# Private Information Retrieval with Partially Known Private Side Information

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

***Abstract*—We consider the problem of private information retrieval (PIR) of a single message out of $K$ messages from $N$ replicated and non-colluding databases where a cache-enabled user of cache-size $M$ messages possesses side information in the form of full messages that are partially known to the databases. In this model, the user and the databases engage in a two-phase scheme, namely, the prefetching phase and the retrieval phase. In the prefetching phase, the user receives $m_n$ full messages from the $n$th database, under the cache memory size constraint $\sum_{n=1}^{N} m_n \leq M$. In the retrieval phase, the user wishes to retrieve a message such that no individual database learns anything about the identity of the desired message. In addition, the identities of the side information messages that the user did not prefetch from a database must remain private against that database. Since the side information provided by each database in the prefetching phase is known by the providing database and the side information must be kept private against the remaining databases, we coin this model as *partially known private side information*. We characterize the capacity of the PIR with partially known private side information to be $C = \left(1+\frac{1}{N}+\cdots+\frac{1}{N^{K-M-1}}\right)^{-1} = \frac{1-\frac{1}{N}}{1-(\frac{1}{N})^{K-M}}$, which is the same if none of the databases knows any of the prefetched side information. Thus, our result implies that there is no loss in using the same databases for both prefetching and retrieval phases.**

## I. Introduction

Private information retrieval (PIR) is a canonical problem to study privacy issues that arise when information is downloaded (retrieved) from public databases. Since its first formulation by Chor et al. in [1], the PIR problem has become a central research topic in the computer science literature, see e.g., [2], [3]. In the classical setting of PIR in [1], a user wishes to retrieve a single message out of $K$ messages replicated across $N$ non-communicating databases without leaking any information about the identity of the retrieved message. Trivially, the user can download the entire database, but this retrieval strategy is highly inefficient. The efficiency of a PIR scheme is measured by the normalized download cost. The goal of the PIR problem is to devise the most efficient retrieval strategy under the privacy and decodability constraints.

The PIR problem has received attention in recent years in the information and coding theory literatures, see e.g., [4]–[9]. In the leading work of Sun-Jafar [10], the classical PIR problem is re-formulated to conform with the conventional information-theoretic arguments, and the notion of PIR capacity is introduced, which is defined as the supremum of retrieval rates over all achievable retrieval schemes. Reference [10] characterizes the capacity of the classical PIR model to be $C = \left(1+\frac{1}{N}+\cdots+\frac{1}{N^{K-1}}\right)^{-1}$. Following the work of Sun-Jafar [10], the capacity of many interesting variants of the classical PIR model have been investigated [11]–[29].

In this paper, we consider the problem of PIR with partially known private side information. Our work is most closely related to [26]–[29][1]. These works investigate the PIR problem when the user (retriever) possesses some form of side information about the contents of the databases. However, the models of [26]–[29] differ in three important aspects, namely, 1) the structure of the side information, 2) the presence or absence of privacy constraints on the side information, and 3) the databases' awareness of the side information at its initial acquisition. Specifically, reference [26] studies the case that the user caches $rLK$ bits in the form of any arbitrary function of the $K$ messages, where $L$ is the message size, and $0 \leq r \leq 1$ is the caching ratio. Reference [26] assumes that the cache content is perfectly known by all the databases. References [27], [28] study the other extreme when the databases are completely unaware of the side information at its initial acquisition. References [27] and [28] differ in terms of the structure of the cached content: [27] considers the case where $M$ full messages are cached, and [28] considers the case where a random $r$ fraction of the symbols of each of $K$ messages is cached. In [28], there is no privacy constraint on the cached content. Reference [27] considers another model where the cached content (in the form of full messages) which is unknown to the databases at the time of initial prefetching, must remain unknown throughout the PIR. The exact capacity for this problem is settled in [29] to be $C = \left(1+\frac{1}{N}+\cdots+\frac{1}{N^{K-M-1}}\right)^{-1}$.

Here, we take a deeper look at the issue of *awareness* or otherwise *unawareness* of the databases about the cached content *at its initial acquisition*. We first note that it is practically challenging to make the side information completely unknown to the databases at its initial acquisition as assumed in [27]–[29]. We also note that the other extreme of the

This work was supported by NSF Grants CNS 13-14733, CCF 14-22111, CNS 15-26608 and CCF 17-13977.

[1]A parallel line of work that studies privacy issues of requests and side information in index coding based broadcast systems can be found in [30], [31].

problem, where the databases are fully aware of the cached content [26], is discouraging as the user cannot benefit from the cached side information. Therefore, a natural model is to use the databases for both prefetching and retrieval phases, such that the databases gain partial knowledge about the side information available to the user, which makes it possible for the user to exploit the remaining side information that is unknown to each individual database to reduce the download cost during the retrieval process.

We investigate the PIR problem when the user and the databases engage in a two-phase scheme, namely, prefetching phase and retrieval phase. In the prefetching phase, the user caches $m_n$ full messages out of the $K$ messages from the $n$th database under a total cache memory size constraint $\sum_{n=1}^{N} m_n \leq M$. Hence, each database has a *partial knowledge* about the side information possessed by the user. In the retrieval phase, the user wants to retrieve a message without leaking any information to any individual database about the desired message or the remaining side information messages that are unknown to each database. We first derive a general lower bound for the normalized download cost that is independent of the prefetching strategy. Then, we prove that this bound is attainable using two achievable schemes. The first achievable scheme, which is proposed in [29] for completely unknown side information, is a valid achievable scheme for our problem with partially known side information for any prefetching strategy.[2] We provide a second achievable scheme for the case of uniform prefetching, i.e., $m_n = \frac{M}{N} \in \mathbb{N}$, which requires smaller sub-packetization and smaller field size for realizing MDS codes. We prove that the exact capacity of this problem is $C = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-M-1}}\right)^{-1}$. Surprisingly, this is the same capacity expression for the PIR problem when the databases are completely unaware of the side information possessed by the user as found in [29]. Thus, our result implies that there is no loss in the capacity if the same databases are employed in both prefetching and retrieval phases.

## II. System Model

We consider a classic PIR problem with $K$ independent messages $W_1, \ldots, W_K$, where each message is of length $L$,

$$H(W_1) = \cdots = H(W_K) = L, \tag{1}$$

$$H(W_1, \ldots, W_K) = H(W_1) + \cdots + H(W_K). \tag{2}$$

There are $N$ non-communicating databases, and each database stores all the $K$ messages. The user (retriever) has a local cache memory which can store up to $M$ messages.

There are two phases: a *prefetching phase* and a *retrieval phase*. In the prefetching phase, $\forall n \in [N]$, where $[N] = \{1, 2, \ldots, N\}$, the user caches $m_n$ out of total $K$ messages from the $n$th database. We denote the indices of the cached messages from the $n$th database as $\mathbb{H}_n$. Therefore, $|\mathbb{H}_n| = m_n$. We denote the indices of all cached messages as $\mathbb{H}$, $\mathbb{H} = \bigcup_{n=1}^{N} \mathbb{H}_n$, where $\mathbb{H}_{n_1} \cap \mathbb{H}_{n_2} = \emptyset$, if $n_1 \neq n_2$. Due to the cache memory size constraint, we require

$$|\mathbb{H}| = \sum_{n=1}^{N} m_n \leq M. \tag{3}$$

Since the user caches $m_n$ messages from the $n$th database, $\mathbb{H}_n$ is known to the $n$th database. Since the databases do not communicate with each other, $\mathbb{H}_n$ is unknown to the other databases. We use $\mathbf{m} = (m_1, \ldots, m_N)$ to represent the prefetching phase. After the prefetching phase, the user learns $|\mathbb{H}|$ messages, denoted as $\mathcal{W}_{\mathbb{H}} = \{W_{i_1}, \ldots, W_{i_{|\mathbb{H}|}}\}$. We refer to $\mathcal{W}_{\mathbb{H}}$ as *partially known private side information*.

In the retrieval phase, the user privately generates a desired message index $\theta \in [K] \backslash \mathbb{H}$, and wishes to retrieve message $W_\theta$ such that no database knows which message is retrieved. Since the desired message index $\theta$ and cached message indices $\mathbb{H}$ are independent of the message contents, for random variables $\theta$, $\mathbb{H}$, and $W_1, \ldots, W_K$, we have

$$\begin{aligned} &H(\theta, \mathbb{H}, W_1, \ldots, W_K) \\ &\quad = H(\theta, \mathbb{H}) + H(W_1) + \cdots + H(W_K). \end{aligned} \tag{4}$$

In order to retrieve $W_\theta$, the user sends $N$ queries $Q_1^{[\theta,\mathbb{H}]}, \ldots, Q_N^{[\theta,\mathbb{H}]}$ to the $N$ databases, where $Q_n^{[\theta,\mathbb{H}]}$ is the query sent to the $n$th database for message $W_\theta$ given the user has partially known private side information $\mathcal{W}_{\mathbb{H}}$. The queries are generated according to $\mathbb{H}$, which is independent of the realizations of the $K$ messages. Therefore, we have

$$I(W_1, \ldots, W_K; Q_1^{[\theta,\mathbb{H}]}, \ldots, Q_N^{[\theta,\mathbb{H}]}) = 0. \tag{5}$$

To ensure that individual databases do not know which message is retrieved and also do not know the cached messages from other databases, i.e., to guarantee the privacy of $(\theta, \mathbb{H} \backslash \mathbb{H}_n)$, we need to satisfy the following privacy constraint, $\forall n \in [N]$, $\forall \mathbb{H}, \mathbb{H}'$ such that $|\mathbb{H}| = |\mathbb{H}'| \leq M$, $\mathbb{H}_n \subset \mathbb{H}$, $\mathbb{H}_n \subset \mathbb{H}'$, and $\forall \theta \in [K] \setminus \mathbb{H}$, $\forall \theta' \in [K] \setminus \mathbb{H}'$,

$$\begin{aligned} &(Q_n^{[\theta,\mathbb{H}]}, A_n^{[\theta,\mathbb{H}]}, W_1, \ldots, W_K, \mathbb{H}_n) \\ &\quad \sim (Q_n^{[\theta',\mathbb{H}']}, A_n^{[\theta',\mathbb{H}']}, W_1, \ldots, W_K, \mathbb{H}_n), \end{aligned} \tag{6}$$

where $A \sim B$ means that $A$ and $B$ are identically distributed.

Upon receiving the query $Q_n^{[\theta,\mathbb{H}]}$, the $n$th database replies with an answering string $A_n^{[\theta,\mathbb{H}]}$, which is a function of $Q_n^{[\theta,\mathbb{H}]}$ and all the $K$ messages. Therefore, $\forall \theta \in [K] \setminus \mathbb{H}, \forall n \in [N]$,

$$H(A_n^{[\theta,\mathbb{H}]} | Q_n^{[\theta,\mathbb{H}]}, W_1, \ldots, W_K) = 0. \tag{7}$$

After receiving the answering strings $A_1^{[\theta,\mathbb{H}]}, \ldots, A_N^{[\theta,\mathbb{H}]}$ from all the $N$ databases, the user needs to decode the desired message $W_\theta$ reliably. By using Fano's inequality, we have the following reliability constraint

$$H\left(W_\theta | \mathcal{W}_{\mathbb{H}}, \mathbb{H}, Q_1^{[\theta,\mathbb{H}]}, \ldots, Q_N^{[\theta,\mathbb{H}]}, A_1^{[\theta,\mathbb{H}]}, \ldots, A_N^{[\theta,\mathbb{H}]}\right) = o(L), \tag{8}$$

where $o(L)$ denotes a function such that $\frac{o(L)}{L} \to 0$ as $L \to \infty$.

For fixed $N$, $K$, and prefetching scheme $\mathbf{m}$ =

[2] We thank Dr. Hua Sun for pointing this out.

$(m_1, \ldots, m_N)$, a pair $(D(\mathbf{m}), L(\mathbf{m}))$ is achievable if there exists a PIR scheme for messages of size $L(\mathbf{m})$ symbols long with partially known private side information satisfying the privacy constraint (6) and the reliability constraint (8), where $D(\mathbf{m})$ represents the expected number of downloaded symbols (over all the queries) from the $N$ databases via the answering strings $A_{1:N}^{[\theta,\mathbb{H}]}$, where $A_{1:N}^{[\theta,\mathbb{H}]} = (A_1^{[\theta,\mathbb{H}]}, \ldots, A_N^{[\theta,\mathbb{H}]})$, i.e.,

$$D(\mathbf{m}) = \sum_{n=1}^{N} H\left(A_n^{[\theta,\mathbb{H}]}\right). \tag{9}$$

In this work, for fixed $N$, $K$, and $M$, we aim to characterize the optimal normalized download cost $D^*$, where

$$D^* = \inf_{\mathbf{m}:(3)} \left\{ \frac{D(\mathbf{m})}{L(\mathbf{m})} : (D(\mathbf{m}), L(\mathbf{m})) \text{ is achievable} \right\}. \tag{10}$$

## III. Main Results

We characterize the exact normalized download cost for the PIR problem with partially known private side information as shown in the following theorem.

**Theorem 1** *In the PIR problem with partially known private side information under the cache memory size constraint $|\mathbb{H}| \leq M$, the optimal normalized download cost is*

$$D^* = 1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-M-1}} = \frac{1 - (\frac{1}{N})^{K-M}}{1 - \frac{1}{N}}. \tag{11}$$

The converse proof for Theorem 1 is given in Section IV, and the achievability proof for Theorem 1 is given in Section V. Theorem 1 does not assume any particular property for the prefetching strategy, i.e., $\mathbf{m}$ is arbitrary except for satisfying the memory size constraint. We have a few comments.

Theorem 1 implies that $C = \frac{1}{D^*} = \frac{1-\frac{1}{N}}{1-(\frac{1}{N})^{K-M}}$. Surprisingly, this capacity expression is exactly the same as the capacity for the PIR problem with completely unknown private side information in [29]. This implies that there is no loss in capacity due to employing the same databases for both prefetching and retrieval phases. The reason for this phenomenon is that although each database has a partial knowledge about some of the cached messages at the user, the privacy constraint on this known side information is relaxed.

The normalized download cost in Theorem 1 is the same as the normalized download cost for the classical PIR problem [10] if the number of messages is $K - M$. That is, a cache of size $M$ messages effectively reduces the total number of messages by $M$. The effective reduction in the number of messages by the cache size results in a significant reduction in the download cost due to the presence of side information at the user even though it is partially known by the databases and it needs to be kept private against other databases.

The optimal prefetching strategy exploits the entire cache memory of the user as the capacity expression is monotonically increasing in $M$.

In Section V, we present the capacity achieving schemes for the partially known private side information. We note that, in general the PIR scheme in [29] is a valid achievable scheme for our problem as well. Nevertheless, in the special case of *uniform prefetching*, i.e., $m_n = \frac{M}{N} = m \in \mathbb{N}$, we provide a different achievable scheme that exploits the prefetching uniformity to work with message size $L = N^{K-m} = N^{K-\frac{M}{N}}$ in contrast to $L = N^K$ needed for the scheme in [29], i.e., the message size is decreased by an exponential factor $N^{\frac{M}{N}}$. Furthermore, we note that although both schemes need an MDS code to reduce the number of downloaded equations, the field size needed to realize this MDS code is significantly smaller with our scheme (if $\frac{M}{N} \in \mathbb{N}$) compared with the field size needed in the scheme in [29]. This implies that although *uniform prefetching* does not affect the PIR capacity, it significantly simplifies the achievable scheme.

## IV. Converse Proof

In this section, we derive a general lower bound for the normalized download cost $D^*$ given in (10). We extend the techniques presented in [10], [29] to the PIR problem with partially known private side information.

For the prefetching vector $\mathbf{m} = (m_1, \ldots, m_N)$ satisfying (3), we note that satisfying the memory size constraint with equality leads to a valid lower bound on (10). Consequently, we first consider the case $\sum_{n=1}^{N} m_n = \tilde{M} \leq M$, i.e., we study the case when the user learns $\tilde{M}$ messages after the prefetching phase. Since we do not specify the prefetching strategy $\mathbf{m}$ in advance, the following lower bound is valid for all $\mathbf{m}$ such that $\sum_{n=1}^{N} m_n = \tilde{M}$. Without loss of generality, we relabel the $\tilde{M}$ cached messages as $W_1, W_2, \ldots, W_{\tilde{M}}$, i.e., $\mathbb{H} = \{1, 2, \ldots, \tilde{M}\}$ and $\mathcal{W}_{\mathbb{H}} = W_{1:\tilde{M}}$. We first need the following lemma, which characterizes a lower bound on the length of the undesired portion of the answering strings as a consequence of the privacy constraint on the retrieved message. The proof of Lemma 1 is provided in [32, Lemma 1].

**Lemma 1 (Interference lower bound)** *For the PIR with partially known private side information, the interference from undesired messages within the answering strings, $D - L$, is lower bounded by,*

$$\begin{aligned} &D - L + o(L) \\ &\quad \geq I\left(W_{\tilde{M}+2:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+1,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1}\right). \end{aligned} \tag{12}$$

If the privacy constraint is absent, the user downloads only $L$ symbols for the desired message, however, when the privacy constraint is present, it should download $D$ symbols. The difference between $D$ and $L$, i.e., $D - L$, corresponds to the undesired portion of the answering strings. Note that Lemma 1 is an extension of [10, Lemma 5] if $\tilde{M} = 0$, i.e., the user has no partially known private side information. Lemma 1 differs from its counterpart in [28, Lemma 1] in two aspects, namely, the left hand side is $D(r) - L(1 - r)$ in [28] as the user requests to download the uncached bits only, and the bound in [28, Lemma 1] constructs $K - 1$ distinct lower bounds by changing $k$ in contrast to one bound here as it always starts from $W_{\tilde{M}+2}$. Finally, we note that a similar argument

to Lemma 1 can be implied from [29]. In the following lemma, we prove an inductive relation for the mutual information term on the right hand side of (12). The proof of Lemma 2 is provided in [32, Lemma 2].

**Lemma 2 (Induction lemma)** *For all $k \in \{\tilde{M}+2, \dots, K\}$, the mutual information term in Lemma 1 can be inductively lower bounded as,*

$$I\left(W_{k:K}; \mathbb{H}, Q_{1:N}^{[k-1,\mathbb{H}]}, A_{1:N}^{[k-1,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:k-1}\right) \geq \frac{1}{N} I\left(W_{k+1:K}; \mathbb{H}, Q_{1:N}^{[k,\mathbb{H}]}, A_{1:N}^{[k,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:k}\right) + \frac{L-o(L)}{N}. \tag{13}$$

Lemma 2 is a generalization of [10, Lemma 6] to our setting. The main difference between Lemma 2 and [29] is that in order to apply the *partial* privacy constraint, the random variable $\mathbb{H}$ should be used in its local form $\mathbb{H}_n$ as it corresponds to the partial knowledge of the $n$th database.

Now, we are ready to derive the lower bound for arbitrary $K$, $N$, and $\tilde{M}$. For fixed $N$, $K$, and $\tilde{M} \leq M$, by applying Lemma 1 and Lemma 2 successively, we have

$$D \overset{(12)}{\geq} L - o(L) + I\left(W_{\tilde{M}+2:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+1,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1}\right) \tag{14}$$

$$\overset{(13)}{\geq} L + \frac{L}{N} - o(L) + \frac{1}{N} I\Big(W_{\tilde{M}+3:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+2,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+2,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:\tilde{M}+2}\Big) \tag{15}$$

$$\overset{(13)}{\geq} \dots \tag{16}$$

$$\overset{(13)}{\geq} L\left(1 + \frac{1}{N} + \dots + \frac{1}{N^{K-\tilde{M}-1}}\right) - o(L), \tag{17}$$

where (14) follows from Lemma 1, (15)-(17) follow from applying Lemma 2 starting from $k = \tilde{M}+2$ to $k = K$, which differs from [10] in terms of the starting point of the induction. We conclude the converse proof by dividing by $L$ and taking $L \to \infty$ in (17). Since $\tilde{M} \leq M$, the lowest lower bound is obtained by taking $\tilde{M} = M$, which yields the final converse bound.

## V. Achievability Proof

We first note that the achievability scheme proposed in [29] for the PIR problem with completely unknown private side information also works for the PIR problem with partially known private side information here. The PIR scheme in [29] is based on MDS codes and consists of two stages. The first stage determines the systematic part of the MDS code according to the queries generated in [10], which protects the privacy of the desired message. In the second stage, the user reduces the number of the downloaded equations by downloading the parity part of the MDS code only. For the case of partially known private side information here, two privacy constraints should be satisfied: the desired message privacy constraint and the side information privacy constraint. For the desired message, we note that the user should guarantee that the queries designed to retrieve any of the $K - m_n$ messages should be indistinguishable at the $n$th database. Due to the first stage, the privacy of the desired message holds as it was designed to protect the privacy of all $K$ messages, which is more restricted. Furthermore, the PIR scheme in [29] also protects the privacy of the side information. In our model, we note that we need to protect the privacy of $M - m_n$ messages from the $n$th database, as the remaining $m_n$ messages are known to the $n$th database. Since the privacy constraint imposed on the side information in our model is less restricted than [29], using the scheme in [29] satisfies the privacy constraint of the side information in our case as well. That is, the $n$th database cannot infer which other $M - m_n$ messages the user holds. The PIR scheme in [29] achieves the normalized download cost in Theorem 1. The PIR scheme in [29] requires a message size of $N^K$ symbols. In the following, we propose another achievability scheme which requires a message size of $N^{K-\frac{M}{N}}$, if $m_n = \frac{M}{N} \in \mathbb{N}$. Thus, this scheme requires smaller sub-packetization and smaller field size for the MDS code.

Our PIR scheme for partially known private side information is based on the PIR schemes in [10], [29]. To protect the privacy of the partially known private side information and the privacy of the desired message, similar to [10], we apply the following three principles recursively: 1) database symmetry, 2) message symmetry within each database, and 3) exploiting undesired messages as side information. We reduce the download cost by utilizing the reconstruction property of MDS codes by exploiting partially known private side information as in [29]. The side information enables the user to request reduced number of equations as a consequence of the user's knowledge of $M$ messages from the prefetching phase. Nevertheless, to protect the privacy of the side information, the user actually queries MDS coded symbols which is mixture of $K - m_n$ messages. The main difference between our achievability scheme and that in [10], [29] is that since the $n$th database knows that the user has prefetched $m_n$ messages, the user does not need to protect the privacy for these $m_n$ messages from the $n$th database. This effectively reduces the number of messages that the scheme in [29] needs to operate on to $K - m_n$ messages in contrast to $K$ in [29]. When $\frac{M}{N} \in \mathbb{N}$, we show that if the user caches the same number of messages from each database, i.e., $m_n = \frac{M}{N}$, for all $n$, then the lower bound in (11) is achievable by this scheme. This scheme reduces the message size requirement from $L = N^K$ in [29] to $L = N^{K-\frac{M}{N}}$ here, simplifying the achievable scheme.

### A. Motivating Example: $N = 2$ Databases, $K = 4$ Messages, and $M = 2$ Cached Messages

Assume that each message is of size 8 symbols. We permute the symbols of messages $W_1$, $W_2$, $W_3$ and $W_4$ randomly and independently, and use $a_i$, $b_i$, $c_i$ and $d_i$, for $i = 1, \dots, 8$, to

denote the symbols of each permutated message, respectively. In this example, in the prefetching phase, the user caches message $W_3$ from database 1, and message $W_4$ from database 2; and in the retrieval phase, the user wishes to retrieve message $W_1$ privately. The user first generates the query table in Table I. In Table I, the user queries 7 symbols. Since the user knows $d_1$ from the cached message $W_4$, in order to use the partially known private side information, the user can in fact reduce the number of queries to 6 equations per database by ignoring $d_1$. However, if the user simply does not download $d_1$, it compromises the privacy of $W_4$ at database 1. Alternatively, the user queries the MDS coded version of the 7 symbols. By using these 7 symbols as the systematic part, we can use a $(13, 7)$ MDS code. By downloading the 6 parity symbols, the user can reconstruct the whole 7 symbols utilizing the knowledge of $d_1$. Therefore, the normalized download cost for our achievability scheme is $\frac{6+6}{8} = \frac{3}{2}$, which matches the lower bound in (11) for this case.

For database 1, the query table in Table I induces the same distribution on the messages $W_1$, $W_2$ and $W_4$. Therefore, we guarantee the privacy of the desired message. The reliability constraint can also be verified. Note that $b_2$ is downloaded from database 2, and $d_2$ is downloaded in the prefetching phase. Therefore, $a_3$ and $a_4$ are decodable. By getting $b_4 + c_3$ from database 2, the user can get $b_4$ due to the private side information $W_3$. Therefore, the user can decode $a_7$ from $a_7 + b_4 + d_4$. Similar arguments follow for database 2.

TABLE I
QUERY TABLE FOR $K = 4$, $N = 2$, $M = 2$.

| DB1 | DB2 |
|---|---|
| $a_1$ | $a_2$ |
| $b_1$ | $b_2$ |
| $d_1$ | $c_1$ |
| $a_3 + b_2$ | $a_5 + b_1$ |
| $a_4 + d_2$ | $a_6 + c_2$ |
| $b_3 + d_3$ | $b_4 + c_3$ |
| $a_7 + b_4 + d_4$ | $a_8 + b_3 + c_4$ |
| $\mathcal{W}_{\mathbb{H}_1}$=$\{W_3\}$ | $\mathcal{W}_{\mathbb{H}_2}$=$\{W_4\}$ |

### B. General Achievable Scheme for $\frac{M}{N} \in \mathbb{N}$

Let $\frac{M}{N} = m$. In the prefetching phase, the user caches $m$ messages from each database. To achieve the lower bound shown in (11), in the retrieval phase, we choose the message size as $L = N^{K-m}$ symbols. The details of the achievable scheme are as follows:

1) *Initialization:* The user permutes each message randomly and independently. After the random permutation, we use $U_i(j)$ to denote the $j$th symbol of the permuted message $W_i$. Suppose the user wishes to retrieve $W_\theta$ privately. We then prepare the query table by first querying $U_\theta(1)$ from database 1. Set the round index to $r = 1$.
2) *Symmetry across databases:* The user queries the same number of equations with the same structure as database 1 from the remaining databases.
3) *Message symmetry:* For each database, to satisfy the privacy constraint, the user should query equal amount of symbols from all other $K - m$ messages. Since the user has cached $m$ messages from each database in the prefetching phase, the user does not need to protect the privacy for these $m$ messages. For the $r$th round, the user queries sums of every $r$ combinations of the $K-m$ messages.
4) *Exploiting side information:* For database 1, the user exploits the side information equations obtained from the other $(N - 1)$ databases to query sum of $r + 1$ combinations of the $K - m$ messages, where sum of $r$ combinations is the side information. If the $r$ combinations contain the cached message from database 1, we replace the overlapping symbols through the symbols cached from other databases.
5) *Repeat* steps 2, 3, 4 after setting $r = r + 1$ until $r = K - m + 1$.
6) *Shuffling the order of queries:* By shuffling the order of queries uniformly, all possible queries can be made equally likely regardless of the message index. This guarantees the privacy of the desired message.
7) *Downloading MDS parity parts:* Now, the query table is finished. For each database, let $p$ be the number of queried symbols in the query table, and let $q$ be the number of queried symbols which are determined by the side information the user cached in the prefetching phase. Apply a $(2p - q, p)$ MDS code to the queried symbols by letting the $p$ symbols to be the systematic part. Finally, the user downloads the parity parts of the MDS-coded answering strings which are $p - q$ symbols for each database.

### C. Normalized Download Cost

We now calculate the total number of downloaded symbols. We first calculate $p$, which is the number of queried symbols in the query table for each database,

$$p = \sum_{i=1}^{K-m} \binom{K-m}{i} (N-1)^{i-1} = \frac{1}{N-1} \left(N^{K-m} - 1\right), \tag{18}$$

where $\binom{K-m}{r}$ in (18) corresponds to the queries of sums of every $r$ combinations of the $K - m$ messages, and $(N - 1)^{r-1}$ corresponds to the number of sets of the available side information from other $(N - 1)$ databases.

We then calculate $q$, which is the number of queried symbols which are determined by the side information the user cached in the prefetching phase,

$$q = \sum_{i=1}^{(N-1)m} \binom{(N-1)m}{i} (N-1)^{i-1} = \frac{\left(N^{(N-1)m} - 1\right)}{N-1}, \tag{19}$$

where $\binom{(N-1)m}{r}$ in (19) corresponds to the queries which can be determined by the partially known private side information, and $(N-1)^{r-1}$ corresponds to the number of sets of queries consisting of $r$ combinations.

Next, we calculate the number of symbols for the desired message,

$$L = N \sum_{i=0}^{K-m-1} \binom{K-m-1}{i} (N-1)^i = N^{K-m}, \quad (20)$$

where $\binom{K-m-1}{r-1}$ in (20) corresponds to the queries containing the desired message and $(N-1)^{r-1}$ corresponds to the number of sets of queries consisting of $r$ combinations.

Therefore, the normalized download cost becomes,

$$\frac{D}{L} = \frac{N(p-q)}{L} = \frac{1}{1-\frac{1}{N}} \times \left[1 - \left(\frac{1}{N}\right)^{K-M}\right], \quad (21)$$

which matches the lower bound in (11).

## VI. Conclusion

In this paper, we have introduced a new PIR model, namely, PIR with partially known private side information as a natural model for studying practical PIR problems with cached side information. In this model, the user and the databases engage in a caching/PIR scenario which consists of two phases, namely, prefetching phase and retrieval phase. The $n$th database provides the user with $m_n$ side information messages in the prefetching phase such that $\sum_{n=1}^{N} m_n \leq M$, hence, each database has *partial knowledge* about the side information in contrast to full knowledge in [26] and no knowledge in [27]–[29]. Based on this side information, the user designs a retrieval scheme that does not reveal the identity of the desired message or the identities of the remaining $M - m_n$ messages to the $n$th database. For this model, we determined the exact capacity to be $C = \frac{1-\frac{1}{N}}{1-(\frac{1}{N})^{K-M}}$. The capacity is attained for any prefetching strategy that satisfies the cache memory size constraint with equality. The achievable scheme in [29] can also be used for this model. We further proposed another PIR scheme which requires smaller sub-packetization and field size for the case of uniform prefetching. Uniform prefetching, when feasible, is optimal. Interestingly, the capacity expression we derive for this problem is exactly the same as the capacity expression for the PIR problem with completely unknown side information [29]. Therefore, our result implies that there is no loss in employing the same databases for prefetching and retrieval purposes.

## References


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

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

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

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

[5] G. Fanti and K. Ramchandran. Efficient private information retrieval over unsynchronized databases. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1229–1239, October 2015.

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

[7] A. Fazeli, A. Vardy, and E. Yaakobi. Codes for distributed PIR with low storage overhead. In *IEEE ISIT*, June 2015.

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

[9] H. Sun and S. A. Jafar. The capacity of private information retrieval. In *IEEE Globecom*, December 2016.

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

[11] H. Sun and S. A. Jafar. The capacity of robust private information retrieval with colluding databases. *IEEE Trans. on Info. Theory*, 2017.

[12] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb. Private information retrieval schemes for coded data with arbitrary collusion patterns. 2017. Available at arXiv:1701.07636.

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

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

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

[16] H. Sun and S. A. Jafar. Multiround private information retrieval: Capacity and storage overhead. 2016. Available at arXiv:1611.02257.

[17] K. Banawan and S. Ulukus. Multi-message private information retrieval: Capacity results and near-optimal schemes. *IEEE Trans. on Info. Theory*. Submitted February 2017. Also available at arXiv:1702.01739.

[18] K. Banawan and S. Ulukus. The capacity of private information retrieval from Byzantine and colluding databases. *IEEE Trans. on Info. Theory*. Submitted June 2017. Also available at arXiv:1706.01442.

[19] Q. Wang and M. Skoglund. Symmetric private information retrieval for MDS coded distributed storage. 2016. Available at arXiv:1610.04530.

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

[21] H. Sun and S. A. Jafar. Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al. 2017. Available at arXiv: 1701.07807.

[22] Y. Zhang and G. Ge. A general private information retrieval scheme for MDS coded databases with colluding servers. 2017. Available at arXiv: 1704.06785.

[23] Y. Zhang and G. Ge. Multi-file private information retrieval from MDS coded databases with colluding servers. 2017. Available at arXiv: 1705.03186.

[24] Q. Wang and M. Skoglund. Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers. 2017. Available at arXiv:1708.05673.

[25] Q. Wang and M. Skoglund. Secure symmetric private information retrieval from colluding databases with adversaries. 2017. Available at arXiv:1707.02152.

[26] R. Tandon. The capacity of cache aided private information retrieval. 2017. Available at arXiv: 1706.07035.

[27] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson. Private information retrieval with side information. 2017. Available at arXiv:1709.00112.

[28] Y.-P. Wei, K. Banawan, and S. Ulukus. Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching. 2017. Available at arXiv:1709.01056.

[29] Z. Chen, Z. Wang, and S. A. Jafar. The capacity of private information retrieval with private side information. 2017. Available at arXiv:1709.03022.

[30] M. Karmoose, L. Song, M. Cardone, and C. Fragouli. Private broadcasting: an index coding approach. 2017. Available at arXiv: 1701.04958.

[31] M. Karmoose, L. Song, M. Cardone, and C. Fragouli. Preserving privacy while broadcasting: $k$-limited-access schemes. 2017. Available at arXiv:1705.08437.

[32] Y.-P. Wei, K. Banawan, and S. Ulukus. The capacity of private information retrieval with partially known private side information. 2017. Available at arXiv:1710.00809.