Semantic Private Information Retrieval: Effects of Heterogeneous Message Sizes and Popularities

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 (semantic PIR). In semantic PIR, a user privately retrieves a message out of K independent messages stored in N replicated and non-colluding databases. The messages come with *different semantics*, i.e., the messages are allowed to have *non-uniform a priori probabilities* denoted by $(p_i > 0,\ i \in [K])$ and *arbitrary message sizes* $(L_i,\ i \in [K])$. We derive the semantic PIR capacity for general K, N. We present two achievable semantic PIR schemes: The first one is a deterministic scheme with non-uniform subpacketization. The second scheme is probabilistic and is based on choosing one query set out of multiple options at random to retrieve the required message without the need for exponential subpacketization. We derive conditions for the semantic PIR capacity to exceed the classical PIR capacity with equal priors and sizes. Our results show that the semantic PIR capacity can be larger than the classical PIR capacity when longer messages have higher popularities. However, when messages are of equal-length, the non-uniform priors cannot be exploited to improve the retrieval rate.**

I. Introduction

Private information retrieval (PIR) describes an elemental privacy setting. In the classical PIR problem, introduced in the seminal paper [1], a user needs to retrieve a message (file), from multiple replicated databases, without revealing any information about the identity of the desired message. In [2], the notion of PIR capacity is introduced as the maximum ratio of the desired message size to the total download size. Reference [2] characterizes the classical PIR capacity using a greedy algorithm which is based on message and database symmetry. Using this performance metric, further practical variants of the problem have been investigated [3]–[36].

In all these works, two assumptions are made: All messages have the same size, L, and all messages are requested uniformly by the users. These assumptions are highly idealistic from a practical point of view. Take a streaming application for instance. The storage database has a catalog of different movies and TV shows. These media files cannot be assumed to have the same level of popularity. In addition, the media files cannot be assumed to be equal in size. Consequently, each message stored in the databases exhibits different *semantics*, in the sense that each message has a different size and a different prior probability of retrieval. With this backdrop, in this paper, we investigate how a PIR scheme should be implemented over databases holding messages with different semantics.

This work was supported by NSF Grants CCF 17-13977 and ECCS 18-07348.

In this work, we introduce the semantic PIR problem. We extend redthe notion of PIR capacity to deal with different message semantics. We define the retrieval rate to be the ratio of the *expected* message size to the *expected* download cost. Due to the privacy constraint, the download cost is the same for all messages. Hence, the retrieval rate is equal to the weighted average of all individual message retrieval rates. We investigate the semantic PIR capacity as a function of the system parameters: number of databases N, number of messages K, message priors p_i, and message lengths L_i. We ask whether there is a PIR capacity gain over classical PIR capacity from exploiting the message semantics.

In this paper, we characterize the exact semantic PIR capacity. To that end, we present two achievable schemes; the first scheme is deterministic, in the sense that the *query structure* is fixed, and the second scheme is stochastic, in the sense that the user picks a query structure *randomly* from a list of possible structures. For the deterministic scheme, we present a systematic method to determine the subpacketization level for each message. This scheme uses non-uniform subpacketization where the block size considered in each download differs from one message to another. The query structure of the deterministic scheme resembles the query structure of [2]. The second achievable scheme is comprised of several query options that the user may use with equal probability to retrieve any message. In this scheme, the messages are divided into several blocks depending on the number of databases. This is similar to the scheme presented in [37] with an extension to more than two databases (see also [38]). We provide a matching converse that takes into account the heterogeneity of message sizes, resulting in settling the semantic PIR capacity.

The semantic PIR capacity implies that for certain message sizes and priors, the classical PIR capacity may be exceeded by exploiting the semantics of the messages even if the zero-padding needed in classical PIR to equalize the message sizes is ignored. Concretely, our results imply: 1) When message lengths are the same, semantic PIR capacity is equal to the classical PIR capacity irrespective of the message priors, i.e., priors cannot be exploited to increase the PIR capacity. 2) For certain cases, such as when the prior probability distribution favors longer files (i.e., longer files are more popular), the semantic PIR capacity exceeds the classical PIR capacity $C_{PIR} = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}}\right)^{-1}$. 3) For all priors and lengths, our scheme achieves a larger PIR rate than the PIR

rate the classical approach would achieve by simply zero-padding the messages to bring them to the same length. Due to space limitations here, proof details, extra examples, remarks, and figures can be found in the longer version in [39].

II. Problem Formulation

We consider a setting, where N replicated and non-colluding databases store K independent messages, $W_1, \ldots, W_K$. The messages exhibit different semantics, i.e., the messages have different sizes and different a priori probabilities of retrieval. The a priori probability of W_i is denoted by p_i, such that $p_i > 0$ for $i = 1, \ldots, K$. The a priori probability distribution is globally known at all entities. All message symbols are picked from a finite field $\mathbb{F}_s$. The message size of the ith message is denoted by L_i. Without loss of generality, we assume that the messages are ordered with respect to their sizes, such that $L_1 \geq L_2 \geq \cdots \geq L_K$. The message sizes can be expressed in s-ary symbols as,

$$H(W_i) = L_i, \quad i = 1, \ldots, K \tag{1}$$

$$H(W_1, \ldots, W_K) = \sum_{i=1}^{K} H(W_i) = \sum_{i=1}^{K} L_i \tag{2}$$

In semantic PIR, a user needs to retrieve a message W_i without revealing the index i to any individual database. The user sends a query $Q_n^{[i]}$ to the nth database to retrieve W_i for $n = 1, \ldots, N$. Initially, the user does not have any information about the message contents. Hence, the queries are independent of the messages,

$$I(W_1, \ldots, W_K; Q_1^{[i]}, \ldots, Q_N^{[i]}) = 0, \quad i = 1, \ldots, K \tag{3}$$

The nth database prepares an answer string $A_n^{[i]}$. which is a deterministic function of the messages $W_1, \ldots, W_K$ and the received query $Q_n^{[i]}$. I.e., for $i = 1, \ldots, K$ and $n = 1, \ldots, N$,

$$H(A_n^{[i]} | Q_n^{[i]}, W_1, \ldots, W_K) = 0 \tag{4}$$

A feasible PIR scheme satisfies the following conditions:

Correctness: The user can correctly reconstruct the desired message based on the received answer strings. Therefore,

$$H(W_i | A_1^{[i]}, \ldots, A_N^{[i]}, Q_1^{[i]}, \ldots, Q_N^{[i]}) = 0, \quad i = 1, \ldots, K \tag{5}$$

Privacy: The queries should not leak any information about i. Formally, the a posteriori probability of the message index i given a query $Q_n^{[i]}$ is equal to the a priori probability of the message index i. That is, denoting the random variable representing the desired message index by θ,

$$P(\theta = i | Q_n^{[i]}) = P(\theta = i),\ i = 1, \ldots, K,\ n = 1, \ldots, N \tag{6}$$

Constraint (6) implies that for $n = 1, \ldots, N,\ i, j = 1, \ldots, K$,

$$(Q_n^{[i]}, A_n^{[i]}, W_1, \ldots, W_K) \sim (Q_n^{[j]}, A_n^{[j]}, W_1, \ldots, W_K) \tag{7}$$

where $\sim$ denotes statistical equivalence.

Due to the heterogeneity of message sizes and a priori probabilities, we define the performance metric, the expected retrieval rate R, as the ratio of the expected retrieved message size to the expected download size, i.e.,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} \tag{8}$$

where $\mathbb{E}[L]$ is the expected number of useful bits downloaded 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 semantic PIR capacity is defined as the supremum of the expected retrieval rates over all achievable PIR schemes, i.e., $C = \sup R$.

III. Main Results and Discussions

Theorem 1 *The semantic PIR capacity with N databases, K messages, message sizes L_i (arranged in decreasing order as $L_1 \geq L_2 \geq \cdots \geq L_K$), and prior probabilities p_i, is*

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N} \frac{L_2}{\mathbb{E}[L]} + \cdots + \frac{1}{N^{K-1}} \frac{L_K}{\mathbb{E}[L]} \right)^{-1} \tag{9}$$

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

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

Next, we give a necessary and sufficient condition for the cases the semantic capacity exceeds the classical PIR capacity. The proof can be found in [39].

Corollary 1 *The semantic PIR capacity is strictly larger than the classical PIR capacity if and only if,*

$$\sum_{i=1}^{K} \frac{1}{N^{i-1}} (L_i - \mathbb{E}[L]) < 0 \tag{10}$$

Remark 1 *More explicit conditions can be derived for specific cases. For example, consider the case $K = 2$, $N = 2$, and assume that $L_1 > L_2$ (strictly larger). Then,* (10) *simplifies to,*

$$(L_1 - (p_1 L_1 + p_2 L_2)) + \frac{1}{2}(L_2 - (p_1 L_1 + p_2 L_2)) < 0 \tag{11}$$

$$p_2(L_1 - L_2) + \frac{1}{2} p_1 (L_2 - L_1) < 0 \tag{12}$$

$$p_2 - \frac{1}{2} p_1 < 0 \tag{13}$$

$$p_1 > \frac{2}{3} \tag{14}$$

where (13) *follows from $L_1 > L_2$. This means that for $N = 2$ and $K = 2$, the capacity of semantic PIR is greater than the capacity of classical PIR when the a priori probability of the longer message is greater than $\frac{2}{3}$ irrespective of the values of L_1 and L_2.*

As a further explicit example, if $p_1 = 4p_2$ and $L_1 = 4L_2$, then the semantic PIR capacity is $C = \frac{34}{45}$ while the classical PIR capacity is $C_{PIR} = \frac{2}{3} < C = \frac{34}{45}$.

Remark 2 *We further expand on Remark 1 above by noting the following fact. The classical PIR capacity is a formula that depends only on the number of databases N and the*

number of messages K, and is not necessarily achievable by the classical PIR scheme for any given message priors and lengths. To see this, we note that the classical PIR scheme requires equal message sizes. In the example in Remark 1 where $p_1 = 4p_2$ and $L_1 = 4L_2$, if we zero-pad the shorter message to make the message lengths the same, we achieve $R_{ach} = p_1 \frac{L_1}{D} + p_2 \frac{L_2}{D} = \frac{17}{30}$ by noting $D = \frac{3}{2}L_1$ as the length of the longer message is the common message length now, and the classical PIR capacity for this case is $\frac{2}{3}$. Thus, we observe $R_{ach} = \frac{17}{30} < C_{PIR} = \frac{2}{3} < C = \frac{34}{45}$ for this case.

Furthermore, the next corollary states that the proposed achievable scheme always outperforms zero-padding shorter messages and applying the classical PIR scheme for so-constructed equal-length messages. The proof is in [39].

Corollary 2 *Semantic PIR capacity outperforms classical PIR rate with zero-padding.*

Remark 3 *If all messages have equal lengths, irrespective of the prior probabilities, the capacity of semantic PIR becomes equal to that of classical PIR. Note, in this case, $L_i = \mathbb{E}[L]$ and the capacity expression in* (9) *reduces to the classical PIR capacity expression in [2]. Thus, in order to exploit variability in priors to achieve a PIR capacity higher than the classical PIR capacity, we need variability in message lengths.*

IV. Achievability Proof

In this section, we present two capacity-achieving PIR schemes.

A. Achievable Semantic PIR Scheme 1

The scheme is based on the iterative structure of the achievable scheme in [2]. In this scheme, the user downloads k-sums from the messages for $k = 1, \dots, K$. The novel component in our scheme is the calculation of the number of stages needed to be downloaded from each message based on the message sizes.

This achievable scheme is parameterized by $(K, N, \{L_i\}_{i=1}^K)$. Based on these parameters, the user prepares queries to retrieve the desired message privately. The basic structure of our achievable scheme is as follows.

1) **Message indexing**: Order the messages in the descending order of message sizes. That is, index 1 is assigned to the longest message and index K is assigned to the shortest message ($L_1 \geq L_2 \geq \cdots \geq L_K$). Calculate retrieval parameters $\upsilon_1, \upsilon_2, \dots, \upsilon_K$ corresponding to each message such that $\upsilon_1 \geq \upsilon_2 \geq \cdots \geq \upsilon_K$. The retrieval parameters denote the number of stages that needs to be downloaded from each message. The explicit expressions for the parameters are:

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \vdots \\ \upsilon_K \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} \frac{1}{N} & -\frac{N-1}{N^2} & \dots & -\frac{N-1}{N^K} \\ 0 & \frac{1}{N^2} & \dots & -\frac{N-1}{N^K} \\ 0 & 0 & \dots & -\frac{N-1}{N^K} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1}{N^K} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ \vdots \\ L_K \end{bmatrix} \quad (15)$$

for some integer α. The proof of this choice can be found in [39]. For the rest of this section, assume that the user wishes to download W_j.

2) **Index preparation:** The user permutes the indices of all messages independently, uniformly, and privately from the databases.
3) **Singletons:** Download υ_k different bits from message W_k from the nth database, where $n = 1, \dots, N$ and $k = 1, \dots, K$.
4) **Sums of two elements (2-sums):** 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, download $(N-1)\min\{\upsilon_i, \upsilon_j\}$ bit-wise sums of W_i and W_j each from the N databases for all $i \neq j$. Each sum comprises of an already downloaded W_i bit from another database and a new bit of W_j. For the second block, for all possible message pairs (W_{i_1}, W_{i_2}) for $i_1 \neq i_2 \neq j$, download $(N-1)\min\{\upsilon_{i_1}, \upsilon_{i_2}\}$ number of bit-wise sums of W_{i_1} and W_{i_2} each from the N databases. Each sum comprises of fresh bits from W_{i_1} and W_{i_2}.
5) **Repeat step 4** for all k-sums where $k = 3, 4, \dots, K$. For each k-sum, download k bit-wise sum from k messages. If one of these messages is the desired message, the remaining $(k-1)$-sum is derived from the previous $(k-1)$th round from a different database. Otherwise, download $(N-1)^{k-1}\min\{\upsilon_{i_1}, \dots, \upsilon_{i_k}\}$ sums from new bits of the undesired messages.

The proof of the retrieval rate, and the privacy can be found in [39].

B. Example 1: $N = 2, K = 2$, $L_1 = 1024$ *bits*, $L_2 = 256$ *bits*

First, the message indices are independently and uniformly permuted. The first and the second messages after permutations are denoted by bits a_i and b_i, respectively.

- Message indexing and calculation of υ_i: Messages are indexed such that the first message is the longer one, and the second message is the shorter one. Below, we will give query tables for downloading W_1 and W_2. We calculate υ_1 and υ_2 as,

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \quad (16)$$

  where $\alpha = \text{gcd}\{\frac{L_1}{2} - \frac{L_2}{4}, \frac{L_2}{4}\}$. Hence,

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} 448 \\ 64 \end{bmatrix} \quad (17)$$

  Hence, $\alpha = \text{gcd}\{448, 64\} = 64$. Therefore, $\upsilon_1 = 7$ and $\upsilon_2 = 1$. The subpacketization levels of W_1 and W_2 are $U_1 = \frac{1024}{64} = 16$ and $U_2 = \frac{256}{64} = 4$, respectively.
- Singletons: Download $\upsilon_1 = 7$ bits of W_1 and $\upsilon_2 = 1$ bit of W_2 from each of the two databases.
- Sums of twos: Download $(N-1)\upsilon_2 = 1$ sum of W_1 and W_2 bits each from both databases. Note that if W_1 is the

desired message, the singletons of W_2 are used as a side information with new W_1 bits in the sum and vice versa.

Tables I and II show the queries sent to the databases to retrieve W_1 and W_2, respectively.

Database 1	Database 2
$a_1, \ldots, a_7$ b_1	$a_8, \ldots, a_{14}$ b_2
$a_{15} + b_2$	$a_{16} + b_1$

TABLE I
THE QUERY TABLE FOR THE RETRIEVAL OF W_1.

Database 1	Database 2
$a_1, \ldots, a_7$ b_1	$a_8, \ldots, a_{14}$ b_2
$a_8 + b_3$	$a_1 + b_4$

TABLE II
THE QUERY TABLE FOR THE RETRIEVAL OF W_2.

The rate achieved by this scheme when downloading W_1 is $R_1 = \frac{16}{18} = \frac{8}{9}$, and the rate achieved by this scheme when downloading W_2 is $R_2 = \frac{4}{18} = \frac{2}{9}$. Therefore, the average rate R achieved by the scheme is,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} = \frac{p_1 L_1 + p_2 L_2}{p_1 D + p_2 D} = p_1 R_1 + p_2 R_2 = \frac{8}{9} p_1 + \frac{2}{9} p_2 \quad (18)$$

This matches the capacity expression in Theorem 1 as,

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N} \frac{L_2}{\mathbb{E}[L]} \right)^{-1} \quad (19)$$

$$= (1024 p_1 + 256 p_2) \left(1024 + \frac{256}{2} \right)^{-1} = \frac{8}{9} p_1 + \frac{2}{9} p_2 \quad (20)$$

The classic PIR capacity for this case with equal priors is,

$$C = \left(1 + \frac{1}{N} \right)^{-1} = \left(1 + \frac{1}{2} \right)^{-1} = \frac{2}{3} \quad (21)$$

The semantic PIR capacity in (20) exceeds the classical PIR capacity in (21) when

$$\frac{8}{9} p_1 + \frac{2}{9} p_2 > \frac{2}{3} \quad (22)$$

which is when $p_1 > \frac{2}{3}$. Consequently, when $p_1 > \frac{2}{3}$, there is a strict gain from exploiting message semantics for PIR.

Remark 4 *By accounting for the zero-padding needed for the shorter message to be of length 1024 bits, the actual retrieval rate is not $\frac{2}{3}$ as the actual message size of W_2 is much less. Specifically, the total download for this scheme is $D = \frac{L}{R} = \frac{1024}{2/3} = 1536$. The actual retrieval rate for the classical PIR problem is,*

$$R_{ach} = \frac{1/2 \times 1024 + 1/2 \times 256}{1536} = \frac{5}{12} < \frac{5}{9} < \frac{6}{9} \quad (23)$$

Thus, even though the semantic PIR capacity $\frac{5}{9}$ is less than the classical PIR capacity $\frac{6}{9}$, the semantic PIR capacity (which is achievable) is larger than the classical PIR rate with zero-padding $\frac{5}{12}$ as asserted by Corollary 2.

C. Achievable PIR Scheme 2

The scheme is stochastic in the sense that the user has a list of different possible query structures and the user picks one of these structures randomly. This is unlike the previous scheme where the structure is deterministic and the randomness comes from the random permutations of indices.

This scheme is developed for arbitrary number of databases and arbitrary message lengths that are multiples of $N-1$; the deterministic scheme in Sections IV-A assumed message lengths that are multiples of N^K. The scheme can be viewed as an extension of the achievable scheme in [37] to work with arbitrary number of databases and heterogeneous message sizes. Our scheme shares similarities with [38]. However, our scheme differs in that it introduces database symmetry to the scheme. The basic structure of the achievable scheme is as follows.

1) **Message indexing:** Index all messages such that $L_1 \geq L_2 \geq \cdots \geq L_K$. Divide all messages into $N-1$ blocks. Let W_i^m be the mth block of W_i.
   For the rest of this section, assume that the user requires to download W_j.
2) **Single blocks:** Use $N-1$ out of the N databases to download each block of W_j and download nothing from the remaining database. Consider all N cyclic shifts of the blocks around the databases to obtain N options for different queries that can be used to download W_j. These N queries require the user to download L_j bits in total, resulting in no side information.
3) **Sums of two blocks/single blocks:** Choose one database to download W_i^1 where $i \neq j$ and download $W_j^m + W_i^1$ for $m = 1, \ldots, N-1$ from the remaining $N-1$ databases. Create N query options in total by considering all N cyclic shifts of the blocks, around the databases. Repeat the procedure for W_i^ℓ where $\ell = 2, \ldots, N-1$. There are a total of $N(N-1)\binom{K-1}{1}$ query options of this type.
4) **Sums of three blocks/sums of two blocks:** Choose one database to download $W_{i_1}^1 + W_{i_2}^1$ where $i_1, i_2 \neq j$ and download $W_j^m + W_{i_1}^1 + W_{i_2}^1$ for $m = 1, \ldots, N-1$ from the remaining $N-1$ databases. Create N query options in total by considering all N cyclic shifts of the blocks around the databases. Repeat the procedure for $W_{i_1}^{\ell_1} + W_{i_2}^{\ell_2}$ where $\ell_1, \ell_2 \in \{2, \ldots, N-1\}$. There are $N(N-1)^2\binom{K-1}{2}$ query options of this type.
5) **Repeat step 4** up to sums of K blocks/sums of $K-1$ blocks.

Once the user chooses a query to be sent to the N databases, out of the N^K options, each database might have to compute sums of messages with different lengths. All messages except the longest in the sum are zero-padded to the left to have equal-length blocks. Then, bit-wise sums are calculated.

Once the answers are received from the databases, the user might need to subtract messages of different lengths to recover

the required message. In this case, according to the design of the scheme, the subtrahend will always be shorter than or equal to the length of the minuend. Hence, the subtraction operation in this context will not be any different than the usual operation.

The retrieval rate, privacy proofs can be found in [39].

D. Example 2: $N = 4, K = 2$, $L_1 = 3000$ *bits*, $L_2 = 1800$ *bits*

Table III shows the sets of queries that the user can utilize with probability $\frac{1}{16}$ in order to retrieve messages W_1. We note that for a given database, the set of possible queries that the user utilizes is the same regardless of the desired message. Whenever a set of queries for the four databases is chosen with probability $\frac{1}{16}$, the required message is retrieved by subtracting the smaller sum from the larger sums, guaranteeing correctness.

In the first block of Table III, W_1 is divided into 3 parts and each part is retrieved from different 3 databases at each query option. In the second block, W_2^1 is used as side information, which is requested from one database, and the three parts of W_1 are retrieved from the other three databases in terms of $W_1^i + W_2^1$ for $i = 1, 2, 3$. The third and fourth blocks are the same as block 2, with W_2^1 replaced by W_2^2 and W_2^3, respectively.

Probability	Database 1	Database 2	Database 3	Database 4
$\frac{1}{16}$	W_1^1	W_1^2	W_1^3	ϕ
$\frac{1}{16}$	W_1^2	W_1^3	ϕ	W_1^1
$\frac{1}{16}$	W_1^3	ϕ	W_1^1	W_1^2
$\frac{1}{16}$	ϕ	W_1^1	W_1^2	W_1^3
$\frac{1}{16}$	$W_1^1 + W_2^1$	$W_1^2 + W_2^1$	$W_1^3 + W_2^1$	W_2^1
$\frac{1}{16}$	$W_1^2 + W_2^1$	$W_1^3 + W_2^1$	W_2^1	$W_1^1 + W_2^1$
$\frac{1}{16}$	$W_1^3 + W_2^1$	W_2^1	$W_1^1 + W_2^1$	$W_1^2 + W_2^1$
$\frac{1}{16}$	W_2^1	$W_1^1 + W_2^1$	$W_1^2 + W_2^1$	$W_1^3 + W_2^1$
$\frac{1}{16}$	$W_1^1 + W_2^2$	$W_1^2 + W_2^2$	$W_1^3 + W_2^2$	W_2^2
$\frac{1}{16}$	$W_1^2 + W_2^2$	$W_1^3 + W_2^2$	W_2^2	$W_1^1 + W_2^2$
$\frac{1}{16}$	$W_1^3 + W_2^2$	W_2^2	$W_1^1 + W_2^2$	$W_1^2 + W_2^2$
$\frac{1}{16}$	W_2^2	$W_1^1 + W_2^2$	$W_1^2 + W_2^2$	$W_1^3 + W_2^2$
$\frac{1}{16}$	$W_1^1 + W_2^3$	$W_1^2 + W_2^3$	$W_1^3 + W_2^3$	W_2^3
$\frac{1}{16}$	$W_1^2 + W_2^3$	$W_1^3 + W_2^3$	W_2^3	$W_1^1 + W_2^3$
$\frac{1}{16}$	$W_1^3 + W_2^3$	W_2^3	$W_1^1 + W_2^3$	$W_1^2 + W_2^3$
$\frac{1}{16}$	W_2^3	$W_1^1 + W_2^3$	$W_1^2 + W_2^3$	$W_1^3 + W_2^3$

TABLE III
THE QUERY TABLE FOR THE RETRIEVAL OF W_1.

The rate achieved by this scheme when retrieving W_1 is,

$$R_1 = \frac{L_1}{\frac{1}{16}\left(4L_1 + 12(\frac{L_1}{3} \times 3 + \frac{L_2}{3})\right)} = \frac{20}{23} \tag{24}$$

The rate achieved by this scheme when retrieving W_2 is,

$$R_2 = \frac{L_2}{\frac{1}{16}\left(4L_2 + 12 \times 4 \times \frac{L_1}{3}\right)} = \frac{12}{23} \tag{25}$$

The overall message retrieval rate for this example is,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} = p_1 R_1 + p_2 R_2 = \frac{20}{23}p_1 + \frac{12}{23}p_2 \tag{26}$$

This matches the semantic PIR capacity expression in Theorem 1,

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N}\frac{L_2}{\mathbb{E}[L]}\right)^{-1} = \frac{20}{23}p_1 + \frac{12}{23}p_2 \tag{27}$$

The classical PIR capacity for this case with equal priors is,

$$C = \left(1 + \frac{1}{N}\right)^{-1} = \left(1 + \frac{1}{4}\right)^{-1} = \frac{4}{5} \tag{28}$$

The semantic PIR capacity in (27) exceeds the classical PIR capacity in (28) when

$$\frac{20}{23}p_1 + \frac{12}{23}p_2 > \frac{4}{5} \tag{29}$$

which is when $p_1 > \frac{4}{5}$. Consequently, when $p_1 > \frac{4}{5}$, there is a strict gain from exploiting message semantics for PIR.

V. CONVERSE PROOF

In this section, we present the converse proof for Theorem 1. We note that our converse proof inherits most of its core ideas from [2]. We extend the proof of [2] to handle non-equal message sizes. First, due to the privacy constraint, we have,

$$A_n^{[i]} \sim A_n^{[j]}, \quad n \in [N], \quad i, j \in [K] \tag{30}$$

Hence, $H(A_n^{[i]}) = H(A_n^{[j]})$ for all $i, j \in [K]$ for all $n \in [N]$.

Choose an arbitrary permutation $\{i_1, \ldots, i_K\}$ for the order of messages. The expected download cost can be written as[1],

$$\begin{aligned}
\mathbb{E}[D] &= \sum_{i=1}^{K} q_i (H(A_1^{[i]}) + \cdots + H(A_N^{[i]})) && (31)\\
&= H(A_1^{[i_1]}) + \cdots + H(A_N^{[i_1]}) && (32)\\
&\geq H(A_{1:N}^{[i_1]}) && (33)\\
&\geq H(A_{1:N}^{[i_1]} | Q_{1:N}^{[i_1]}) && (34)\\
&= I(W_{i_1:i_K}; A_{1:N}^{[i_1]} | Q_{1:N}^{[i_1]}) && (35)\\
&= I(W_{i_1}; A_{1:N}^{[i_1]} | Q_{1:N}^{[i_1]}) + I(W_{i_2:i_K}; A_{1:N}^{[i_1]} | Q_{1:N}^{[i_1]}, W_{i_1}) && (36)\\
&= L_{i_1} + I(W_{i_2:i_K}; A_{1:N}^{[i_1]}, Q_{1:N}^{[i_1]} | W_{i_1}) && (37)\\
&\geq L_{i_1} + I(W_{i_2:i_K}; A_1^{[i_1]}, Q_1^{[i_1]} | W_{i_1}) && (38)\\
&= L_{i_1} + I(W_{i_2:i_K}; A_1^{[i_1]} | Q_1^{[i_1]}, W_{i_1}) && (39)\\
&= L_{i_1} + H(A_1^{[i_1]} | Q_1^{[i_1]}, W_{i_1}) && (40)\\
&= L_{i_1} + H(A_1^{[i_2]} | Q_1^{[i_2]}, W_{i_1}) && (41)
\end{aligned}$$

Since (38) holds true for any pair of query, answer string, the last inequality (41) is also true for any $(Q_n^{[i_2]}, A_n^{[i_2]})$, hence,

$$\mathbb{E}[D] \geq L_{i_1} + H(A_n^{[i_2]} | Q_n^{[i_2]}, W_{i_1}), \quad n = 1, \ldots, N \tag{42}$$

By summing all N inequalities corresponding to (42),

$$N\mathbb{E}[D] \geq N L_{i_1} + \sum_{n=1}^{N} H(A_n^{[i_2]} | Q_n^{[i_2]}, W_{i_1}) \tag{43}$$

[1]For compact notation, we denote a collection of random variables as $X_{1:N} = \{X_1, X_2, \cdots, X_N\}$.

$$\geq NL_{i_1} + H(A_{1:N}^{[i_2]}|Q_{1:N}^{[i_2]}, W_{i_1}) \tag{44}$$

The term $H(A_{1:N}^{[i_2]}|Q_{1:N}^{[i_2]}, W_{i_1})$ can be bounded by repeating the previous arguments for W_{i_2} (with conditioning on W_{i_1}) to have,

$$N\mathbb{E}[D] \geq NL_{i_1} + L_{i_2} + H(A_1^{[i_3]}|Q_1^{[i_3]}, W_{i_1}, W_{i_2}) \tag{45}$$

By summing the corresponding inequalities and continuing with the same procedure for $W_{i_3}, \ldots, W_{i_K}$, we have,

$$\begin{aligned} N^{K-1}\mathbb{E}[D] \geq\ & N^{K-1}L_{i_1} + N^{K-2}L_{i_2} + \cdots + NL_{i_{K-1}} \\ & + I(W_{i_K}; A_{1:N}^{[i_K]}|Q_{1:N}^{[i_K]}, W_{i_1:i_K}) \end{aligned} \tag{46}$$

and therefore, we have,

$$\mathbb{E}[D] \geq L_{i_1} + \frac{1}{N}L_{i_2} + \cdots + \frac{1}{N^{K-1}}L_{i_K} \tag{47}$$

This gives,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} \leq \left(\frac{L_{i_1}}{\mathbb{E}[L]} + \frac{1}{N}\frac{L_{i_2}}{\mathbb{E}[L]} + \cdots + \frac{1}{N^{K-1}}\frac{L_{i_K}}{\mathbb{E}[L]}\right)^{-1} \tag{48}$$

The last upper bound holds for any permutations of the messages. Since the messages are ordered such that $L_1 \geq L_2 \geq \cdots \geq L_K$, the tightest upper bound is attained at $\{i_1, \ldots, i_K\} = \{1, \ldots, K\}$. Thus,

$$R \leq \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N}\frac{L_2}{\mathbb{E}[L]} + \cdots + \frac{1}{N^{K-1}}\frac{L_K}{\mathbb{E}[L]}\right)^{-1} \tag{49}$$

completing the converse proof.

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] X. Yao, N. Liu, and W. Kang. The capacity of private information retrieval under arbitrary collusion patterns. Available at arXiv:2001.03843.

[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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[25] 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.

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

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

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

[29] 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.

[30] Q. Wang and M. Skoglund. On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers. *IEEE Trans. on Info. Theory*, 65(5):3183–3197, May 2019.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] I. Samy, M. A. Attia, R. Tandon, and L. Lazos. Latent-variable private information retrieval. Available at arXiv: 2001.05998.

[37] I. Samy, R. Tandon, and L. Lazos. On the capacity of leaky private information retrieval. In *IEEE ISIT*, pages 1262–1266, July 2019.

[38] 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.

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