{K}\right)^{i-1} \upsilon_i\right], \quad (12)$$

where $\gamma = \frac{N}{K} - 1$. Thus, the subpacketiation for $W_i$ can be defined as $\frac{U_j}{K}$, which represents the number of rows of $W_j$, that can be retrieved by a single use of the scheme.

Since the total number of rows of $W_j, j \in \{1,\dots,M\}$ have to be a common multiple of their own subpacketizations,

$$L_j = \alpha \frac{U_j}{K}, \quad j = 1,\dots,M \tag{13}$$

for some $\alpha \in \mathbb{N}$. Solving (12), (13) for the retrieval parameters $\upsilon_1,\dots,\upsilon_M$ with $\gamma = \frac{N}{K} - 1$ leads to:

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \vdots \\ \upsilon_M \end{bmatrix} = \frac{1}{K\alpha} \begin{bmatrix} \frac{K}{N} & -\left(\frac{K}{N}\right)^2 \gamma & \dots & -\left(\frac{K}{N}\right)^M \gamma \\ 0 & \left(\frac{K}{N}\right)^2 & \dots & -\left(\frac{K}{N}\right)^M \gamma \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \left(\frac{K}{N}\right)^M \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_M \end{bmatrix} \tag{14}$$

In order for the values of $\upsilon_i, i \in \{1,\dots,M\}$ to be integers, this scheme requires all $L_i$'s to be a multiples of $N^M$. Here, $\alpha$ should be chosen to be the greatest common divisor (gcd) of the elements of the vector resulting from multiplying the matrix and the vector on the right side of (14). This allows the shortest subpacketization levels for all messages.

The total and useful numbers of bits downloaded ((9) and (12) respectively) are both within one subpacketization level. These downloads are repeated $\alpha$ times to download the entire message; see also (13). Thus, the achievable rate is given by,

$$R = \frac{K\mathbb{E}[L]}{\mathbb{E}[D]} = \frac{K\sum_{i=1}^{M} p_i L_i}{\alpha K^2 \sum_{i=1}^{M} \frac{N^i}{K^i} \upsilon_i} \tag{15}$$

$$= \frac{\mathbb{E}[L]}{\alpha K \frac{1}{K\alpha} \sum_{i=1}^{M} \frac{N^i}{K^i} \left[ \left(\frac{K^i}{N^i}\right) L_i - \left(\frac{N}{K} - 1\right) \sum_{t=i+1}^{M} \left(\frac{K^t}{N^t}\right) L_t \right]} \tag{16}$$

$$= \frac{\mathbb{E}[L]}{\sum_{i=1}^{M} \left[ L_i - \left(\frac{N}{K} - 1\right) \sum_{t=i+1}^{M} \left(\frac{K^{t-i}}{N^{t-i}}\right) L_t \right]} \tag{17}$$

$$= \frac{\mathbb{E}[L]}{L_1 + L_2\left(\frac{K}{N}\right) + \sum_{i=3}^{M} L_i \left[1 - \left(1 - \frac{K^{i-1}}{N^{i-1}}\right)\right]} \tag{18}$$

$$= \left( \frac{L_1}{\mathbb{E}[L]} + \left(\frac{K}{N}\right) \frac{L_2}{\mathbb{E}[L]} + \dots + \left(\frac{K}{N}\right)^{M-1} \frac{L_M}{\mathbb{E}[L]} \right)^{-1} \tag{19}$$

### C. A Representative Example for the Proposed Scheme

In this example, we consider a $(5,3)$ code with $M = 2$, $L_1 = 100$ and $L_2 = 50$. First, the row indices of $W_i \in \mathbb{F}_q^{L_i \times K}$ for $i = 1,2$ are independently and uniformly permuted. The rows of the first and the second messages after permutations are denoted by $x_r^{[i]}$ for $r \in \{1,\dots,L_i\}$ and $i = 1,2$. Messages are indexed such that the first message is the longer. The calculation of $\upsilon_i$, $i = 1,2$ is as follows:

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \end{bmatrix} = \frac{1}{3\alpha} \begin{bmatrix} \frac{3}{5} & -\frac{6}{25} \\ 0 & \frac{9}{25} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \tag{20}$$

where $\alpha = \text{gcd}\{\frac{L_1}{5} - \frac{2L_2}{25}, \frac{3L_2}{25}\} = \text{gcd}\{16,6\} = 2$, $\upsilon_1 = 8$ and $\upsilon_2 = 3$. The subpacketization levels of $W_1$ and $W_2$ in terms of the number of rows are $\frac{U_1}{3} = \frac{100}{2} = 50$ and $\frac{U_2}{3} = \frac{50}{2} = 25$, respectively. Assume that $W_1$ is the desired message in the sequel. Table I shows the queries sent to the databases to retrieve $W_1$.

First, download $\upsilon_1 = 8$ different coded bits of $W_1$ from each database as singletons. There are $40$ coded bits of this form in total, corresponding to $40$ different rows of $W_1$. Download $\upsilon_2 = 3$ coded bits of $W_2$ from each database such that three coded bits corresponding to the same row of $W_2$ is downloaded from three different databases for correct decoding of the downloaded rows of $W_2$. There are $15$ different coded bits of this form corresponding to $5$ different rows of $W_2$.

Next, download $\left(\frac{N}{K} - 1\right)\upsilon_2 = 2$ sums of the form $(x_{r_1}^{[1]} + x_{r_2}^{[2]})h_n$ from the $n$th database for $n = 1,\dots,N$. Each $x_{r_1}^{[1]}$ corresponds to a new row of $W_1$ and each $x_{r_2}^{[2]}$ corresponds to an already decoded row of $W_2$ in the previous step.

Finally, we repeat the above queries two extra times with the queries that include rows of $W_1$ shifted to the neighboring database and by choosing new rows of $W_2$ in each repetition.

The rate achieved by this scheme when downloading $W_1$ is $R_1 = \frac{3\times 50}{3\times 65} = \frac{10}{13}$, and the rate achieved by this scheme when downloading $W_2$ is $R_2 = \frac{3\times 25}{3\times 65} = \frac{5}{13}$. Therefore, the average rate $R$ achieved by the scheme is,

$$R = \frac{K(p_1 L_1 + p_2 L_2)}{p_1 D + p_2 D} = p_1 R_1 + p_2 R_2 = \frac{10}{13} p_1 + \frac{5}{13} p_2 \tag{21}$$

This matches the capacity expression in Theorem 1.

## V. Converse Proof

In this section, we provide an upper bound on the rate of semantic PIR from MDS-coded databases.

First, note that the answer strings generated by any set of $K$ databases are independent from each other, i.e.,

$$H(A_\Omega^{[m]} | Q_\Omega^{[m]}, W_\Delta) = \sum_{n \in \Omega} H(A_n^{[m]} | Q_n^{[m]}, W_\Delta), \tag{22}$$

where $\Omega \subset [N]$ such that $|\Omega| = K$, and $W_\Delta$ is any subset of messages. The proof can be found in [44, Lemma 1].

We begin the derivation of the upper bound on the rate from the expression for $\mathbb{E}[D]$. To satisfy the privacy constraint, $\mathbb{E}[D]$ is the same for any desired message. Without loss of generality, assume that the user required message is $W_1$,

$$\mathbb{E}[D] = \sum_{i=1}^{N} H(A_i^{[1]}) \tag{23}$$

$$\geq H(A_{1:N}^{[1]} | Q_{1:N}^{[1]}) \tag{24}$$

$$= I(W_{1:M}; A_{1:N}^{[1]} | Q_{1:N}^{[1]}) \tag{25}$$

$$= H(W_1) + I(W_{2:M}; A_{1:N}^{[1]}, Q_{1:N}^{[1]} | W_1) \tag{26}$$

$$\geq KL_1 + \frac{1}{\binom{N}{K}} \sum_{\Omega:|\Omega|=K} I(W_{2:M}; A_\Omega^{[1]}, Q_\Omega^{[1]} | W_1) \tag{27}$$

$$= KL_1 + \frac{1}{\binom{N}{K}} \sum_{\Omega:|\Omega|=K} H(A_\Omega^{[1]} | Q_\Omega^{[1]}, W_1) \tag{28}$$

$$= KL_1 + \frac{1}{\binom{N}{K}} \sum_{\Omega:|\Omega|=K} \sum_{n \in \Omega} H(A_n^{[1]} | Q_n^{[1]}, W_1) \tag{29}$$

$$= KL_1 + \frac{1}{\binom{N}{K}} \sum_{\Omega:|\Omega|=K} \sum_{n \in \Omega} H(A_n^{[2]} | Q_n^{[2]}, W_1) \tag{30}$$

| Database 1 | Database 2 | Database 3 | Database 4 | Database 5 |
|---|---|---|---|---|
| $x_1^{[1]}h_1$ | $x_9^{[1]}h_2$ | $x_{17}^{[1]}h_3$ | $x_{25}^{[1]}h_4$ | $x_{33}^{[1]}h_5$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_8^{[1]}h_1$ | $x_{16}^{[1]}h_2$ | $x_{24}^{[1]}h_3$ | $x_{32}^{[1]}h_4$ | $x_{40}^{[1]}h_5$ |
| $x_1^{[2]}h_1$ | $x_1^{[2]}h_2$ | $x_1^{[2]}h_3$ | $x_2^{[2]}h_4$ | $x_2^{[2]}h_5$ |
| $x_2^{[2]}h_1$ | $x_3^{[2]}h_2$ | $x_3^{[2]}h_3$ | $x_3^{[2]}h_4$ | $x_4^{[2]}h_5$ |
| $x_4^{[2]}h_1$ | $x_4^{[2]}h_2$ | $x_5^{[2]}h_3$ | $x_5^{[2]}h_4$ | $x_5^{[2]}h_5$ |
| $(x_{41}^{[1]}+x_3^{[2]})h_1$ | $(x_{43}^{[1]}+x_2^{[2]})h_2$ | $(x_{45}^{[1]}+x_2^{[2]})h_3$ | $(x_{47}^{[1]}+x_1^{[2]})h_4$ | $(x_{49}^{[1]}+x_1^{[2]})h_5$ |
| $(x_{42}^{[1]}+x_5^{[2]})h_1$ | $(x_{44}^{[1]}+x_5^{[2]})h_2$ | $(x_{46}^{[1]}+x_4^{[2]})h_3$ | $(x_{48}^{[1]}+x_4^{[2]})h_4$ | $(x_{50}^{[1]}+x_3^{[2]})h_5$ |
| $x_{33}^{[1]}h_1$ | $x_1^{[1]}h_2$ | $x_9^{[1]}h_3$ | $x_{17}^{[1]}h_4$ | $x_{25}^{[1]}h_5$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{40}^{[1]}h_1$ | $x_8^{[1]}h_2$ | $x_{16}^{[1]}h_3$ | $x_{24}^{[1]}h_4$ | $x_{32}^{[1]}h_5$ |
| $x_6^{[2]}h_1$ | $x_6^{[2]}h_2$ | $x_6^{[2]}h_3$ | $x_7^{[2]}h_4$ | $x_7^{[2]}h_5$ |
| $x_7^{[2]}h_1$ | $x_8^{[2]}h_2$ | $x_8^{[2]}h_3$ | $x_8^{[2]}h_4$ | $x_9^{[2]}h_5$ |
| $x_9^{[2]}h_1$ | $x_9^{[2]}h_2$ | $x_{10}^{[2]}h_3$ | $x_{10}^{[2]}h_4$ | $x_{10}^{[2]}h_5$ |
| $(x_{49}^{[1]}+x_8^{[2]})h_1$ | $(x_{41}^{[1]}+x_7^{[2]})h_2$ | $(x_{43}^{[1]}+x_7^{[2]})h_3$ | $(x_{45}^{[1]}+x_6^{[2]})h_4$ | $(x_{47}^{[1]}+x_6^{[2]})h_5$ |
| $(x_{50}^{[1]}+x_{10}^{[2]})h_1$ | $(x_{42}^{[1]}+x_{10}^{[2]})h_2$ | $(x_{44}^{[1]}+x_9^{[2]})h_3$ | $(x_{46}^{[1]}+x_9^{[2]})h_4$ | $(x_{48}^{[1]}+x_8^{[2]})h_5$ |
| $x_{25}^{[1]}h_1$ | $x_{33}^{[1]}h_2$ | $x_1^{[1]}h_3$ | $x_9^{[1]}h_4$ | $x_{17}^{[1]}h_5$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{32}^{[2]}h_1$ | $x_{40}^{[2]}h_2$ | $x_8^{[2]}h_3$ | $x_{16}^{[2]}h_4$ | $x_{24}^{[2]}h_5$ |
| $x_{11}^{[2]}h_1$ | $x_{11}^{[2]}h_2$ | $x_{11}^{[2]}h_3$ | $x_{12}^{[2]}h_4$ | $x_{12}^{[2]}h_5$ |
| $x_{12}^{[2]}h_1$ | $x_{13}^{[2]}h_2$ | $x_{13}^{[2]}h_3$ | $x_{13}^{[2]}h_4$ | $x_{14}^{[2]}h_5$ |
| $x_{14}^{[2]}h_1$ | $x_{14}^{[2]}h_2$ | $x_{15}^{[2]}h_3$ | $x_{15}^{[2]}h_4$ | $x_{15}^{[2]}h_5$ |
| $(x_{47}^{[1]}+x_{13}^{[2]})h_1$ | $(x_{49}^{[1]}+x_{12}^{[2]})h_2$ | $(x_{41}^{[1]}+x_{12}^{[2]})h_3$ | $(x_{43}^{[1]}+x_{11}^{[2]})h_4$ | $(x_{45}^{[1]}+x_{11}^{[2]})h_5$ |
| $(x_{48}^{[1]}+x_{15}^{[2]})h_1$ | $(x_{50}^{[1]}+x_{15}^{[2]})h_2$ | $(x_{42}^{[1]}+x_{14}^{[2]})h_3$ | $(x_{44}^{[1]}+x_{14}^{[2]})h_4$ | $(x_{46}^{[1]}+x_{13}^{[2]})h_5$ |

TABLE I
THE QUERY TABLE FOR THE RETRIEVAL OF $W_1$.

$$= KL_1 + \frac{1}{\binom{N}{K}} \sum_{\Omega:|\Omega|=K} H(A_\Omega^{[2]}|Q_\Omega^{[2]}, W_1) \quad (31)$$

$$\geq KL_1 + \frac{1}{\binom{N}{K}} \sum_{\Omega:|\Omega|=K} H(A_\Omega^{[2]}|Q_{1:N}^{[2]}, W_1) \quad (32)$$

$$\geq KL_1 + \frac{K}{N} H(A_{1:N}^{[2]}|Q_{1:N}^{[2]}, W_1) \quad (33)$$

$$= KL_1 + \frac{K}{N} I(W_{2:M}; A_{1:N}^{[2]}, Q_{1:N}^{[2]}|W_1) \quad (34)$$

$$= KL_1 + \frac{K}{N} I(W_{2:M}; W_2, A_{1:N}^{[2]}, Q_{1:N}^{[2]}|W_1) \quad (35)$$

$$= KL_1 + \frac{K}{N}(KL_2 + I(W_{3:M}; A_{1:N}^{[2]}, Q_{1:N}^{[2]}|W_{1:2})) \quad (36)$$

$$= KL_1 + \frac{K}{N} KL_2 + \frac{K}{N} I(W_{3:M}; A_{1:N}^{[2]}, Q_{1:N}^{[2]}|W_{1:2}), \quad (37)$$

where (29), (30), and (33) follow from (22), the privacy constraint in (3), and Han's inequality $\frac{1}{\binom{N}{K}} \sum_{\Omega:|\Omega|=K} H(A_\Omega^{[m]}|W_\Delta, Q_{1:N}^{[m]}) \geq \frac{K}{N} H(A_{1:N}^{[m]}|W_\Delta, Q_{1:N}^{[m]})$, respectively. The recursive application of (26)-(37) on the last two terms in (37) gives

$$R = \frac{K\mathbb{E}[L]}{\mathbb{E}[D]} \leq \frac{K\mathbb{E}[L]}{K\left(L_1 + \frac{K}{N}L_2 + \cdots + \frac{K^{M-1}}{N^{M-1}}L_M\right)} \quad (38)$$

leading to the capacity expression in Theorem 1.

## VI. DISCUSSION

In this paper, we presented a capacity-achieving scheme for semantic PIR from MDS-coded databases. An alternative description of the scheme in Section IV is as follows. Each MDS coded database contains $L_i$ coded bits representing $W_i$ with $L_1 \geq L_2 \geq \cdots \geq L_M$. Initially, consider the first $L_M$ coded bits of all $M$ messages and perform equal-length PIR as described in [44]. Then, consider the next $L_{M-1} - L_M$ coded bits of all messages except $W_M$ and perform equal-length PIR with the remaining $M-1$ messages. Continue this process until the last $L_1 - L_2$ coded bits of $W_1$ and perform PIR with a single message. This process yields the same outcome described in the scheme in Section IV.

Nevertheless, the description of the scheme in Section IV provides a *systematic* method of calculating the subpacketization based on the message sizes $\{L_i\}_{i=1}^M$, such that the scheme described on a single subpacket is applied repeatedly throughout the retrieval process in the same way. This is in contrast to the alternative description, where the subpacketization changes from one block to another (blocks of sizes $L_M$ and $L_i - L_{i+1}$ bits for $i = 1, \ldots, M-1$). In summary, the description of the scheme in Section IV uses a fixed subpacketization, while that of the alternative scheme uses irregular subpacketization.

## REFERENCES


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

[2] H. Sun and S. A. Jafar. The capacity of private information retrieval. *IEEE Trans. on Info. Theory*, 63(7):4075–4088, July 2017.

[3] H. Sun and S. A. Jafar. The capacity of robust private information retrieval with colluding databases. *IEEE Trans. on Info. Theory*, 64(4):2361–2370, April 2018.

[4] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb. Private information retrieval schemes for coded data with arbitrary collusion patterns. In *IEEE ISIT*, June 2017.

[5] R. Bitar and S. El Rouayheb. Staircase-PIR: Universally robust private information retrieval. In *IEEE ITW*, pages 1–5, November 2018.

[6] H. Sun and S. A. Jafar. The capacity of symmetric private information retrieval. *IEEE Transactions on Information Theory*, 65(1):322–329, January 2019.

[7] Q. Wang, H. Sun, and M. Skoglund. Symmetric private information retrieval with mismatched coded messages and randomness. In *IEEE ISIT*, pages 365–369, July 2019.

[8] T. Guo, R. Zhou, and C. Tian. On the information leakage in private information retrieval systems. Available at arXiv: 1909.11605.

[9] K. Banawan and S. Ulukus. Multi-message private information retrieval: Capacity results and near-optimal schemes. *IEEE Trans. on Info. Theory*, 64(10):6842–6862, October 2018.

[10] K. Banawan and S. Ulukus. The capacity of private information retrieval from Byzantine and colluding databases. *IEEE Trans. on Info. Theory*, 65(2):1206–1219, February 2019.

[11] R. Tandon. The capacity of cache aided private information retrieval. In *Allerton Conference*, October 2017.

[12] Y.-P. Wei, K. Banawan, and S. Ulukus. Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching. *IEEE Trans. on Info. Theory*, 65(5):3215–3232, May 2019.

[13] Y.-P. Wei, K. Banawan, and S. Ulukus. Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits. *IEEE JSAC*, 36(6):1126–1139, June 2018.

[14] S. Kumar, A. G. i Amat, E. Rosnes, and L. Senigagliesi. Private information retrieval from a cellular network with caching at the edge. *IEEE Trans. on Communications*, 67(7):4900–4912, July 2019.

[15] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson. Private information retrieval with side information. *IEEE Trans. on Info. Theory*, 66(4):2032–2043, April 2020.

[16] Z. Chen, Z. Wang, and S. Jafar. The capacity of $T$-private information retrieval with private side information. Available at arXiv:1709.03022.

[17] Y.-P. Wei, K. Banawan, and S. Ulukus. The capacity of private information retrieval with partially known private side information. *IEEE Trans. on Info. Theory*, 65(12):8222–8231, December 2019.

[18] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali. Multi-message private information retrieval with private side information. In *IEEE ITW*, pages 1–5, November 2018.

[19] S. Li and M. Gastpar. Converse for multi-server single-message PIR with side information. Available at arXiv:1809.09861.

[20] H. Sun and S. A. Jafar. The capacity of private computation. *IEEE Trans. on Info. Theory*, 65(6):3880–3897, June 2019.

[21] M. Mirmohseni and M. A. Maddah-Ali. Private function retrieval. In *IWCIT*, pages 1–6, April 2018.

[22] Z. Chen, Z. Wang, and S. Jafar. The asymptotic capacity of private search. In *IEEE ISIT*, June 2018.

[23] M. A. Attia, D. Kumar, and R. Tandon. The capacity of private information retrieval from uncoded storage constrained databases. Available at arXiv:1805.04104v2.

[24] C. Tian, H. Sun, and J. Chen. Capacity-achieving private information retrieval codes with optimal message size and upload cost. *IEEE Trans. on Info. Theory*, 65(11):7613–7627, Nov 2019.

[25] Y.-P. Wei and S. Ulukus. The capacity of private information retrieval with private side information under storage constraints. *IEEE Trans. on Info. Theory*, 66(4):2023–2031, April 2020.

[26] K. Banawan, B. Arasli, and S. Ulukus. Improved storage for efficient private information retrieval. In *IEEE ITW*, August 2019.

[27] C. Tian. On the storage cost of private information retrieval. Available at arXiv:1910.11973.

[28] Y.-P. Wei, B. Arasli, K. Banawan, and S. Ulukus. The capacity of private information retrieval from decentralized uncoded caching databases. *Information*, 10, December 2019.

[29] K. Banawan, B. Arasli, Y. P. Wei, and S. Ulukus. The capacity of private information retrieval from heterogeneous uncoded caching databases. *IEEE Trans. on Info. Theory*, 66(6):3407–3416, 2020.

[30] N. Raviv and I. Tamo. Private information retrieval in graph based replication systems. In *IEEE ISIT*, June 2018.

[31] K. Banawan and S. Ulukus. Private information retrieval from non-replicated databases. In *IEEE ISIT*, pages 1272–1276, July 2019.

[32] K. Banawan and S. Ulukus. Private information retrieval through wiretap channel II: Privacy meets security. *IEEE Trans. on Info. Theory*, 66(7):4129–4149, 2020.

[33] H. Sun and S. A. Jafar. Optimal download cost of private information retrieval for arbitrary message length. *IEEE Trans. on Info. Forensics and Security*, 12(12):2920–2932, December 2017.

[34] Q. Wang, H. Sun, and M. Skoglund. The capacity of private information retrieval with eavesdroppers. *IEEE Trans. on Info. Theory*, 65(5):3198–3214, May 2019.

[35] H. Yang, W. Shin, and J. Lee. Private information retrieval for secure distributed storage systems. *IEEE Trans. on Info. Forensics and Security*, 13(12):2953–2964, December 2018.

[36] Z. Jia, H. Sun, and S. Jafar. Cross subspace alignment and the asymptotic capacity of $X$-secure $T$-private information retrieval. *IEEE Trans. on Info. Theory*, 65(9):5783–5798, September 2019.

[37] R. Zhou, C. Tian, H. Sun, and T. Liu. Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size. Available at arXiv: 1903.08229.

[38] K. Banawan and S. Ulukus. Asymmetry hurts: Private information retrieval under asymmetric-traffic constraints. *IEEE Trans. on Info. Theory*, 65(11):7628–7645, November 2019.

[39] K. Banawan and S. Ulukus. Noisy private information retrieval: On separability of channel coding and information retrieval. *IEEE Trans. on Info. Theory*, 65(12):8232–8249, December 2019.

[40] R. G. L. D'Oliveira and S. El Rouayheb. One-shot PIR: Refinement and lifting. *IEEE Trans. on Info. Theory*, 66(4):2443–2455, April 2020.

[41] R. Tajeddine, A. Wachter-Zeh, and C. Hollanti. Private information retrieval over random linear networks. Available at arXiv:1810.08941.

[42] Z. Wang, K. Banawan, and S. Ulukus. Private set intersection: A multi-message symmetric private information retrieval perspective. Available at arXiv: 1912.13501.

[43] Z. Wang, K. Banawan, and S. Ulukus. Multi-party private set intersection: An information-theoretic approach. Available at arXiv: 2008.07504.

[44] K. Banawan and S. Ulukus. The capacity of private information retrieval from coded databases. *IEEE Trans. on Info. Theory*, 64(3):1945–1956, March 2018.

[45] R. Freij-Hollanti, O. Gnilke, C. Hollanti, and D. Karpuk. Private information retrieval from coded databases with colluding servers. *SIAM Journal on Applied Algebra and Geometry*, 1(1):647–664, 2017.

[46] H. Sun and S. A. Jafar. Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al. *IEEE Trans. on Info. Theory*, 64(2):1000–1022, February 2018.

[47] Y. Zhang and G. Ge. A general private information retrieval scheme for MDS coded databases with colluding servers. *Designs, Codes and Cryptography*, 87(11), November 2019.

[48] Y. Zhang and G. Ge. Multi-file private information retrieval from MDS coded databases with colluding servers. Available at arXiv: 1705.03186.

[49] R. Tandon, M. Abdul-Wahid, F. Almoualem, and D. Kumar. PIR from storage constrained databases - coded caching meets PIR. *IEEE ICC*, May 2018.

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

[51] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb. Private information retrieval from mds coded data in distributed storage systems. *IEEE Tran. on Info.Theory*, 64(11):7081–7093, November 2018.

[52] R. Tajeddine and S. El Rouayheb. Robust private information retrieval on coded data. In *IEEE ISIT*, June 2017.

[53] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti. Private information retrieval from coded storage systems with colluding, Byzantine, and unresponsive servers. *IEEE Trans. on Info. Theory*, 65(6):3898–3906, June 2019.

[54] S. Vithana, K. Banawan, and S. Ulukus. Semantic private information retrieval. Available at arXiv: 2003.13667.