Semantic Private Information Retrieval

Sajani Vithana, *Graduate Student Member, IEEE*, Karim Banawan, *Member, IEEE*, and Sennur Ulukus, *Fellow, IEEE*

***Abstract*— We investigate the problem of semantic private information retrieval (semantic PIR). In semantic PIR, a user retrieves a message out of K independent messages stored in N replicated and non-colluding databases without revealing the identity of the desired message to any individual database. The messages come with *different semantics*, i.e., the messages are allowed to have *non-uniform a priori probabilities* denoted by $(p_i > 0,\ i \in [K])$, which are a proxy for their respective popularity of retrieval, and *arbitrary message sizes* $(L_i, i \in [K])$. This is a generalization of the classical private information retrieval (PIR) problem, where messages are assumed to have equal message sizes. We derive the semantic PIR capacity for general K, N. The results show that the semantic PIR capacity depends on the number of databases N, the number of messages K, the a priori probability distribution of messages p_i, and the message sizes L_i. We present two achievable semantic PIR schemes: The first one is a deterministic scheme which is based on message asymmetry. This scheme employs 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 necessary and sufficient 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 equal-length, the non-uniform priors cannot be exploited to improve the retrieval rate over the classical PIR capacity. We provide two extensions of the semantic PIR problem, namely, the semantic PIR from MDS-coded databases and the semantic PIR from colluding databases. For both extensions, we derive the exact PIR capacity in addition to providing a corresponding optimal scheme.**

***Index Terms*— Private information retrieval (PIR), arbitrary message lengths, arbitrary message popularities, PIR capacity.**

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), out of several messages, from multiple replicated databases, without revealing any information about the identity of the desired message. This problem has attracted significant recent interest in information theory where the fundamental limits of the problem based on absolute guarantees (in contrast to computational guarantees as in [2]) have been investigated. In [3], the notion of PIR capacity is introduced as the maximum ratio of the desired message size to the total download size. Reference [3] has characterized 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 in different settings, such as, colluding databases [4]–[8], coded storage [9]–[11], coded and colluding databases [12]–[15], Byzantine databases [16]–[18], storage constrained databases and other storage related settings [19]–[26], multiple message PIR [27], [28], symmetric PIR [29]–[32], PIR with side information [33]–[40], cache-aided PIR [41]–[45], information leakage in PIR [46]–[48], private computation [49]–[51], security constraints and effects of adversaries and eavesdroppers on PIR [52]–[57], studies of optimal costs in PIR [58]–[61] and PIR under different channel configurations [62]–[64].

In all these works, two assumptions are made: All messages have the same size,[1] 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, i.e., it is unlikely that all files are equally probable to be downloaded by a user. The streaming service, in this case, has an a priori probability distribution over all the files, for example, from box office revenues and online rating systems. In addition, the media files cannot be assumed to be equal in size; some movies are longer, some are shorter. 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.

Manuscript received March 28, 2020; revised November 8, 2021; accepted December 2, 2021. Date of publication December 20, 2021; date of current version March 17, 2022. This work was supported by NSF under Grant CCF 17-13977 and Grant ECCS 18-07348. An earlier version of this paper was presented in part at the 2020 IEEE Global Communications Conference [DOI: 10.1109/GLOBECOM42002.2020.9348234] and in part at the 2021 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT45174.2021.9517765]. *(Corresponding author: Sennur Ulukus.)*

Sajani Vithana and Sennur Ulukus are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: spallego@umd.edu; ulukus@umd.edu).

Karim Banawan is with the Electrical Engineering Department, Faculty of Engineering, Alexandria University, Alexandria 21544, Egypt (e-mail: kbanawan@umd.edu).

Communicated by J. Kliewer, Associate Editor for Coding Techniques.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3136583.

Digital Object Identifier 10.1109/TIT.2021.3136583

[1] With the exception of [29], which characterizes the capacity of the symmetric PIR (SPIR) problem for heterogeneous file sizes (without considering a priori probabilities of retrieval) to be $R_k = \frac{L_k}{\max_i L_i}\left(1 - \frac{1}{N}\right)$, where R_k is the rate of retrieving message k. The achievable scheme follows by dividing the files into partitions of length $N-1$ and repeating the original SPIR scheme in each partition. This scheme zero-pads shorter messages so that their lengths are equal to that of the longest message.

In this paper, we introduce the semantic PIR problem. We extend the notion of the PIR capacity to deal with the heterogeneity of message sizes and prior probabilities. 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 needs to be the same for all messages; thus, the expected download cost is equal to the download cost for each individual message. Hence, the retrieval rate achieved by a given scheme 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 how semantic PIR capacity compares to classical PIR capacity, and whether there is a PIR capacity gain from exploiting the message semantics.

In this paper, we characterize the exact semantic PIR capacity for arbitrary parameters. 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. Note that this is crucial in our semantic problem due to the heterogeneous message sizes, unlike the majority of the literature that utilizes uniform subpacketization within their schemes [61]. 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 [3], in that, our scheme uses the same k-sums idea of [3]. 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. The message is retrieved using a single set of queries, which is chosen uniformly randomly from the query options to ensure privacy. This is similar to the scheme presented in [47] with an extension to more than two databases (see also [60]). We provide a matching converse that takes into account the heterogeneity of message sizes, resulting in settling the semantic PIR capacity. Additionally, we provide two extensions of the semantic PIR problem, namely, the semantic PIR from MDS-coded databases and the semantic PIR from colluding databases. For both extensions, we derive the exact PIR capacity in addition to providing a corresponding optimal scheme.

The semantic PIR capacity is a function of the message sizes and the a priori probability distribution. The expression 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 no matter what the message priors are, i.e., priors cannot be exploited to increase the PIR capacity if the message lengths are the same. 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 which depends only on the number of databases and the number of messages. Note that, by classical PIR capacity, we mean the classical PIR capacity expression, which may not be attainable for heterogeneous file sizes. 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, as it assumes.

II. Problem Formulation

We consider a setting, where N non-colluding databases store K independent messages (files), $W_1, \ldots, W_K$, in a replicated fashion. 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[2] p_i, such that $p_i > 0$ for $i = 1, \ldots, K$. The a priori probability distribution is globally known at the databases and the user. We assume that all message symbols are picked from a finite field[3] $\mathbb{F}_s$. 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,[4] such that $L_1 \geq L_2 \geq \cdots \geq L_K$. We assume that the messages stored in databases are mutually independent (which in turn implies pairwise independence). Hence, assuming that the message sizes are expressed in s-ary symbols,

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

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

In semantic PIR, a user needs to retrieve a message W_i without revealing the index i to any individual database. To that end, the user sends a query to each database. The query sent to the nth database to retrieve W_i is denoted by $Q_n^{[i]}$ for $n = 1, \ldots, N$. Prior to retrieval, the user does not have any information about the message contents. Hence, queries sent to the databases to retrieve messages are independent of the messages, i.e., the mutual information between messages and queries is zero,

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

Once the databases receive the queries, they generate answer strings to send back to the user. Specifically, the nth database prepares an answer string $A_n^{[i]}$ which is a deterministic function of the stored messages $W_1, \ldots, W_K$ and the received query $Q_n^{[i]}$. Therefore,

$$H(A_n^{[i]} | Q_n^{[i]}, W_1, \ldots, W_K) = 0, \ i = 1, \ldots, K, \ n = 1, \ldots, N \tag{4}$$

[2]We assume that $p_i > 0$ for all $i \in [K]$ without loss of generality, as $p_j = 0$ for some j implies that this message, W_j, is either non-existent or never requested by the user. Hence, the setting can be reduced to a semantic PIR problem with $K - 1$ messages, each with $p_i > 0$.

[3]In this work, it suffices to work with the binary field, hence, symbols can be interpreted as bits.

[4]This is for ease of expression of the capacity formula in (9). The largest length should have the largest coefficient in the expression in (9) in order to have the largest achievable rate and the tightest converse.

For a feasible PIR scheme, two conditions need to be satisfied, namely, the correctness and the privacy constraints. These are formally described as follows.

Correctness: The user should be able to perfectly retrieve the desired message as soon as the answer strings to the queries are received from the respective databases. Therefore,

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

Privacy: To protect the privacy of the desired message index i, the queries should not leak any information about i. Formally, for the nth database, the a posteriori probability of the message index i given a query $Q_n^{[i]}$ should be equal to the a priori probability of the message index i. That is, the random variable representing the desired message index, θ, should be independent of the received set of queries. Therefore,

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

The privacy constraint (6) along with the independence of messages and queries (3) implies,

$$(Q_n^{[i]},A_n^{[i]},W_1,\dots,W_K)\sim(Q_n^{[j]},A_n^{[j]},W_1,\dots,W_K),\quad n=1,\dots,N,\ i,j=1,\dots,K,\ i\neq j \tag{7}$$

An achievable semantic PIR scheme π is a scheme that satisfies the correctness constraint (5) and the privacy constraint (6). Due to the heterogeneity of message sizes and a priori probabilities, in this work, we define the performance metric, the expected retrieval rate $R(\pi)$ for any scheme $\pi\in\Pi$, where Π is the set of all PIR schemes satisfying the correctness and privacy constraints given in (5) and (6), as the ratio of the expected retrieved message size to the expected download size, i.e.,

$$R(\pi)=\frac{\mathbb{E}[L]}{\mathbb{E}[D]},\quad \pi\in\Pi \tag{8}$$

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 $\mathbb{E}[\cdot]$ in $\mathbb{E}[L]$ is with respect to the a priori probability distribution. Note that $\mathbb{E}[L]$ is fixed for any scheme as it is completely determined by the set of message lengths and prior probabilities which are given in the semantic PIR setting. The expectation $\mathbb{E}[\cdot]$ in $\mathbb{E}[D]$ is with respect to the distribution of the queries. Note that $\mathbb{E}[D]$ does not depend on the prior distribution as for any desired message, the download cost must remain the same to preserve privacy. Therefore, $\mathbb{E}[D]$ of a given scheme is completely determined by the structure of the scheme. The semantic PIR capacity is defined as the supremum of the expected retrieval rates over all achievable PIR schemes in Π, i.e., $C=\sup_{\pi\in\Pi}R(\pi)$. Moreover, the optimal semantic PIR scheme $\pi^*\in\Pi$ is an achievable scheme that minimizes the expected download cost, i.e., $\pi^*=\arg\min_{\pi\in\Pi}\mathbb{E}[D]$.

III. Main Results and Discussions

In this section, we present the main results of the paper. Our first result is a complete characterization of the semantic PIR capacity. The semantic PIR capacity depends on the message sizes and prior probability distribution.

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

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

$$=\left(\frac{L_1}{\sum_{i=1}^K p_iL_i}+\frac{1}{N}\frac{L_2}{\sum_{i=1}^K p_iL_i}+\cdots+\frac{1}{N^{K-1}}\frac{L_K}{\sum_{i=1}^K p_iL_i}\right)^{-1} \tag{10}$$

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

The achievability proof of Theorem 1 is presented in Section IV and the converse proof is presented in Section V. Next, we have a few corollaries and remarks.

The following corollary gives a necessary and sufficient condition for the cases at which the semantic capacity exceeds the classical PIR capacity.

Corollary 1 (A Necessary and Sufficient Condition for Semantic Capacity Gain): The semantic PIR capacity is strictly larger than the classical PIR capacity (with uniform priors and message sizes) if and only if,

$$\sum_{i=1}^K\frac{1}{N^{i-1}}(L_i-\mathbb{E}[L])<0 \tag{11}$$

which is further equivalent to,

$$\sum_{i=1}^K\sum_{j=1}^K\frac{p_j}{N^{i-1}}(L_i-L_j)<0 \tag{12}$$

Proof: The proof follows from comparing the semantic PIR capacity expression in (9) and the classical PIR capacity, C_{PIR}, in [3],

$$C_{PIR}=\left(1+\frac{1}{N}+\cdots+\frac{1}{N^{K-1}}\right)^{-1} \tag{13}$$

Hence, $C>C_{PIR}$ implies

$$\frac{L_1}{\mathbb{E}[L]}+\frac{1}{N}\frac{L_2}{\mathbb{E}[L]}+\cdots+\frac{1}{N^{K-1}}\frac{L_K}{\mathbb{E}[L]}<1+\frac{1}{N}+\cdots+\frac{1}{N^{K-1}} \tag{14}$$

Ordering the terms leads to,

$$\sum_{i=1}^K\frac{1}{N^{i-1}}(L_i-\mathbb{E}[L])<0 \tag{15}$$

Noting $L_i=\sum_{j=1}^K p_jL_i$, since p_j sum to 1, and $\mathbb{E}[L]=\sum_{j=1}^K p_jL_j$ by definition of expectation,

$$\sum_{i=1}^K\sum_{j=1}^K\frac{p_j}{N^{i-1}}(L_i-L_j)<0 \tag{16}$$

■

Remark 1: The condition in (11) is a statement about the sum weighted (by $\frac{1}{N^{i-1}}$) deviation of message size from its

expected value. Note that the expected value of the message size $\mathbb{E}[L]$ is a function of the message sizes L_i and the prior distribution p_i for $i = 1, \ldots, K$.

Remark 2: The intuition behind the condition in Corollary 1 is as follows. The set of message lengths and prior probabilities need to result in a large enough expected message length, which further implies that the longer messages need to be more popular, in order for the semantic PIR rate to outperform the classical PIR rate.

Remark 3: 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, (11) simplifies to,

$$(L_1 - (p_1 L_1 + p_2 L_2)) + \frac{1}{2}(L_2 - (p_1 L_1 + p_2 L_2)) < 0 \quad (17)$$

$$p_2(L_1 - L_2) + \frac{1}{2} p_1 (L_2 - L_1) < 0 \quad (18)$$

$$p_2 - \frac{1}{2} p_1 < 0 \quad (19)$$

$$p_1 > \frac{2}{3} \quad (20)$$

where (19) 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 the more likely message is 4 times more likely and 4 times longer than the less likely message, i.e., 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 $C_{PIR} = \frac{2}{3} = \frac{30}{45}$. That is, for this case, $C_{PIR} = \frac{2}{3} < C = \frac{34}{45}$.

Remark 4: We further expand on Remark 3 above by noting the following fact. The classical PIR capacity is a formula, as given in (13), 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 3 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.

As a follow up to Remark 4, we note that the achievable scheme proposed in this paper always outperforms zero-padding shorter messages and applying the classical PIR scheme for so-constructed equal-length messages. This is proved in the following corollary.

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

Proof: We first calculate the general achievable rate for the classical PIR scheme with zero-padding, R_{ach}. Noting $L_1 \geq L_2 \geq \cdots \geq L_K$, we zero-pad messages $2, \ldots, K$ until the message sizes are all equal to L_1. Next, we apply the classical PIR scheme with the common message size L_1. Then, the download cost (and the expected download cost) becomes,

$$\mathbb{E}[D] = D = \frac{L_1}{C_{PIR}} \quad (21)$$

Now, using C_{PIR} in (13) in equation (21) above, we obtain,

$$R_{ach} = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} \quad (22)$$

$$= \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N} \frac{L_1}{\mathbb{E}[L]} + \cdots + \frac{1}{N^{K-1}} \frac{L_1}{\mathbb{E}[L]} \right)^{-1} \quad (23)$$

Note repeated L_1 in the expression in (23). Comparing R_{ach} in (23) with the semantic PIR capacity in (9), we deduce that $R_{ach} \leq C$ as $L_1 \geq L_2 \geq \cdots \geq L_K$. ■

Remark 5: 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 (13). 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.[5]

Remark 6: Similar to classical PIR, the semantic PIR capacity increases with the number of databases, N. As the number of databases approaches infinity, the capacity approaches $\frac{\mathbb{E}[L]}{L_1}$. The reason why this asymptotic capacity is less than 1 is that the download cost must remain constant at L_1 (as the longest message achieves a rate of 1) irrespective of the desired message. The semantic PIR capacity decreases as the number of messages, K, increases. As K approaches infinity, the semantic PIR capacity is lower bounded by

$$C > \frac{\mathbb{E}[L]}{L_1} \left(1 - \frac{1}{N} \right) \quad (24)$$

IV. Achievability Proof

In this section, we present two PIR schemes that achieve the semantic PIR capacity given in Theorem 1. For each scheme, we first formally present the scheme, then we verify its correctness and privacy, calculate its achievable rate, and give explicit examples for illustration.

A. Achievable Semantic PIR Scheme 1

The scheme is based on the iterative structure of the achievable scheme in [3]. In this scheme, the user downloads k-sums from the messages for $k = 1, \ldots, K$. 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

[5] It is worth noting that classical PIR schemes need to be designed to satisfy the privacy constraint irrespective of the prior distribution. Nevertheless, the performance of the classical PIR schemes does not depend on the prior distribution as they consider uniform message sizes. This is in contrast to the semantic PIR problem, where the heterogeneity of the message sizes can be exploited to enhance the retrieval rate based on the properties of the prior distribution.

TABLE I
SINGLETON QUERIES

Message	Database 1	Database 2	...	Database N
W_j	$a_1, \ldots, a_{\upsilon_j}$	$a_{\upsilon_j+1}, \ldots, a_{2\upsilon_j}$	...	$a_{(N-1)\upsilon_j+1}, \ldots, a_{N\upsilon_j}$
$W_i, i \neq j$	$b_1, \ldots, b_{\upsilon_i}$	$b_{\upsilon_i+1}, \ldots, b_{2\upsilon_i}$	...	$b_{(N-1)\upsilon_i+1}, \ldots, b_{N\upsilon_i}$

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[6] $\upsilon_1, \upsilon_2, \ldots, \upsilon_K$ 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 these parameters are as follows:

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \vdots \\ \upsilon_K \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} \frac{1}{N} & -\frac{N-1}{N^2} & -\frac{N-1}{N^3} & \ldots & -\frac{N-1}{N^K} \\ 0 & \frac{1}{N^2} & -\frac{N-1}{N^3} & \ldots & -\frac{N-1}{N^K} \\ 0 & 0 & \frac{1}{N^3} & \ldots & -\frac{N-1}{N^K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{N^K} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ \vdots \\ L_K \end{bmatrix} \tag{25}$$

where α can be chosen as the gcd of the vector elements resulting from the matrix multiplication in the right hand side of (25). This choice will become clear in Section IV-A.1.

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. I.e., if the number of elements in a subpacket of W_i is ℓ_i, let W_i be denoted by, $W_i = (W_i(1), \ldots, W_i(\ell_i))$ for $i \in \{1, \ldots, K\}$. For each message W_i, the user uniformly and randomly chooses a permutation of the ℓ_i indices out of the $\ell_i!$ options, indicated by $(\gamma_i(1), \gamma_i(2), \ldots, \gamma_i(\ell_i))$, which is independent of all other message permutations. Then, the permutation of the elements of W_i is given by, $\Gamma(W_i(1), \ldots, W_i(\ell_i)) = (W_i(\gamma_i(1)), \ldots, W_i(\gamma_i(\ell_i)))$. This process simply shuffles the elements of message vectors uniformly and randomly irrespective of the message requirement. All queries generated by the user in the scheme are based on these permuted indices.

3) *Singletons:* Download υ_k different bits from message W_k from the nth database, where $n = 1, \ldots, N$ and $k = 1, \ldots, K$. Table I shows the singletons downloaded from the required message W_j and any other message W_i, $i \neq j$. Note that the permuted elements of W_j and W_i are denoted by a's and b's respectively.

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 an already downloaded W_i bit from another database and a new bit of W_j. I.e., if $\upsilon_j > \upsilon_i$ user sends queries of the form $(a_{N\upsilon_j+1} + b_{\upsilon_i+1})$, ..., $(a_{N\upsilon_j+\upsilon_i} + b_{2\upsilon_i})$, ..., $(a_{N\upsilon_j+(N-2)\upsilon_i+1} + b_{(N-1)\upsilon_i+1})$, ..., $(a_{N\upsilon_j+(N-1)\upsilon_i} + b_{N\upsilon_i})$ to database 1. Note that each $\min\{\upsilon_i, \upsilon_j\} = \upsilon_i$ side information bit downloaded from each of the databases 2 to N in the previous step have been utilized exactly once in the 2-sums of database 1. Queries of the same form are sent to all databases, which contain new bits of W_j and all the already downloaded bits of $W_i, i \neq j$ from the rest of the databases. Each side information bit from the previous step if utilized only once in a given database.

If $\min\{\upsilon_i, \upsilon_j\} = \upsilon_j$ user can randomly pick any υ_j side information bits out of the υ_i bits from each database and follow the same steps as above, ensuring that any given side information bit from a different database in the previous step is utilized only once in a given database.

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 $k = 3, 4, \ldots, K$. 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 $(N-1)^{k-1}\min\{\upsilon_{i_1}, \ldots, \upsilon_{i_k}\}$ sums from new bits of the undesired messages.

1) Rate of Semantic PIR Scheme 1: In this PIR scheme, the total number of downloaded bits remains constant for all message requirements of the user in order to guarantee privacy. Therefore, $\mathbb{E}[D]$ in (8) can be calculated by counting the total number of bits in the set of queries sent to the databases by the user to download any message. Within the set of queries, there are $\sum_{i=1}^{K} N\upsilon_i$ number of singletons and $\sum_{i=t}^{K} N(N-1)^{t-1}\upsilon_i\binom{i-1}{t-1}$ number of sums of t elements. Therefore,

$$\mathbb{E}[D] = \sum_{i=1}^{K} N\upsilon_i + \sum_{t=2}^{K}\sum_{i=t}^{K} N(N-1)^{t-1}\upsilon_i\binom{i-1}{t-1} \tag{26}$$

$$= N\left[\sum_{i=1}^{K}\upsilon_i + \sum_{i=2}^{K}\sum_{t=2}^{i}(N-1)^{t-1}\upsilon_i\binom{i-1}{t-1}\right] \tag{27}$$

[6]This set of parameters determines the nonuniform subpacketization of a given semantic PIR setting with arbitrary message lengths. It also controls the numbers of stages in the next steps of the scheme (numbers of ℓ-sums, $\ell \in \{1, \ldots, K\}$) such that the scheme is private and capacity achieving.

$$= N\left[\sum_{i=1}^{K} v_i + \sum_{i=2}^{K} v_i \left(\sum_{t=0}^{i}(N-1)^t \binom{i-1}{t} - 1\right)\right] \tag{28}$$

$$= N\left[\sum_{i=1}^{K} v_i + \sum_{i=2}^{K} v_i \left(N^{i-1} - 1\right)\right] \tag{29}$$

$$= \sum_{i=1}^{K} v_i N^i \tag{30}$$

In order to calculate $\mathbb{E}[L]$, assume that the desired message is W_j. There are Nv_j number of singletons of W_j in the set of queries sent to the databases to retrieve W_j. The scheme can recover $N(N-1)^{t-1}v_j\binom{j-1}{t-1} + N(N-1)^{t-1}\sum_{i=j+1}^{K} v_i\binom{i-2}{t-2}$ number of W_j bits using the tth block of the scheme (sum of t elements) when $t \leq j$, where the first term in the sum corresponds to t-sums with the shortest message being W_j and the second term corresponds to t-sums with the shortest message being some other message ($\neq W_j$). When $t > j$ this scheme is able to retrieve $\sum_{i=t}^{K} N(N-1)^{t-1}v_i\binom{i-2}{t-2}$ number of W_j bits as there should be at least $t-j$ number of messages in the sum that are shorter than L_j. Therefore, the total number of useful bits of W_j retrieved, U_j, is given by,

$$\begin{aligned}
U_j =& Nv_j + \sum_{t=2}^{j}\left(N(N-1)^{t-1}v_j\binom{j-1}{t-1}\right.\\
&\left. + \sum_{i=j+1}^{K} N(N-1)^{t-1}v_i\binom{i-2}{t-2}\right)\\
&+ \sum_{t=j+1}^{K}\sum_{i=t}^{K} N(N-1)^{t-1}v_i\binom{i-2}{t-2} \qquad (31)\\
=& Nv_j\left(1 + \sum_{t=2}^{j}(N-1)^{t-1}\binom{j-1}{t-1}\right)\\
&+ \sum_{t=2}^{j}\sum_{i=j+1}^{K} N(N-1)^{t-1}v_i\binom{i-2}{t-2}\\
&+ \sum_{t=j+1}^{K}\sum_{i=t}^{K} N(N-1)^{t-1}v_i\binom{i-2}{t-2} \qquad (32)\\
=& Nv_j\left(1 + (N-1)\binom{j-1}{1} + (N-1)^2\binom{j-1}{2}\right.\\
&\left. + \cdots + (N-1)^{j-1}\binom{j-1}{j-1}\right)\\
&+ Nv_{j+1}\left(\sum_{t=2}^{j}(N-1)^{t-1}\binom{j-1}{t-2}\right)\\
&+ Nv_{j+2}\left(\sum_{t=2}^{j}(N-1)^{t-1}\binom{j}{t-2}\right) + \cdots\\
&+ Nv_K\left(\sum_{t=2}^{j}(N-1)^{t-1}\binom{K-2}{t-2}\right)\\
&+ Nv_{j+1}(N-1)^j\binom{j-1}{j-1}\\
&+ Nv_{j+2}\left((N-1)^j\binom{j}{j-1} + (N-1)^{j+1}\binom{j}{j}\right)\\
&+ \cdots + Nv_K\left((N-1)^j\binom{K-2}{j-1}\right.\\
&+ (N-1)^{j+1}\binom{K-2}{j} + \ldots\\
&\left. +(N-1)^{K-1}\binom{K-2}{K-2}\right) \qquad (33)\\
=& Nv_j(N-1+1)^{j-1}\\
&+ Nv_{j+1}\left((N-1)\binom{j-1}{0} + (N-1)^2\binom{j-1}{1}\right.\\
&\left. + \cdots + (N-1)^j\binom{j-1}{j-1}\right) + Nv_{j+2}\left((N-1)\binom{j}{0}\right.\\
&\left. +(N-1)^2\binom{j}{1} + \cdots + (N-1)^{j+1}\binom{j}{j}\right) + \ldots\\
&+ Nv_k\left((N-1)\binom{K-2}{0} + (N-1)^2\binom{K-2}{1}\right.\\
&\left. + \cdots + (N-1)^{K-1}\binom{K-2}{K-2}\right) \qquad (34)\\
=& N^j v_j + N(N-1)(N-1+1)^{j-1}v_{j+1}\\
&+ N(N-1)(N-1+1)^j v_{j+2} + \ldots\\
&+ N(N-1)(N-1+1)^{K-2}v_K \qquad (35)\\
=& N^j v_j + (N-1)\sum_{i=j+1}^{K} N^{i-1}v_i \qquad (36)
\end{aligned}$$

Thus, the scheme retrieves $N^j v_j + (N-1)\sum_{i=j+1}^{K} N^{i-1}v_i$ number of useful bits of the required message at a time. Hence, we define *subpacketization* for message W_j as U_j, where

$$U_j = N^j v_j + (N-1)\sum_{i=j+1}^{K} N^{i-1}v_i, \quad j = 1, \ldots, K \tag{37}$$

We then need the message sizes to be a common multiple of their own subpacketizations,

$$L_j = \alpha U_j, \quad j = 1, \ldots, K \tag{38}$$

We note that α should be the same for all j in (38) to guarantee privacy.

The requirements in (37) and (38) can be written succinctly as a matrix equation,

$$\begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_K \end{bmatrix} = \alpha \begin{bmatrix} N & N(N-1) & \ldots & N^{K-1}(N-1) \\ 0 & N^2 & \ldots & N^{K-1}(N-1) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & N^K \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_K \end{bmatrix} \tag{39}$$

Since $L_1, \ldots, L_K$ are parameters (inputs) to the scheme, the internal parameters $v_1, \ldots, v_K$ can be calculated by inverting the matrix as,

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_K \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} \frac{1}{N} & -\frac{N-1}{N^2} & -\frac{N-1}{N^3} & \ldots & -\frac{N-1}{N^K} \\ 0 & \frac{1}{N^2} & -\frac{N-1}{N^3} & \ldots & -\frac{N-1}{N^K} \\ 0 & 0 & \frac{1}{N^3} & \ldots & -\frac{N-1}{N^K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{N^K} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ \vdots \\ L_K \end{bmatrix} \tag{40}$$

Here, α 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 (40). This allows the shortest subpacketization levels for all messages for increased flexibility.

The total number of bits downloaded calculated in (125) and the number of useful bits downloaded calculated in (36) are both within one subpacketization level. This subpacketization level downloads are repeated α times to download the entire file; see also (38). Thus, we calculate the achievable rate of this scheme as,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} \tag{41}$$

$$= \frac{\sum_{i=1}^{K} p_i U_i}{\sum_{i=1}^{K} N^i \upsilon_i} \tag{42}$$

$$= \frac{\frac{1}{\alpha}\sum_{i=1}^{K} p_i L_i}{\sum_{i=1}^{K} \frac{1}{\alpha} N^i (N^{-i} L_i - \sum_{j=i+1}^{K} (N-1) N^{-j} L_j)} \tag{43}$$

$$= \frac{\mathbb{E}[L]}{\sum_{i=1}^{K} L_i - (N-1)\sum_{i=1}^{K}\sum_{j=i+1}^{K} N^{-j} L_j N^i} \tag{44}$$

$$= \mathbb{E}[L] / \left(\sum_{i=1}^{K} L_i - (N-1) \left(\sum_{j=2}^{K} N^{-j+1} L_j + \sum_{j=3}^{K} N^{-j+2} L_j + \cdots + N^{-1} L_K \right) \right) \tag{45}$$

$$= \mathbb{E}[L] / \Big(L_1 + L_2 \left(1 - (N-1)N^{-1}\right) + \ldots + L_K \left(1 - (N-1)(N^{-(K-1)} + \cdots + N^{-1})\right) \Big) \tag{46}$$

$$= \frac{\mathbb{E}[L]}{L_1 + \frac{L_2}{N} + \frac{L_3}{N^2} + \cdots + \frac{L_K}{N^{K-1}}} \tag{47}$$

$$= \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N}\frac{L_2}{\mathbb{E}[L]} + \cdots + \frac{1}{N^{K-1}}\frac{L_K}{\mathbb{E}[L]} \right)^{-1} \tag{48}$$

where (43) follows by applying (38) in the numerator and writing υ_i in terms of L_j using (40) in the denominator. This concludes the derivation of the achievable rate.

Remark 7: We assume that message L_i has a length which is a multiple of N^i to aid smooth computation of $\upsilon_1, \ldots, \upsilon_K$. This is automatically satisfied by the assumption of all message lengths being multiples of N^K in [3].

2) Proof of Privacy: Since $L_1 \geq L_2 \geq \cdots \geq L_K$ we have $\upsilon_1 \geq \upsilon_2 \geq \cdots \geq \upsilon_K$. A given database receives a set of queries for $\upsilon_1, \upsilon_2, \ldots, \upsilon_K$ numbers of bits of $W_1, W_2, \ldots, W_K$, respectively, as singletons and $(N-1)^{t-1} \min\{\upsilon_{i_1}, \ldots, \upsilon_{i_t}\}$ bit-wise t-sums of $W_{i_1}, \ldots, W_{i_t}$, for $t = 2, \ldots, K$. According to the query generation procedure, no bit of any message is requested from a given database more than once as a singleton or as an element of a sum. Any given database receives the exact same set of queries in type, irrespective of the desired message of the user. Therefore, two sets of queries corresponding to two different message requirements received by a given database can only differ from the permutations used in each message in the index preparation step. Let $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K)$ be a sample realization of permutations used in the query generation process when the message requirement θ is W_k, where $\mathbf{w}_i = \Gamma_i(W_i(1), W_i(2), \ldots, W_i(\ell_i))$ for $i \in \{1, \ldots, K\}$ with permutation functions Γ_i that independently and randomly permute the ℓ_i elements of W_i, where ℓ_i is the subpacketization of W_i. Therefore, the probability of sending the set of queries q for a given message requirement $\theta = k$ is equal to the probability of choosing the corresponding sample realization of permutations of the message bits when downloading W_k. This probability is calculated by,

$$P(Q_n = q|\theta = k) = P(\text{permutation} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K)|\theta = k) \tag{49}$$

$$= \prod_{i=1}^{K} P(\text{permutation of } W_i = \mathbf{w_i}|\theta = k) \tag{50}$$

$$= \prod_{i=1}^{K} \left(\frac{1}{\ell_i}\right)\left(\frac{1}{\ell_i - 1}\right)\cdots\left(\frac{1}{\ell_i - \ell_i + 1}\right) \tag{51}$$

for $n \in \{1, \ldots, N\}$, where Q_n is the random variable representing the set of queries sent to database n. This yields,

$$P(Q_n = q|\theta = i) = P(Q_n = q|\theta = j), \quad \forall i, j \in \{1, \ldots, K\}, \quad n \in \{1, \ldots, N\} \tag{52}$$

as $P(Q_n = q|\theta = k)$ is independent of k by the above calculation. The a posteriori probability of the user needing W_i given a realization of the set of queries received by any given database is given by,

$$P(\theta = i|Q = q) = \frac{P(Q = q|\theta = i)P(\theta = i)}{\sum_{j=1}^{K} P(Q = q|\theta = j)P(\theta = j)} \tag{53}$$

Using (52),

$$P(\theta = i|Q = q) = \frac{P(Q = q|\theta = i)P(\theta = i)}{\sum_{j=1}^{K} P(Q = q|\theta = i)P(\theta = j)} \tag{54}$$

$$= P(\theta = i) \tag{55}$$

which ensures that this scheme is private, since it implies that θ and Q are independent.

B. Examples of Semantic PIR Scheme 1

1) 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,

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \tag{56}$$

where $\alpha = \gcd\{\frac{L_1}{2} - \frac{L_2}{4}, \frac{L_2}{4}\}$. By direct substitution, we get,

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} 448 \\ 64 \end{bmatrix} \tag{57}$$

TABLE II
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
$a_{15} + b_2$	$a_{16} + b_1$

TABLE III
THE QUERY TABLE FOR THE RETRIEVAL OF W_2

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

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 each from the two databases.
- Sums of twos: Download $(N-1)\upsilon_2 = 1$ sum of W_1 and W_2 bits each from the two 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 II and III show the queries sent to the databases to retrieve W_1 and W_2, respectively.

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,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} = \frac{p_1 L_1 + p_2 L_2}{p_1 D + p_2 D} = p_1 \frac{L_1}{D} + p_2 \frac{L_2}{D} \tag{58}$$

$$= p_1 R_1 + p_2 R_2 = \frac{8}{9} p_1 + \frac{2}{9} p_2 \tag{59}$$

This matches the capacity expression in Theorem 1 as,

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N} \frac{L_2}{\mathbb{E}[L]} \right)^{-1} \tag{60}$$

$$= (1024 p_1 + 256 p_2) \left(1024 + \frac{256}{2} \right)^{-1} \tag{61}$$

$$= \frac{8}{9} p_1 + \frac{2}{9} p_2 \tag{62}$$

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

$$C = \left(1 + \frac{1}{N} \right)^{-1} = \left(1 + \frac{1}{2} \right)^{-1} = \frac{2}{3} \tag{63}$$

The semantic PIR capacity in (62) exceeds the classical PIR capacity in (63) when

$$\frac{8}{9} p_1 + \frac{2}{9} p_2 > \frac{2}{3} \tag{64}$$

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, in this case.

Remark 8: Although it is apparent in this example that the rate of semantic PIR is lower than the capacity of classical PIR for $p_1 < \frac{2}{3}$, as discussed in Remark 3 and Remark 4, there is a subtle aspect that should be addressed for a fair comparison. To see this, let us take the case of uniform a priori distribution, i.e., $p_1 = p_2 = \frac{1}{2}$, i.e., a case where $p_1 < \frac{2}{3}$. In this case, the semantic PIR capacity using (62) is $\frac{5}{9}$. In order to properly use the classical PIR scheme in [3], messages need to be of equal size. One way to do this is to zero-pad the shorter message to be of length 1024 bits as well. In this case, 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,

$$R_{ach} = \frac{1/2 \times 1024 + 1/2 \times 256}{1536} = \frac{5}{12} < \frac{5}{9} < \frac{6}{9} \tag{65}$$

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

2) Example 2: $N = 4, K = 3$, $L_1 = 8192$ Bits, $L_2 = 2048$ Bits, $L_3 = 512$ Bits: First, the message indices are independently and uniformly permuted. The first, second, and third messages after permutations are denoted by bits a_i, b_i and c_i, respectively.

- Message indexing and calculation of υ_i: Messages are indexed such that the first message is the longest one, and the third message is the shortest one. Below, we will give the query table for downloading W_2, i.e., the medium-length message. The bits of W_2 are represented by b_i. We calculate υ_1, υ_2 and υ_3 as,

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} \frac{1}{4} & -\frac{3}{16} & -\frac{3}{64} \\ 0 & \frac{1}{16} & -\frac{3}{64} \\ 0 & 0 & \frac{1}{64} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} \tag{66}$$

where $\alpha = \text{gcd}\{\frac{L_1}{4} - \frac{3L_2}{16} - \frac{3L_3}{64}, \frac{L_2}{16} - \frac{3L_3}{64}, \frac{L_3}{64}\}$. By direct substitution, we get,

$$\begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \end{bmatrix} = \frac{1}{\alpha} \begin{bmatrix} 1640 \\ 104 \\ 8 \end{bmatrix} \tag{67}$$

Hence, $\alpha = \text{gcd}\{1640, 104, 8\} = 8$. Therefore, $\upsilon_1 = 205$, $\upsilon_2 = 13$ and $\upsilon_3 = 1$. The subpacketization levels of W_1, W_2 and W_3 are $U_1 = \frac{8192}{8} = 1024$, $U_2 = \frac{2048}{8} = 256$ and $U_3 = \frac{512}{8} = 64$, respectively.

- Singletons: Download $\upsilon_1 = 205$ bits of W_1, $\upsilon_2 = 13$ bits of W_2 and $\upsilon_3 = 1$ bits of W_3 each from the four databases.
- Sums of twos: Download $(N-1)\upsilon_2 = 39$ sums of W_1 and W_2 and $(N-1)\upsilon_3 = 3$ sums of W_2 and W_3 bits each from the four databases. Use the downloaded singletons from W_1, W_3 as side information with new W_2 bits. Download $(N-1)\upsilon_3 = 3$ bit-wise sums of W_1 and W_3 each from the four databases using fresh bits of both messages.

TABLE IV
THE QUERY TABLE FOR THE RETRIEVAL OF W_2

Database 1	Database 2	Database 3	Database 4
$a_1, \ldots, a_{205}$ $b_1, \ldots, b_{13}$ c_1	$a_{206}, \ldots, a_{410}$ $b_{14}, \ldots, b_{26}$ c_2	$a_{411}, \ldots, a_{615}$ $b_{27}, \ldots, b_{39}$ c_3	$a_{616}, \ldots, a_{820}$ $b_{40}, \ldots, b_{52}$ c_4
$a_{206} + b_{53}$	$a_{411} + b_{92}$	$a_{616} + b_{131}$	$a_1 + b_{170}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_{218} + b_{65}$	$a_{423} + b_{104}$	$a_{628} + b_{143}$	$a_{13} + b_{182}$
$a_{411} + b_{66}$	$a_{616} + b_{105}$	$a_1 + b_{144}$	$a_{206} + b_{183}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_{423} + b_{78}$	$a_{628} + b_{117}$	$a_{13} + b_{156}$	$a_{218} + b_{195}$
$a_{616} + b_{79}$	$a_1 + b_{118}$	$a_{206} + b_{157}$	$a_{411} + b_{196}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_{628} + b_{91}$	$a_{13} + b_{130}$	$a_{218} + b_{169}$	$a_{423} + b_{208}$
$b_{209} + c_2$	$b_{212} + c_3$	$b_{215} + c_4$	$b_{218} + c_1$
$b_{210} + c_3$	$b_{213} + c_4$	$b_{216} + c_1$	$b_{219} + c_2$
$b_{211} + c_4$	$b_{214} + c_1$	$b_{217} + c_2$	$b_{220} + c_3$
$a_{821} + c_5$	$a_{824} + c_8$	$a_{827} + c_{11}$	$a_{830} + c_{14}$
$a_{822} + c_6$	$a_{825} + c_9$	$a_{828} + c_{12}$	$a_{831} + c_{15}$
$a_{823} + c_7$	$a_{826} + c_{10}$	$a_{829} + c_{13}$	$a_{832} + c_{16}$
$a_{824} + b_{221} + c_8$	$a_{827} + b_{230} + c_{11}$	$a_{830} + b_{239} + c_{14}$	$a_{821} + b_{248} + c_5$
$a_{825} + b_{222} + c_9$	$a_{828} + b_{231} + c_{12}$	$a_{831} + b_{240} + c_{15}$	$a_{822} + b_{249} + c_6$
$a_{826} + b_{223} + c_{10}$	$a_{829} + b_{232} + c_{13}$	$a_{832} + b_{241} + c_{16}$	$a_{823} + b_{250} + c_7$
$a_{827} + b_{224} + c_{11}$	$a_{830} + b_{233} + c_{14}$	$a_{821} + b_{242} + c_5$	$a_{824} + b_{251} + c_8$
$a_{828} + b_{225} + c_{12}$	$a_{831} + b_{234} + c_{15}$	$a_{822} + b_{243} + c_6$	$a_{825} + b_{252} + c_9$
$a_{829} + b_{226} + c_{13}$	$a_{832} + b_{235} + c_{16}$	$a_{823} + b_{244} + c_7$	$a_{826} + b_{253} + c_{10}$
$a_{830} + b_{227} + c_{14}$	$a_{821} + b_{236} + c_5$	$a_{824} + b_{245} + c_8$	$a_{827} + b_{254} + c_{11}$
$a_{831} + b_{228} + c_{15}$	$a_{822} + b_{237} + c_6$	$a_{825} + b_{246} + c_9$	$a_{828} + b_{255} + c_{12}$
$a_{832} + b_{229} + c_{16}$	$a_{823} + b_{238} + c_7$	$a_{826} + b_{247} + c_{10}$	$a_{829} + b_{256} + c_{13}$

- Sums of threes: Download $(N-1)^2 \upsilon_3 = 9$ bit-wise sums involving all three messages from each database utilizing the downloaded sums of W_1 and W_3 from the other databases in the previous step as side information.

Table IV shows the queries sent to the databases to retrieve W_2.

The rate achieved by this scheme when downloading W_2 is $R_2 = \frac{256}{1092} = \frac{64}{273}$, and the rates achieved when downloading W_1 and W_3 are $R_1 = \frac{1024}{1092} = \frac{256}{273}$ and $R_3 = \frac{64}{1092} = \frac{16}{273}$, respectively. Therefore, the average rate R achieved by this scheme is,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} = \frac{p_1 L_1 + p_2 L_2 + p_3 L_3}{p_1 D + p_2 D + p_3 D} \tag{68}$$

$$= p_1 \frac{L_1}{D} + p_2 \frac{L_2}{D} + p_3 \frac{L_3}{D} = p_1 R_1 + p_2 R_2 + p_3 R_3 \tag{69}$$

$$= \frac{256}{273} p_1 + \frac{64}{273} p_2 + \frac{16}{273} p_3 \tag{70}$$

This matches the capacity expression in Theorem 1 as,

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N} \frac{L_2}{\mathbb{E}[L]} + \frac{1}{N^2} \frac{L_3}{\mathbb{E}[L]} \right)^{-1} \tag{71}$$

$$= (8192 p_1 + 2048 p_2 + 512 p_3) \left(8192 + \frac{2048}{4} + \frac{512}{4^2} \right)^{-1} \tag{72}$$

$$= \frac{256}{273} p_1 + \frac{64}{273} p_2 + \frac{16}{273} p_3 \tag{73}$$

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

$$C = \left(1 + \frac{1}{N} + \frac{1}{N^2} \right)^{-1} = \left(1 + \frac{1}{4} + \frac{1}{4^2} \right)^{-1} = \frac{16}{21} \tag{74}$$

The semantic PIR capacity in (73) exceeds the classical PIR capacity in (74) when

$$\frac{256}{273} p_1 + \frac{64}{273} p_2 + \frac{16}{273} p_3 > \frac{16}{21} \tag{75}$$

which is equivalent to

$$p_1 + \frac{1}{5} p_2 > \frac{4}{5} \tag{76}$$

C. Alternative Description of Semantic PIR Scheme 1

In this section, we present an alternative description to the semantic PIR scheme presented in Section IV-A.

The two descriptions are identical in terms of the queries generated considering the retrieval of the entire required message (all subpackets). However, the two descriptions differ in subpacketization and the scheme used within a subpacket.

Consider the general semantic PIR setting with K messages with arbitrary message lengths $L_1 \geq L_2 \geq \ldots \geq L_K$ and arbitrary probabilities of retrieval p_i, $i \in \{1, \ldots, K\}$. The alternative description requires the messages to be partitioned in to K segments, such that the first segment contains the first L_K bits of all messages, the second segment contains the next $L_{K-1} - L_K$ bits of messages $W_1, \ldots, W_{K-1}$ and the ℓth segment for $\ell \in \{3, \ldots, K\}$ contains $L_{K-\ell+1} - L_{K-\ell+2}$ bits of $W_1, \ldots, W_{K-\ell+1}$ that follow the bits in the $(\ell-1)$st segment.

Apply the classical PIR scheme in [3] to the 1) first segment with K messages with a subpacketization of N^K, 2) second segment with $K-1$ messages with a subpacketization of N^{K-1} and 3) ℓth segment with $K-\ell+1$ messages with a subpacketization of $N^{K-\ell+1}$, for $\ell \in \{3, \ldots, K\}$. The above three steps need to be followed irrespective of the message requirement for privacy. Note that the schemes used in each segment are private [3], and the fact that the K schemes corresponding to the K segments are always used, even though the required message may not be within a given segment, guarantees privacy. The achievable rate of this scheme is calculated as follows. The fixed download cost is given by,

$$D = \frac{L}{R} \tag{77}$$

$$= \frac{L_K}{\left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}}\right)^{-1}} + \frac{L_{K-1} - L_K}{\left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-2}}\right)^{-1}} + \ldots + \frac{L_2 - L_3}{\left(1 + \frac{1}{N}\right)^{-1}} + \frac{L_1 - L_2}{1} \tag{78}$$

$$= L_K \frac{1}{N^{K-1}} + L_{K-1} \frac{1}{N^{K-2}} + \ldots + L_2 \frac{1}{N} + L_1 \tag{79}$$

Therefore, the achievable rate is,

$$R = \frac{\mathbb{E}[L]}{D} \tag{80}$$

$$= \frac{\mathbb{E}[L]}{L_K \frac{1}{N^{K-1}} + L_{K-1} \frac{1}{N^{K-2}} + \ldots + L_2 \frac{1}{N} + L_1} \tag{81}$$

which is the capacity of semantic PIR in (9). Note that the description in Section IV-A provides a *systematic* way of calculating the nonuniform subpacketization based on the given set of message lengths. The scheme is then described on a single subpacket, which is repeatedly applied throughout the retrieval process in the same way. On the other hand, the alternative description has different uniform subpacketizations for different segments. Therefore, the scheme needs to be specified for each segement separately. This is illustrated in Fig. 1.

D. 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 and IV-B assumed message lengths that are multiples of N^K. The scheme can be viewed as an extension of the achievable scheme in [47] to work with arbitrary number of databases and heterogeneous message sizes. Our scheme shares similarities with [60]. 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.

The above steps describe all the N^K query options, out of which the user selects one with equal probability to retrieve the required message. Note that due to the cyclic shifts of all queries, this scheme has database symmetry, and the exact same set of queries constitutes the possible set of queries received by any given database, irrespective of the desired message of the user.

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

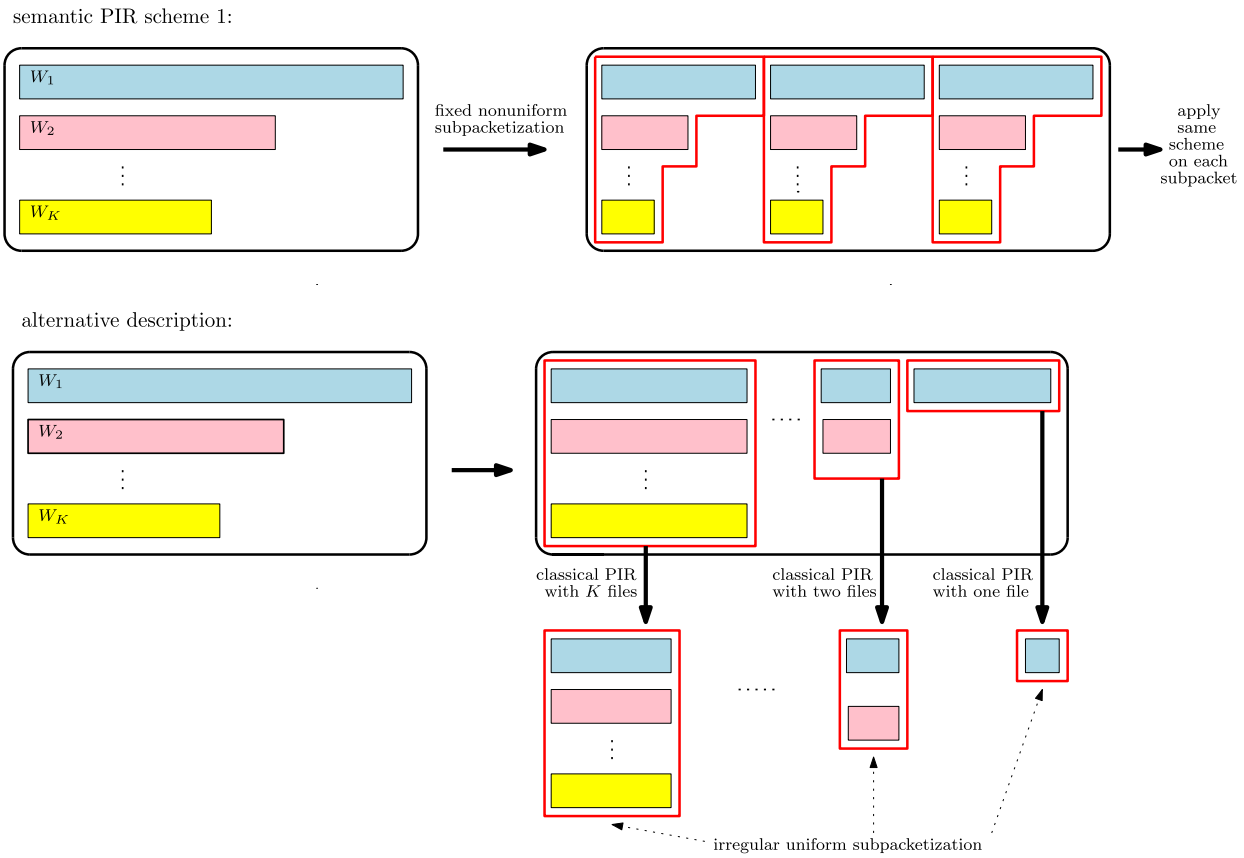

Fig. 1. Comparison of the two descriptions of semantic PIR scheme 1.

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.

Remark 9: Each query is chosen with probability $\frac{1}{N^K}$ as there are $\sum_{t=0}^{K}(N-1)^t\binom{K}{t} = N^K$ number of query options in total. Each element of the sum corresponds to the number of t-sums within the set of all possible queries that can be sent to a given database.

1) Rate of Semantic PIR Scheme 2: In this PIR scheme, each query option is utilized by the user with a probability of $\frac{1}{N^K}$ to download any desired message. When analyzing all possible queries that can be sent to all databases, we note that they have the same entries (in a shuffled way) irrespective of the desired message. Since all query entries are equally probable to be sent to the databases, we calculate $\mathbb{E}[D]$ by,

$$\mathbb{E}[D] = \sum_{i=1}^{K} p_i \frac{1}{N^K}\left(\sum_{t=1}^{K}\sum_{j=1}^{K-t+1} L_j (N-1)^{t-1}\binom{K-j}{t-1}\right) N \tag{82}$$

$$= \frac{1}{N^{K-1}}\sum_{j=1}^{K}\sum_{t=1}^{K-j+1} L_j (N-1)^{t-1}\binom{K-j}{t-1} \tag{83}$$

$$= \frac{1}{N^{K-1}}\sum_{j=1}^{K} L_j \sum_{t=0}^{K-j}(N-1)^t\binom{K-j}{t} \tag{84}$$

$$= \frac{1}{N^{K-1}}\sum_{j=1}^{K} L_j N^{K-j} \tag{85}$$

$$= L_1 + \frac{L_2}{N} + \frac{L_3}{N^2} + \cdots + \frac{L_K}{N^{K-1}} \tag{86}$$

where the second and third sums in (82) correspond to different t-sums and all possible longest messages within the t-sum, respectively. The p_i terms are ignored in (83) as the expected number of downloads per query set does not depend on the desired message.

For a given desired message, the number of downloaded useful bits is the length of the desired message (ignoring zero-padding, as it is ignored by the user upon receiving the answer strings). This remains constant regardless of the query set utilized by the user. Hence,

$$\mathbb{E}[L] = \sum_{i=1}^{K} p_i L_i \tag{87}$$

Thus, combining (86) and (87), the achievable rate of this scheme becomes,

$$R = \frac{\mathbb{E}[L]}{\mathbb{E}[D]} \tag{88}$$

$$= \frac{\mathbb{E}[L]}{L_1 + \frac{L_2}{N} + \frac{L_3}{N^2}\cdots + \frac{L_K}{N^{K-1}}} \tag{89}$$

$$= \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N}\frac{L_2}{\mathbb{E}[L]} + \cdots + \frac{1}{N^{K-1}}\frac{L_K}{\mathbb{E}[L]}\right)^{-1} \tag{90}$$

This concludes the derivation of the achievable rate.

TABLE V
THE QUERY TABLE FOR THE RETRIEVAL OF W_1

Probability	Database 1	Database 2	Database 3
$\frac{1}{27}$	W_1^1	W_1^2	ϕ
$\frac{1}{27}$	W_1^2	ϕ	W_1^1
$\frac{1}{27}$	ϕ	W_1^1	W_1^2
$\frac{1}{27}$	$W_1^1+W_2^1$	$W_1^2+W_2^1$	W_2^1
$\frac{1}{27}$	$W_1^2+W_2^1$	W_2^1	$W_1^1+W_2^1$
$\frac{1}{27}$	W_2^1	$W_1^1+W_2^1$	$W_1^2+W_2^1$
$\frac{1}{27}$	$W_1^1+W_2^2$	$W_1^2+W_2^2$	W_2^2
$\frac{1}{27}$	$W_1^2+W_2^2$	W_2^2	$W_1^1+W_2^2$
$\frac{1}{27}$	W_2^2	$W_1^1+W_2^2$	$W_1^2+W_2^2$
$\frac{1}{27}$	$W_1^1+W_3^1$	$W_1^2+W_3^1$	W_3^1
$\frac{1}{27}$	$W_1^2+W_3^1$	W_3^1	$W_1^1+W_3^1$
$\frac{1}{27}$	W_3^1	$W_1^1+W_3^1$	$W_1^2+W_3^1$
$\frac{1}{27}$	$W_1^1+W_3^2$	$W_1^2+W_3^2$	W_3^2
$\frac{1}{27}$	$W_1^2+W_3^2$	W_3^2	$W_1^1+W_3^2$
$\frac{1}{27}$	W_3^2	$W_1^1+W_3^2$	$W_1^2+W_3^2$
$\frac{1}{27}$	$W_1^1+W_2^1+W_3^1$	$W_1^2+W_2^1+W_3^1$	$W_2^1+W_3^1$
$\frac{1}{27}$	$W_1^2+W_2^1+W_3^1$	$W_2^1+W_3^1$	$W_1^1+W_2^1+W_3^1$
$\frac{1}{27}$	$W_2^1+W_3^1$	$W_1^1+W_2^1+W_3^1$	$W_1^2+W_2^1+W_3^1$
$\frac{1}{27}$	$W_1^1+W_2^2+W_3^1$	$W_1^2+W_2^2+W_3^1$	$W_2^2+W_3^1$
$\frac{1}{27}$	$W_1^2+W_2^2+W_3^1$	$W_2^2+W_3^1$	$W_1^1+W_2^2+W_3^1$
$\frac{1}{27}$	$W_2^2+W_3^1$	$W_1^1+W_2^2+W_3^1$	$W_1^2+W_2^2+W_3^1$
$\frac{1}{27}$	$W_1^1+W_2^1+W_3^2$	$W_1^2+W_2^1+W_3^2$	$W_2^1+W_3^2$
$\frac{1}{27}$	$W_1^2+W_2^1+W_3^2$	$W_2^1+W_3^2$	$W_1^1+W_2^1+W_3^2$
$\frac{1}{27}$	$W_2^1+W_3^2$	$W_1^1+W_2^1+W_3^2$	$W_1^2+W_2^1+W_3^2$
$\frac{1}{27}$	$W_1^1+W_2^2+W_3^2$	$W_1^2+W_2^2+W_3^2$	$W_2^2+W_3^2$
$\frac{1}{27}$	$W_1^2+W_2^2+W_3^2$	$W_2^2+W_3^2$	$W_1^1+W_2^2+W_3^2$
$\frac{1}{27}$	$W_2^2+W_3^2$	$W_1^1+W_2^2+W_3^2$	$W_1^2+W_2^2+W_3^2$

2) Proof of Privacy: The scheme is constructed in such a way that any given database always receives a query out of the set of queries given by, $\{\phi, \{W_i^\ell, i \in \{1,\ldots,K\}, \ell \in \{1,\ldots,N-1\}\}, \{W_{i_1}^{\ell_1} + \ldots + W_{i_t}^{\ell_t}, \text{ for } i_1,\ldots,i_t \in \{1,\ldots,K\}, \ell_1,\ldots,\ell_t \in \{1,\ldots,N-1\}, t \in \{2,\ldots,K\}\}\}$ with equal probability $\frac{1}{N^K}$ irrespective of the message requirement. Therefore, from a given database's perspective, the a posteriori probability of the user needing message j, upon receiving a query q from a user can be calculated by,

$$P(\theta=i|Q=q) = \frac{P(Q=q|\theta=i)P(\theta=i)}{\sum_{j=1}^{K} P(Q=q|\theta=j)P(\theta=j)} \quad (91)$$

$$= \frac{\frac{1}{N^K}P(\theta=i)}{\sum_{j=1}^{K}\frac{1}{N^K}P(\theta=j)} \quad (92)$$

$$= P(\theta=i) \quad (93)$$

which ensures that this scheme is private, since it implies that θ and Q are independent.

E. Example of Semantic PIR Scheme 2

1) Example 3: $N=3, K=3$, $L_1=400$ Bits, $L_2=300$ Bits and $L_3=100$ Bits: Table V shows the query options that the user may use with probability $\frac{1}{27}$, to download W_1. Whenever a set of queries for the three databases is chosen with probability $\frac{1}{27}$, the required message is retrieved by subtracting the smaller sum from the larger sums, guaranteeing correctness.

The queries in the first block have zero side information, and retrieve the $N-1=2$ parts of W_1 using $N-1$ different databases. The second block uses W_2^1 as side information, and retrieve the two parts of W_1 (in terms of a sum of itself and side information) using the other two databases. The same procedure is carried out in blocks 3, 4 and 5, with W_2^1 replaced by W_2^2, W_3^1 and W_3^2. Last four blocks of Table V use $W_2^i+W_3^j$ for $j \in 1,2$ as side information and use sums of three elements ($W_1^k+W_2^i+W_3^j$ for $k=1,2$) to retrieve the two parts of W_1.

The rate achieved by this scheme when retrieving W_1 is,

$$R_1 = \frac{L_1}{\frac{1}{27}\left(3L_1 + 18(\frac{L_1}{2}\times 2 + \frac{L_2}{2}) + 6(\frac{L_1}{2}\times 2 + \frac{L_3}{2})\right)} \quad (94)$$

$$= \frac{L_1}{\frac{1}{27}(27L_1 + 9L_2 + 3L_3)} \quad (95)$$

$$= \frac{400}{\frac{1}{27}(27\times 400 + 9\times 300 + 3\times 100)} = \frac{36}{46} \quad (96)$$

The rate achieved by this scheme when retrieving W_2 is,

$$R_2 = \frac{L_2}{\frac{1}{27}\left(3L_2 + 18\times 3\times \frac{L_1}{2} + 6\times(L_2 + \frac{L_3}{2})\right)} \quad (97)$$

$$=\frac{L_2}{\frac{1}{27}(27L_1+9L_2+3L_3)} \tag{98}$$

$$=\frac{300}{\frac{1}{27}(27\times 400+9\times 300+3\times 100)}=\frac{27}{46} \tag{99}$$

The rate achieved by this scheme when retrieving W_3 is,

$$R_3=\frac{L_3}{\frac{1}{27}\left(3L_3+18\times 3\times\frac{L_1}{2}+6\times 3\times\frac{L_2}{2}\right)} \tag{100}$$

$$=\frac{L_3}{\frac{1}{27}(27L_1+9L_2+3L_3)} \tag{101}$$

$$=\frac{100}{\frac{1}{27}(27\times 400+9\times 300+3\times 100)}=\frac{9}{46} \tag{102}$$

The overall message retrieval rate for this example is,

$$R=\frac{\mathbb{E}[L]}{\mathbb{E}[D]}=p_1\frac{L_1}{D}+p_2\frac{L_2}{D}+p_3\frac{L_3}{D} \tag{103}$$

$$=p_1R_1+p_2R_2+p_3R_3=\frac{36}{46}p_1+\frac{27}{46}p_2+\frac{9}{46}p_3 \tag{104}$$

This matches the semantic PIR capacity expression in Theorem 1,

$$C=\left(\frac{L_1}{\mathbb{E}[L]}+\frac{1}{N}\frac{L_2}{\mathbb{E}[L]}+\frac{1}{N^2}\frac{L_3}{\mathbb{E}[L]}\right)^{-1} \tag{105}$$

$$=(400p_1+300p_2+100p_3)\left(400+\frac{300}{3}+\frac{100}{9}\right)^{-1} \tag{106}$$

$$=\frac{36}{46}p_1+\frac{27}{46}p_2+\frac{9}{46}p_3 \tag{107}$$

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

$$C=\left(1+\frac{1}{N}+\frac{1}{N^2}\right)^{-1}=\left(1+\frac{1}{3}+\frac{1}{9}\right)^{-1}=\frac{9}{13} \tag{108}$$

The semantic PIR capacity in (107) exceeds the classical PIR capacity in (108) when

$$\frac{36}{46}p_1+\frac{27}{46}p_2+\frac{9}{46}p_3>\frac{9}{13} \tag{109}$$

which is equivalent to

$$p_1+\frac{2}{3}p_2>\frac{11}{13} \tag{110}$$

Remark 10: We note again that the rate calculation presented here for the semantic PIR capacity takes into consideration the zero-padding needed to be added to the shorter message block in order to perform bit-wise message addition for any query realization. The classical PIR capacity expression in (108) assumes that all messages are of equal size and hence the extra zero-padding is not reflected in that expression. Hence, the actual rate of classical PIR scheme is indeed less than the reported PIR capacity if the messages are of unequal size.

Remark 11: The second scheme presented above is an extension to more than two databases of the path-based scheme presented in [47]. It is also similar to the scheme provided in [60], except for the fact that the above scheme has database symmetry as opposed to the scheme presented in [60].

V. Converse Proof

In this section, we present the converse proof for Theorem 1. This proof is a slight modification of the converse proof presented in [3]. The central intuition of our proof is the fact that the expected length of the answer string generated by a given database should remain the same, irrespective of the identity of the desired message as a consequence of the privacy constraint. The major difference of our proof compared to [3] is the handling of the non-equal message sizes.

We begin the proof of Theorem 1 by the definition of message retrieval rate,

$$R=\frac{\mathbb{E}[L]}{\mathbb{E}[D]} \tag{111}$$

We choose some permutation $\{i_1,\ldots,i_K\}$ as an arbitrary order of the messages. The denominator of (111) can be expanded as follows,

$$\mathbb{E}[D]=\sum_{i=1}^{K}p_i(H(A_1^{[i]})+\cdots+H(A_N^{[i]})) \tag{112}$$

$$=H(A_1^{[i_1]})+\cdots+H(A_N^{[i_1]}) \tag{113}$$

Following the same steps in the converse proof of [3] $\mathbb{E}[D]$ can be lower bounded as,

$$\mathbb{E}[D]\geq L_{i_1}+H(A_n^{[i_2]}|Q_n^{[i_2]},W_{i_1})\quad n=1,\ldots,N \tag{114}$$

By summing all N inequalities corresponding to (114) and repeating the previous steps for W_{i_2} (with conditioning on W_{i_1}) leads to,

$$N\mathbb{E}[D]\geq NL_{i_1}+L_{i_2}+H(A_n^{[i_3]}|Q_n^{[i_3]},W_{i_1},W_{i_2}) \tag{115}$$

for $n=1,\ldots,N$. By summing the corresponding inequalities and continuing with the same procedure for $W_{i_3},\ldots,W_{i_K}$ yields,

$$N^{K-1}\mathbb{E}[D]\geq N^{K-1}L_{i_1}+N^{K-2}L_{i_2}+\cdots+NL_{i_{K-1}}$$
$$+I(W_{i_K};A_1^{[i_K]},\ldots,A_N^{[i_K]}|Q_1^{[i_K]},\ldots,Q_N^{[i_K]},W_{i_1},\ldots,W_{i_{K-1}}) \tag{116}$$

and therefore, we have,

$$\mathbb{E}[D]\geq L_{i_1}+\frac{1}{N}L_{i_2}+\cdots+\frac{1}{N^{K-2}}L_{i_{K-1}}+\frac{1}{N^{K-1}}L_{i_K} \tag{117}$$

which further gives,

$$R\leq\left(\frac{L_{i_1}}{\mathbb{E}[L]}+\frac{1}{N}\frac{L_{i_2}}{\mathbb{E}[L]}+\cdots+\frac{1}{N^{K-1}}\frac{L_{i_K}}{\mathbb{E}[L]}\right)^{-1} \tag{118}$$

The upper bound in (118) holds for any permutation $\{i_1,\ldots,i_K\}$, hence, the tightest upper bound can be obtained by minimizing over all permutations.[7] Consequently,

$$R\leq\min_{\{i_1,\ldots,i_K\}}\left(\frac{L_{i_1}}{\mathbb{E}[L]}+\frac{1}{N}\frac{L_{i_2}}{\mathbb{E}[L]}+\cdots+\frac{1}{N^{K-1}}\frac{L_{i_K}}{\mathbb{E}[L]}\right)^{-1} \tag{119}$$

[7]Note that the order does not matter in the case of equal message lengths in [3].

Since the messages are ordered such that $L_1 \geq L_2 \geq \cdots \geq L_K$, the minimum upper bound is attained at $\{i_1, \ldots, i_K\} = \{1, \ldots, K\}$ as it gives the largest number to the largest coefficient in the lower bound on the download cost. Thus,

$$R \leq \left(\frac{L_1}{\mathbb{E}[L]} + \frac{1}{N} \frac{L_2}{\mathbb{E}[L]} + \cdots + \frac{1}{N^{K-1}} \frac{L_K}{\mathbb{E}[L]} \right)^{-1} \tag{120}$$

completing the converse proof.

VI. Extensions of Semantic PIR

A. Semantic PIR From MDS-Coded Databases

In this section, we present a complete characterization of the capacity of semantic PIR from MDS-coded databases, along with an optimal scheme. The optimal scheme is an extension of the scheme presented in [9]. We consider an (N, M) MDS coded distributed storage system containing K independent messages. The messages are allowed to have different semantics (lengths and prior probabilities). Each message W_i is represented as a matrix in $\mathbb{F}_q^{L_i \times M}$, where the elements of the matrix are uniformly and randomly chosen from $\mathbb{F}_q$. The generator matrix of the (N, M) code is $H = [h_1, \ldots, h_N]$, where $h_i \in \mathbb{F}_q^M$, $i \in [N]$. The MDS property implies that any combination of up to M columns of H is linearly independent.

Let the jth row of W_i be denoted by $W_j^{[i]}$. Each database n, $n \in [N]$ stores $W_j^{[i]} h_n$ for $j \in [L_i]$, $i \in [K]$. The objective is to download a required message without revealing its index to any of the databases. In order to retrieve W_i, user sends query $Q_n^{[i]}$ to database n, $n \in [N]$ and receives the answer $A_n^{[i]}$ which is a deterministic function of the contents of the database and $Q_n^{[i]}$. The correctness and privacy conditions are the same as (5) and (6) respectively, and the rate is calculated by,

$$R = \frac{M\mathbb{E}[L]}{\mathbb{E}[D]} \tag{121}$$

Theorem 2 gives the exact PIR capacity for the semantic PIR problem.

Theorem 2: The capacity of semantic PIR with (N, M) MDS-coded databases with N databases, K messages, message sizes ML_i (arranged as $L_1 \geq L_2 \geq \ldots \geq L_K$) and prior probabilities p_i is given by,

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \left(\frac{M}{N}\right) \frac{L_2}{\mathbb{E}[L]} + \cdots + \left(\frac{M}{N}\right)^{K-1} \frac{L_K}{\mathbb{E}[L]} \right)^{-1} \tag{122}$$

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

The achievable scheme is an extension to the first scheme presented in Section IV-A. The steps of the achievable scheme are as follows. Assume that the user requires to download W_j.

1) *Message Indexing:* Assign indices to messages in the descending order message sizes, i.e., $L_1 \geq L_2 \geq \ldots \geq L_K$. Permute the rows of all messages randomly and independently, privately from the databases.
2) *Single Blocks:* Using (130), download υ_j different coded bits of W_j from each database. Download υ_i coded bits of W_i, $i \neq j$ from each database such that the coded bits of M different databases correspond to the same row of W_i. This is required to decode the rows of W_i that are used as side information. Therefore, $N\upsilon_i$ coded bits of W_i, $i \neq j$ are downloaded in this step, that belong to $\frac{N\upsilon_i}{M}$ different rows of W_i.
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 (singles 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, $i \neq j$. Download $\left(\frac{N}{M} - 1\right) \min\{\upsilon_i, \upsilon_j\}$ 2-sums of the form $(W_{r_n}^{[j]} + W_{s_n}^{[i]})h_n$ from database n, $n \in [N]$ where $W_{r_n}^{[j]}$ are new rows of W_j and $W_{s_n}^{[i]}$ are already decoded rows of W_i in the previous step. Note that the set of M databases that were used to decode $W_{s_n}^{[i]}$ in the previous step does not include database n. The second block of 2-sums contains coded bits corresponding to W_{i_1} and W_{i_2}, where $i_1 \neq i_2 \neq j$. Download $\left(\frac{N}{M} - 1\right) \min\{\upsilon_i, \upsilon_j\}$ 2-sums of the form $(W_{t_n}^{[i_1]} + W_{v_n}^{[i_2]})h_n$ from database n, $n \in [N]$ where $W_{t_n}^{[i_1]}$ and $W_{v_n}^{[i_2]}$ are new rows of W_{i_1} and W_{i_2}. Note that coded bits corresponding to the same pair of rows (t_n, v_n) needs to be downloaded from M different databases in order to correctly decode the side information $W_{t_n}^{[i_1]} + W_{v_n}^{[i_2]}$. Thus, the second block contains $N\left(\frac{N}{M} - 1\right) \min\{i_1, i_2\}$ coded 2-sums corresponding to W_{i_1} and W_{i_2} belonging to $\frac{N\left(\frac{N}{M}-1\right)\min\{i_1,i_2\}}{M}$ different pairs of rows.
4) *Sums of ℓ Elements:* There are two types of blocks similar to sums of two. The first block contains queries of the form $(W_r^{[j]} + W_{r_1}^{[i_1]} + \ldots + W_{r_{\ell-1}}^{[i_{\ell-1}]})h_n$, $i_1 \neq , \ldots, \neq i_{\ell-1} \neq j$, where $W_r^{[j]}$ is a new row of W_j and $W_{r_1}^{[i_1]} + \ldots + W_{r_{\ell-1}}^{[i_{\ell-1}]}$ is an already decoded $(\ell-1)$-sum from the previous step. For a given $(\ell-1)$-tuple $(i_1, \ldots, i_{\ell-1})$, download $\left(\frac{N}{M} - 1\right)^{\ell-1} \upsilon_{\min\{j, i_1, \ldots, i_{\ell-1}\}}$ such ℓ-sums from each database. The second block contains queries of the form $(W_{t_1}^{[i_1]} + \ldots + W_{t_\ell}^{[i_\ell]})h_n$, $i_1 \neq \cdots \neq i_\ell \neq j$, where $W_{t_1}^{[i_1]}, \ldots, W_{t_\ell}^{[i_\ell]}$ are new rows of $W_{i_1}, \ldots, W_{i_\ell}$. Download $\left(\frac{N}{M} - 1\right)^{\ell-1} \upsilon_{\min\{i_1, \ldots, i_\ell\}}$ such ℓ-sums from each database such that the coded bits corresponding to a given ℓ-tuple of rows $(r_1, \ldots, r_\ell)$ is downloaded from M different databases. A total of $N\left(\frac{N}{M} - 1\right)^{\ell-1} \upsilon_{\min\{i_1, \ldots, i_\ell\}}$ coded bits of this form will be downloaded corresponding to $\frac{\left(\frac{N}{M}-1\right)^{\ell-1} \upsilon_{\min\{i_1, \ldots, i_\ell\}}}{M}$ different ℓ-tuples of rows of $W_{i_1}, \ldots, W_{i_\ell}$.
5) *Repeat the Process:* up to sums of K elements.
6) *Query Repetition:* To decode each row of W_j, repeat the above process M 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, $i \in \{1, \ldots, K\}$, $i \neq j$ in each repetition. The M different linear combinations of each row of W_j allow us to correctly decode W_j.

The achievable rate of the above scheme is calculated as follows. First, note that the download cost remains the same irrespective of the message requirement in order to guarantee privacy. Therefore, the $\mathbb{E}[D]$ term in (121) is calculated by summing the number of downloads in each step of the scheme. Within one round of queries, there are $\sum_{i=1}^{K} N\upsilon_i$ singletons and $N\left(\frac{N}{M}-1\right)^{\ell-1}\sum_{i=\ell}^{K}\binom{i-1}{\ell-1}\upsilon_i$ sums of ℓ-elements. Therefore,

$$\frac{\mathbb{E}[D]}{M} = \sum_{i=1}^{K} N\upsilon_i + \sum_{\ell=2}^{K}\sum_{i=\ell}^{K} N\left(\frac{N}{M}-1\right)^{\ell-1}\upsilon_i\binom{i-1}{\ell-1} \tag{123}$$

$$= N\left[\sum_{i=1}^{K}\upsilon_i + \sum_{\ell=2}^{K}\upsilon_\ell\sum_{i=2}^{\ell}\left(\frac{N}{M}-1\right)^{i-1}\binom{\ell-1}{i-1}\right] \tag{124}$$

$$= M\left[\frac{N}{M}\upsilon_1 + \sum_{\ell=2}^{K}\upsilon_\ell\left(\frac{N}{M}\right)^{\ell}\right] = M\sum_{\ell=1}^{K}\left(\frac{N}{M}\right)^{\ell}\upsilon_\ell \tag{125}$$

For the $\mathbb{E}[L]$ term in (121), we sum the number of useful bits downloaded in each step of the scheme. Based on the scheme described above, $N\upsilon_j$ rows of W_j are retrieved as singletons, $N\left(\frac{N}{M}-1\right)^{\ell-1}\binom{j-1}{\ell-1}\upsilon_j$ rows of W_j are retrieved as ℓ-sums with W_j being the shortest message and $N\left(\frac{N}{M}-1\right)^{\ell-1}\binom{i-2}{\ell-2}\upsilon_i$ rows of W_j are retrieved as ℓ-sums with W_i, $i\neq j$ being the shortest message in the sum. Denoting U_j as the total number of useful bits downloaded, the number of rows of W_j retrieved is calculated by,

$$\begin{aligned}\frac{U_j}{M} =& N\upsilon_j + \sum_{\ell=2}^{j} N\left(\frac{N}{M}-1\right)^{\ell-1}\binom{j-1}{\ell-1}\upsilon_j \\ &+ \sum_{\ell=2}^{j}\sum_{i=j+1}^{K} N\left(\frac{N}{M}-1\right)^{\ell-1}\binom{i-2}{\ell-2}\upsilon_i \\ &+ \sum_{\ell=j+1}^{K}\sum_{i=\ell}^{K} N\left(\frac{N}{M}-1\right)^{\ell-1}\binom{i-2}{\ell-2}\upsilon_i \qquad (126)\\ =& N\upsilon_j\sum_{\ell=0}^{j-1}\gamma^\ell\binom{j-1}{\ell} + N\upsilon_{j+1}\gamma\sum_{\ell=0}^{j-1}\gamma^\ell\binom{j-1}{\ell} \\ &+ N\upsilon_{j+2}\gamma\sum_{\ell=0}^{j}\gamma^\ell\binom{j}{\ell} + \cdots + N\upsilon_K\gamma\sum_{\ell=0}^{K-2}\gamma^\ell\binom{K-2}{\ell} \qquad (127)\\ =& M\left[\left(\frac{N^j}{M^j}\right)\upsilon_j + \gamma\sum_{i=j+1}^{K}\left(\frac{N}{M}\right)^{i-1}\upsilon_i\right] \qquad (128)\end{aligned}$$

where $\gamma=\frac{N}{M}-1$. Thus, the subpacketization of W_j is defined as $\frac{U_j}{M}$, 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, $j\in\{1,\ldots,K\}$ have to be a common multiple of their own subpacketizations,

$$L_j = \alpha\frac{U_j}{M},\quad j\in\{1,\ldots,K\} \tag{129}$$

for some $\alpha\in\mathbb{N}$. Solving (128) and (129) for $\upsilon_1,\ldots,\upsilon_K$ gives,

$$\begin{bmatrix}\upsilon_1\\ \upsilon_2\\ \vdots\\ \upsilon_K\end{bmatrix} = \frac{1}{M\alpha}\begin{bmatrix}\frac{M}{N} & -\left(\frac{M}{N}\right)^2\gamma & \ldots & -\left(\frac{M}{N}\right)^K\gamma\\ 0 & \left(\frac{M}{N}\right)^2 & \ldots & -\left(\frac{M}{N}\right)^K\gamma\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \ldots & \left(\frac{M}{N}\right)^K\end{bmatrix}\begin{bmatrix}L_1\\ L_2\\ \vdots\\ L_K\end{bmatrix} \tag{130}$$

In order for the values of $\upsilon_i, i\in\{1,\ldots,K\}$ to be integers, this scheme requires each L_i to be a multiple of N^i. Here, α is the greatest common divisor (gcd) of the elements of the vector resulting from multiplying the matrix and the vector on the right side of (130). This allows the shortest subpacketization levels for all messages.

The total and useful numbers of bits downloaded (in (125) and (128), respectively) are both within one subpacketization level. These downloads are repeated α times to download the entire message. Thus, the achievable rate is given by,

$$\begin{aligned}R =& \frac{M\mathbb{E}[L]}{\mathbb{E}[D]} = \frac{M\sum_{i=1}^{K}p_iL_i}{\alpha M^2\sum_{i=1}^{K}\frac{N^i}{M^i}\upsilon_i} \qquad (131)\\ =& \frac{\mathbb{E}[L]}{\alpha M\frac{1}{M\alpha}\sum_{i=1}^{K}\frac{N^i}{M^i}\left[\left(\frac{M^i}{N^i}\right)L_i-\left(\frac{N}{M}-1\right)\sum_{t=i+1}^{K}\left(\frac{M^t}{N^t}\right)L_t\right]} \qquad (132)\\ =& \frac{\mathbb{E}[L]}{\sum_{i=1}^{K}\left[L_i-\left(\frac{N}{M}-1\right)\sum_{t=i+1}^{K}\left(\frac{M^{t-i}}{N^{t-i}}\right)L_t\right]} \qquad (133)\\ =& \frac{\mathbb{E}[L]}{L_1+L_2\left(\frac{M}{N}\right)+\sum_{i=3}^{K}L_i\left[1-\left(1-\frac{M^{i-1}}{N^{i-1}}\right)\right]} \qquad (134)\\ =& \left(\frac{L_1}{\mathbb{E}[L]}+\left(\frac{M}{N}\right)\frac{L_2}{\mathbb{E}[L]}+\cdots+\left(\frac{M}{N}\right)^{K-1}\frac{L_K}{\mathbb{E}[L]}\right)^{-1} \qquad (135)\end{aligned}$$

A given database always receives queries of the same type (i.e., $\left(\frac{N}{M}-1\right)^{\ell-1}\upsilon_{\min\{i_1,\ldots,i_\ell\}}$, $\forall\{i_1,\ldots,i_\ell\}\subset[K]$, ℓ-sums for $\ell\in\{1,\ldots,K\}$) irrespective of the message requirement. According to the query generation procedure, no bit of any message is requested from a given database more than once as a singleton or as an element of a sum. Therefore, a proof similar to what is presented in Section IV-A.2 is used to show that this scheme is private.

The above scheme can be alternatively described using the same ideas presented in Section IV-C. The alternative description is as follows. Database n, $n\in\{1,\ldots,N\}$ contains coded bits corresponding to each row of W_i, $i\in\{1,\ldots,K\}$ given by $W_r^{[i]}h_n$, $r\in\{1,\ldots,L_i\}$. Therefore, each database stores L_i coded bits of W_i, where $L_1\geq L_2\geq\ldots\geq L_K$. Considering the first L_K coded bits of all messages, the classical MDS-coded PIR scheme in [9] is applied as the first step of the scheme. Then, apply the classical coded PIR scheme using the next $L_{K-1}-L_K$ coded bits of messages W_1 to W_{K-1}. In general, in the ℓth step, the classical coded PIR scheme needs to be applied on the $L_{K-\ell+1}-L_{K-\ell+2}$ coded bits of W_1 to $W_{K-\ell+1}$. The complete scheme should be used irrespective of the message requirement.

The alternative description differs from the main description in subpacketization, and in the scheme used within a subpacket as explained in Section IV-C. However, the two descriptions are equivalent when considering the entire retrieval process (all subpackets). The rate achieved by the alternative scheme is given by,

$$R = \mathbb{E}[L] \Bigg/ \Bigg(L_K \left(1 + \frac{M}{N} + \ldots + \frac{M^{K-1}}{N^{K-1}}\right) + (L_{K-1} - L_K)\left(1 + \frac{M}{N} + \ldots + \frac{M^{K-2}}{N^{K-2}}\right) + \ldots + L_1 - L_2 \Bigg) \tag{136}$$

which is the same as (135). A converse proof similar to what is presented in Section V with the ideas of [9] is used to prove an upper bound on the retrieval rate of semantic PIR from MDS-coded databases, which is the same as (135). This proves the capacity expression in (122).

B. Semantic PIR From Colluding Databases

In this section, we present a complete characterization of the capacity of semantic PIR from colluding databases, along with an optimal scheme. This is an extension of the results presented in [4]. We consider K independent messages (W_i, $i \in \{1, \ldots, K\}$) with arbitrary lengths L_i and prior probabilities p_i, stored in N replicated databases. Out of the N databases, any subset up to T databases are allowed to collude. The objective here is to download a user-required message without revealing its index to any T-colluding databases.

Theorem 3: The capacity of semantic PIR from colluding databases, with K messages, message lengths L_i (arranged as $L_1 \geq L_2 \geq \ldots \geq L_K$), prior probabilities p_i and N databases out of which any T are colluding, is given by,

$$C = \left(\frac{L_1}{\mathbb{E}[L]} + \frac{L_2}{\mathbb{E}[L]} \left(\frac{T}{N}\right) + \ldots + \frac{L_K}{\mathbb{E}[L]} \left(\frac{T}{N}\right)^{K-1} \right)^{-1} \tag{137}$$

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

The optimal scheme is an extension of the scheme presented in Section IV-A. The scheme is as follows. Assume that the required message is W_j. Once the messages are indexed based on the decreasing order of lengths, the user needs to generate a set of linear combinations of the message indices given by,

$$x_j = S_j W_j \tag{138}$$

where S_j is a random full rank matrix drawn uniformly and independently from all such matrices in $\mathbb{F}_q^{\ell_j \times \ell_j}$ where ℓ_i is the subpacketization of W_j. For each W_m, $m \neq j$, let m_t denote the number of t-sums in the scheme involving W_m but not W_j. Let $m_{t,j}$ be the number of t-sums in the scheme involving both W_m and W_j.[8] Then, the linear combinations of W_i, $i \in [K]$, $i \neq j$ are generated by,

$$\text{first } (m_1 + m_{2,j}) \text{ bits of } x_i = \text{MDS}_{(m_1+m_{2,j})\times m_1} S_i[(1 : m_1), :] W_i \tag{139}$$

$$\text{next } (m_2 + m_{3,j}) \text{ bits of } x_i = \text{MDS}_{(m_2+m_{3,j})\times m_2} S_i[(m_1 + 1 : m_1 + m_2), :] W_i \tag{140}$$

$$\vdots$$

$$\text{last } (m_{K-1} + m_{K,j}) \text{ bits of } x_i = \text{MDS}_{(m_{K-1}+m_{K,j})\times m_{K-1}} S_i[(\ell_i - m_{K-1} + 1 : \ell_i), :] W_i \tag{141}$$

where S_i, $i \in \{1, \ldots, K\}$ are random full rank matrices of $\mathbb{F}_q^{\ell_i \times \ell_i}$ and $\text{MDS}_{a\times b}$ are globally known generator matrices of (a, b) MDS-codes. The first step of the scheme is to calculate υ_i, $i \in \{1, \ldots, K\}$ using (130) with M replaced by T. Then, download υ_i, $i \in \{1, \ldots, K\}$ bits of each x_i from each database. Next, from each database, download $\left(\frac{N}{T} - 1\right) \upsilon_{\min\{i_1,\ldots,i_t\}}$ t-sums, $t \in \{2, \ldots, K\}$ involving new bits of $x_{i_1}, \ldots, x_{i_t}$, $\forall \{i_1, \ldots, i_t\} \subset \{1, \ldots, K\}$. This completes the scheme.

For a given t-sum of the form $x_{i_1}(\cdot) + \ldots + x_{i_t}(\cdot)$ with $i_1 \geq i_2 \geq \ldots \geq i_t$, which does include any bit of W_j, let the generator matrix corresponding to each element x_{i_k} in the sum be denoted by G_{i_k}. Then, each G_{i_k} must satisfy, $G_{i_k} = \begin{bmatrix} G_{i_{k+1}} \\ \ldots \\ X \end{bmatrix}$, $k \in \{1, \ldots, t\}$ where X denotes the set of extra rows in the larger generator matrix. This is required for interference alignment. The proof of privacy in [4] applies to this scheme as well. The fact that the required message is coded differently, in a non redundant manner, ensures the correctness of the scheme as explained in [4].

The optimal scheme above can be alternatively described as follows. In each database, segment the set of messages into K partitions, such that the first segment contains the first L_K bits of all K messages, the second segment contains the next set of $L_{K_1} - L_K$ bits of messages W_1 to W_{K-1} and so on. Then, apply the classical colluded PIR scheme in [4] to the 1) the first segment with K messages, 2) the second segment with $K-1$ messages, 3) the third segment with $K-2$ messages, and so on. Make sure that the complete scheme is used irrespective of the desired message for privacy. The achievable rate of the scheme is equal to the capacity in (137). The converse is proved using similar ideas provided in the converse proofs of Section V and [4].

VII. Conclusion and Discussion

In this work, we introduced the problem of semantic PIR. In this problem, the stored messages are allowed to have non-uniform popularities, which is captured via an a priori probability distribution (p_i, $i \in [K]$), and heterogeneous sizes (L_i, $i \in [K]$). We derived the exact semantic PIR capacity as a function of $\{L_i\}_{i=1}^K$ and the expected message size $\mathbb{E}[L]$. The result implies that the semantic PIR capacity is equal to the classical PIR capacity if all messages have equal sizes $L_i = L$ for all $i \in [K]$. We derived a necessary and sufficient

[8] The values of m_t and $m_{t,j}$ for $t \in \{1, \ldots, K\}$ are immediate from the steps of scheme which are described later. These values do not depend on the linear combinations.

condition for the semantic PIR capacity to exceed the classical PIR capacity. In particular, we showed that if the longer messages are retrieved more often, there is a strict retrieval rate gain from exploiting the message semantics.[9] Furthermore, we proved that for all message sizes and priors, the semantic PIR capacity exceeds the achievable rate of classical PIR with zero-padding, which zero-pads all messages to equalize their sizes.

To that end, we proposed two achievable schemes for achieving the semantic PIR capacity. The first one has a deterministic query structure. We have proposed a systematic way of calculating the needed subpacketization levels for the messages. We also provided an alternative description to this scheme which implements the classical PIR scheme in a segmented manner. The similarities and differences between the two descriptions were also discussed. The second scheme has a stochastic query structure, where the user picks one query structure at random from an ensemble of structures. The first scheme has the advantage of having a fixed download cost for all messages for all query structures unlike the stochastic scheme, which has the same expected download cost. Nevertheless, the first scheme suffers from exponential subpacketization levels in contrast to the linear counterpart in the stochastic scheme. We derived a matching converse that extends the converse scheme of [3] to take into account the heterogeneous message sizes and prior probabilities. Finally, the extensions of semantic PIR to coded databases and colluding databases were analyzed separately, where the complete characterizations of the capacities of the two cases were presented along with the corresponding optimal schemes.

REFERENCES

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

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

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

[4] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.

[5] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. E. Rouayheb, "Private information retrieval schemes for coded data with arbitrary collusion patterns," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1908–1912.

[6] Z. Jia, H. Sun, and S. A. Jafar, "The capacity of private information retrieval with disjoint colluding sets," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.

[7] X. Yao, N. Liu, and W. Kang, "The capacity of private information retrieval under arbitrary collusion patterns," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1041–1046.

[8] R. Bitar and S. E. Rouayheb, "Staircase-PIR: Universally robust private information retrieval," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[9] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

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

[9]This does not necessarily mean that $p_1 \geq p_2 \geq \ldots \geq p_K$. It essentially means that the $\mathbb{E}[L]$ should be large enough such that (11) is satisfied.

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

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

[13] Y. Zhang and G. Ge, "A general private information retrieval scheme for MDS coded databases with colluding servers," *Des., Codes Cryptogr.*, vol. 87, no. 11, pp. 2611–2623, Nov. 2019.

[14] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij–Hollanti," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[15] L. Holzbaur, R. Freij-Hollanti, J. Li, and C. Hollanti, "Towards the capacity of private information retrieval from coded and colluding servers," 2019, *arXiv:1903.12552*.

[16] K. Banawan and S. Ulukus, "The capacity of private information retrieval from Byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[17] 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. Inf. Theory*, vol. 65, no. 6, pp. 3898–3906, Jun. 2019.

[18] X. Yao, N. Liu, and W. Kang, "The capacity of multi-round private information retrieval from Byzantine databases," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2124–2128.

[19] R. Tandon, M. Abdul-Wahid, F. Almoualem, and D. Kumar, "PIR from storage constrained databases–coded caching meets PIR," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

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

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

[22] Y.-P. Wei, B. Arasli, K. Banawan, and S. Ulukus, "The capacity of private information retrieval from decentralized uncoded caching databases," *Information*, vol. 10, no. 12, p. 372, Nov. 2019.

[23] K. Banawan, B. Arasli, and S. Ulukus, "Improved storage for efficient private information retrieval," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.

[24] C. Tian, "On the storage cost of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7539–7549, Dec. 2020.

[25] N. Raviv and I. Tamot, "Private information retrieval is graph based replication systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1739–1743.

[26] K. Banawan and S. Ulukus, "Private information retrieval from non-replicated databases," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1272–1276.

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

[28] Y. Zhang and G. Ge, "Private information retrieval from MDS coded databases with colluding servers under several variant models," 2017, *arXiv:1705.03186*.

[29] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[30] Q. Wang, H. Sun, and M. Skoglund, "Symmetric private information retrieval with mismatched coded messages and randomness," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 365–369.

[31] Q. Wang and M. Skoglund, "Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5160–5175, Aug. 2019.

[32] Z. Wang, K. Banawan, and S. Ulukus, "Private set intersection: A multi-message symmetric private information retrieval perspective," 2020, *arXiv:1912.13501*.

[33] S. Kadhe, B. Garcia, A. Heidarzadeh, S. E. Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2032–2043, Apr. 2020.

[34] Z. Chen, Z. Wang, and S. Jafar, "The capacity of T-private information retrieval with private side information," 2017, *arXiv:1709.03022*.

[35] Y.-P. Wei, K. Banawan, and S. Ulukus, "The capacity of private information retrieval with partially known private side information," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8222–8231, Dec. 2019.

[36] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[37] A. Heidarzadeh, B. Garcia, S. Kadhe, S. El Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput.*, Oct. 2018, pp. 180–187.

[38] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput.*, Oct. 2018, pp. 173–179.

[39] S. Li and M. Gastpar, "Converse for multi-server single-message PIR with side information," in *Proc. 54th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2020, pp. 1–6.

[40] Y.-P. Wei and S. Ulukus, "The capacity of private information retrieval with private side information under storage constraints," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2023–2031, Apr. 2020.

[41] R. Tandon, "The capacity of cache aided private information retrieval," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 1078–1082.

[42] M. Kim, H. Yang, and J. Lee, "Cache-aided private information retrieval," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, Oct. 2017, pp. 398–402.

[43] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3215–3232, May 2019.

[44] Y.-P. Wei, K. Banawan, and S. Ulukus, "Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1126–1139, Jun. 2018.

[45] 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. Commun.*, vol. 67, no. 7, pp. 4900–4912, Jul. 2019.

[46] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2999–3012, 2020.

[47] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1262–1266.

[48] I. Samy, M. A. Attia, R. Tandon, and L. Lazos, "On the capacity of latent variable private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 1907–1912.

[49] H. Sun and S. A. Jafar, "The capacity of private computation," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3880–3897, Jun. 2019.

[50] M. Mirmohseni and M. A. Maddah-Ali, "Private function retrieval," in *Proc. Iran Workshop Commun. Inf. Theory (IWCIT)*, Apr. 2018, pp. 1–6.

[51] Z. Chen, Z. Wang, and S. Jafar, "The asymptotic capacity of private search," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 2122–2126.

[52] Q. Wang and M. Skoglund, "On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3183–3197, May 2019.

[53] Q. Wang, H. Sun, and M. Skoglund, "The capacity of private information retrieval with eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3198–3214, May 2019.

[54] K. Banawan and S. Ulukus, "Private information retrieval through wiretap channel II: Privacy meets security," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4129–4149, Jul. 2020.

[55] H. Yang, W. Shin, and J. Lee, "Private information retrieval for secure distributed storage systems," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 12, pp. 2953–2964, Dec. 2018.

[56] Z. Jia, H. Sun, and S. Jafar, "Cross subspace alignment and the asymptotic capacity of X-secure T-private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5783–5798, Sep. 2019.

[57] R. G. L. D'Oliveira and S. E. Rouayheb, "One-shot PIR: Refinement and lifting," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2443–2455, Apr. 2020.

[58] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 2920–2932, Dec. 2017.

[59] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.

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

[61] J. Xu and Z. Zhang, "Building capacity-achieving PIR schemes with optimal sub-packetization over small fields," in *Proc. ISIT*, Jun. 2018, pp. 1749–1753.

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

[63] K. Banawan and S. Ulukus, "Noisy private information retrieval: On separability of channel coding and information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8232–8249, Dec. 2019.

[64] R. Tajeddine, A. Wachter-Zeh, and C. Hollanti, "Private information retrieval over random linear networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 790–799, 2020.

Sajani Vithana (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from the University of Peradeniya, Sri Lanka, in 2017. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. Her research interests include information theory, private information retrieval, and machine learning.

Karim Banawan (Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) in electrical engineering from Alexandria University, Alexandria, Egypt, in 2008 and 2012, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 2017 and 2018, respectively, with his Ph.D. thesis on private information retrieval and security in networks.

In 2019, he joined the Department of Electrical Engineering, Alexandria University, as an Assistant Professor. His research interests include information theory, wireless communications, physical layer security, private information retrieval, and applications of machine learning in wireless networks. He was a recipient of the Distinguished Dissertation Fellowship from the Department of Electrical and Computer Engineering, University of Maryland, for his Ph.D. thesis work.

Sennur Ulukus (Fellow, IEEE) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University, and the Ph.D. degree in electrical and computer engineering from WINLAB, Rutgers University.

She is the Anthony Ephremides Professor in information sciences and systems with the Department of Electrical and Computer Engineering, University of Maryland, College Park, where she also holds a joint appointment with the Institute for Systems Research (ISR). Prior to joining UMD, she was a Senior Technical Staff Member at AT&T Labs-Research. Her research interests are in information theory, wireless communications, machine learning, signal processing and networks, with recent focus on private information retrieval, age of information, group testing, distributed coded computing, machine learning for wireless, energy harvesting communications, physical layer security, and wireless energy and information transfer.

Dr. Ulukus is a Distinguished Scholar-Teacher with the University of Maryland. She received the 2003 IEEE Marconi Prize Paper Award in Wireless Communications, the 2019 IEEE Communications Society Best Tutorial Paper Award, the 2020 IEEE Communications Society Women in Communications Engineering (WICE) Outstanding Achievement Award, the 2020 IEEE Communications Society Technical Committee on Green Communications and Computing (TCGCC) Distinguished Technical Achievement Recognition Award, the 2005 NSF CAREER Award, the 2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Outstanding Teaching Award. She was a Distinguished Lecturer of the IEEE Information Theory Society from 2018 to 2019. She was the TPC Co-Chair of 2019 IEEE ITW, 2017 IEEE ISIT, 2016 IEEE GLOBECOM, 2014 IEEE PIMRC, and 2011 IEEE CTW. She is the TPC Chair of 2021 IEEE GLOBECOM. She was an Area Editor of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING from 2016 to 2020, an Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS-SERIES ON GREEN COMMUNICATIONS AND NETWORKING from 2015 to 2016, an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2007 to 2010, and an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS from 2003 to 2007. She was a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in 2008, 2015, and 2021, *Journal of Communications and Networks* in 2012, and the IEEE TRANSACTIONS ON INFORMATION THEORY in 2011. She has been an Area Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2019. She has been a Senior Editor of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING since 2020.