The Capacity of Private Information Retrieval With Partially Known Private Side Information

Yi-Peng Wei, *Student Member, IEEE*, Karim Banawan, *Member, IEEE*, and Sennur Ulukus, *Fellow, IEEE*

***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 (retriever) of cache-size M 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 where the user acquires side information and the retrieval phase where the user downloads desired information. In the prefetching phase, the user receives m_n full messages from the nth 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 (which is not present in its memory) 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 $C = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-M-1}}\right)^{-1} = \frac{1-\frac{1}{N}}{1-(\frac{1}{N})^{K-M}}$. Interestingly, this result is the same if none of the databases knows any of the prefetched side information, i.e., when the side information is obtained externally, a problem posed by Kadhe et al. and settled by Chen-Wang-Jafar recently. Thus, our result implies that there is no loss in using the same databases for both prefetching and retrieval phases.**

***Index Terms*— Private information retrieval, PIR capacity, side information, caching.**

I. Introduction

THE private information retrieval (PIR) problem 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]–[5]. In the classical setting of PIR in [1], a user wishes to retrieve a single message (or a file) out of K messages replicated across N non-communicating databases without leaking any information about the identity of the retrieved message. To that end, the user submits a query to each database. Each database responds truthfully with an answer string. The user reconstructs the desired message from the collected answer strings. Trivially, the user can download the entire database and incur a linear (in number of messages) download cost, but this retrieval strategy is highly inefficient. The efficiency of a PIR scheme is measured by the normalized download cost, which is the cost of privately downloading one bit of the desired message.[1] 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., [6]–[11]. In the leading work of Sun-Jafar [12], 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 [12] characterizes the capacity of the classical PIR model to be $C = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}}\right)^{-1}$ using a greedy achievable scheme that is closely related to blind interference alignment [13] and an induction-based converse argument. Following the work of Sun-Jafar [12], the capacity of many interesting variants of the classical PIR model have been investigated, such as, PIR from colluding databases, robust PIR, symmetric PIR, PIR from MDS-coded databases, PIR for arbitrary message lengths, multi-round PIR, multi-message PIR, PIR from Byzantine databases, secure symmetric PIR with adversaries, and their several combinations [14]–[28].

In this paper, we consider the problem of PIR with partially known private side information. Our work is most closely related to [29]–[32].[2] 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 [29]–[32] differ in three important aspects, namely, 1) the structure of the side information, 2) the presence

Manuscript received November 26, 2017; revised May 7, 2019; accepted July 31, 2019. Date of publication August 19, 2019; date of current version November 20, 2019. This work was supported in part by the NSF Grant CNS 13-14733, Grant CCF 14-22111, Grant CNS 15-26608, and Grant CCF 17-13977. This article was presented in part at the 2018 CISS.

Y.-P. Wei was with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA. He is now with Google Inc., Mountain View, CA 94043 USA (e-mail: ypwei@umd.edu).

K. Banawan was with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA. He is now with the Electrical Engineering Department, Faculty of Engineering, Alexandria University, Alexandria 21544, Egypt (e-mail: kbanawan@umd.edu).

S. Ulukus is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: ulukus@umd.edu).

Communicated by A. Ramamoorthy, Associate Editor for Coding Techniques.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2936023

[1] Since the upload cost does not scale with the message size, we do not consider the upload cost here.

[2] A parallel line of work that studies privacy issues of requests and side information in index coding based broadcast systems can be found in [33], [34].

or absence of privacy constraints on the side information, and 3) the databases' awareness of the side information at its initial acquisition. Here, structure of the side information refers to whether the side information is in the form of full messages or parts of messages or whether messages are mixed through functions (coded/uncoded); privacy of the side information refers to whether the user further aims to keep the side information private from the databases; and databases' awareness of the side information refers to whether the databases knew the initially prefetched side information.

Specifically, reference [29] studies the capacity of the cache-aided PIR where 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 [29] assumes that the cache content is perfectly known by all the databases, and hence there is no need to protect the privacy of the cached content. Reference [29] determines the optimal download cost for this model to be $D^*(r) = \frac{1}{C(r)} = (1-r)\left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}}\right)$ using a memory-sharing achievable scheme and a converse that utilizes Han's inequality. This conclusion is somewhat pessimistic in that the user cannot exploit the cached content as useful side information during PIR to reduce the download cost, since the databases are fully aware of it; the optimum $D^*(r)$ formula indicates that the user should download the uncached part of the content, i.e., $(1-r)$ fraction, via the optimum PIR scheme in [12]. This result motivates [30], [31] to study the other extreme when the databases are completely unaware of the side information at its initial acquisition. References [30] and [31] differ in terms of the structure of the cached content: [30] considers the case where rK full messages are cached, and [31] considers the case where a random r fraction of the symbols of each of K messages is cached. In this case, [31] shows a significant reduction in the download cost over [29], as the user can now leverage the cached bits as side information, since they are unknown to the databases. In [31], there is no privacy constraint on the cached content.

Reference [30] further introduces 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, i.e., the queries of the user should not leak any information about the cached content to the databases. The exact capacity for this problem is settled in [32] to be $C = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-M-1}}\right)^{-1}$. The optimal achievable scheme in this case starts from the traditional achievable scheme without side information in [12] and reduces the download cost by utilizing the reconstruction property of MDS codes.

In this paper, 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 [30]–[32]. One way to do this could be to employ one of the databases for prefetching the side information and exclude it from the retrieval process. Therefore, for the remaining $N-1$ databases, the side information is completely unknown. This solution is strictly sub-optimal as the capacity expression in [32] (shown as C in the previous paragraph) is monotonically decreasing in N. An alternative solution could be to devise a refreshing mechanism that ensures that the cached content is essentially random from the perspective of each database [29], which may be challenging to implement. We also note that the other extreme of the problem, where the databases are fully aware of the cached content [29], 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. This poses the following questions: Can we propose efficient joint prefetching-retrieval strategies that exploit the limited knowledge of each database to drive down the download cost? How much is the loss from the full unawareness case in [30], [32]?

In this paper, 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 nth 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, namely, the part of the side information that this database has provided during the prefetching phase. In the retrieval phase, the user wants to retrieve a message (which is not present in its memory) without leaking any information to any individual database about the desired message or the remaining side information messages that are unknown to each database. The goal of this work is to design a joint prefetching-retrieval scheme that minimizes the download cost in the retrieval phase.

To that end, we first derive a general lower bound for the normalized download cost that is independent of the prefetching strategy.[3] Then, we prove that this bound is attainable using two achievable schemes. The first achievable scheme, which is proposed in [32] for completely unknown side information, is a valid achievable scheme for our problem with partially known side information for any prefetching strategy.[4] 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. While the first achievable scheme [32] requires a message size of $L = N^K$, the second achievable scheme proposed here requires a message size of $L = N^{K-\frac{M}{N}}$, which scales down the message size by an exponential factor $N^{\frac{M}{N}}$. This, in turn, simplifies the achievable scheme and minimizes the total number of downloaded bits without sacrificing from

[3] The lower bound in [32] cannot be directly applied to this work. Here, each database knows the messages that are prefetched from themselves. Therefore, the user cannot further utilize these prefetched side information messages. However, at the same time, the user does not need to keep the prefetched messages private, since each database already knows them.

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

the capacity. 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 [32] recently. Therefore, 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 consists of L symbols with each symbol over field $\mathbb{F}_q$,[5]

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

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

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 nth database. We denote the indices of the cached messages from the nth database as $\mathbb{H}_n$. Therefore, $|\mathbb{H}_n| = m_n$. We denote the indices of all cached messages as $\mathbb{H}$,

$$\mathbb{H} = \bigcup_{n=1}^{N} \mathbb{H}_n, \tag{3}$$

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

$$|\mathbb{H}| = \sum_{n=1}^{N} m_n \leq M. \tag{4}$$

Since the user caches m_n messages from the nth database, $\mathbb{H}_n$ is known to the nth 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_θ such that no database knows which message is retrieved. Since the desired message index θ and cached message indices $\mathbb{H}$ are independent of the message contents, for random variables θ, $\mathbb{H}$, and $W_1, \ldots, W_K$, we have

$$\begin{aligned} &H(\theta, \mathbb{H}, W_1, \ldots, W_K) \\ &\quad = H(\theta, \mathbb{H}) + H(W_1) + \cdots + H(W_K). \end{aligned} \tag{5}$$

In order to retrieve W_θ, 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 nth database for message W_θ 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

$$I(W_1, \ldots, W_K; Q_1^{[\theta,\mathbb{H}]}, \ldots, Q_N^{[\theta,\mathbb{H}]}) = 0. \tag{6}$$

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 $\forall \theta \in [K] \setminus \mathbb{H}$, $\forall \theta' \in [K] \setminus \mathbb{H}'$,

$$\begin{aligned} &(Q_n^{[\theta,\mathbb{H}]}, A_n^{[\theta,\mathbb{H}]}, W_1, \ldots, W_K, \mathbb{H}_n) \\ &\quad \sim (Q_n^{[\theta',\mathbb{H}']}, A_n^{[\theta',\mathbb{H}']}, W_1, \ldots, W_K, \mathbb{H}_n), \end{aligned} \tag{7}$$

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

Upon receiving the query $Q_n^{[\theta,\mathbb{H}]}$, the nth 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, $\forall \theta \in [K] \setminus \mathbb{H}, \forall n \in [N]$,

$$H(A_n^{[\theta,\mathbb{H}]} | Q_n^{[\theta,\mathbb{H}]}, W_1, \ldots, W_K) = 0. \tag{8}$$

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_θ reliably. By using Fano's inequality, we have the following reliability constraint

$$H\left(W_\theta | \mathcal{W}_{\mathbb{H}}, \mathbb{H}, Q_1^{[\theta,\mathbb{H}]}, \ldots, Q_N^{[\theta,\mathbb{H}]}, A_1^{[\theta,\mathbb{H}]}, \ldots, A_N^{[\theta,\mathbb{H}]}\right) = o(L), \tag{9}$$

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

For fixed N, K, and pretching scheme $\mathbf{m} = (m_1, \ldots, m_N)$, 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 (7) and the reliability constraint (9), 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.,

$$D(\mathbf{m}) = \sum_{n=1}^{N} H\left(A_n^{[\theta,\mathbb{H}]}\right). \tag{10}$$

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

$$D^* = \inf_{\mathbf{m}:(4)} \left\{ \frac{D(\mathbf{m})}{L(\mathbf{m})} : (D(\mathbf{m}), L(\mathbf{m})) \text{ is achievable} \right\}. \tag{11}$$

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*

$$D^* = 1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-M-1}} \tag{12}$$

$$= \frac{1 - (\frac{1}{N})^{K-M}}{1 - \frac{1}{N}}. \tag{13}$$

[5] Here, we use logarithm with respect to base q in the entropy functions.

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 remarks.

Remark 1 *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 [32]. 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.*

Remark 2 *Although Theorem 1 shows no loss in capacity compared to the setting in [32], the privacy constraints for these two settings are slightly different. Here, since each database knows the* m_n *messages that were prefetched from it during the prefetching phase, the privacy constraint for the desired message is over the remaining* $K-m_n$ *messages during the retrieval phase. However, in [32], since the databases are unaware of the prefetched messages, the privacy constraint for the desired message is over all* K *messages during the retrieval phase.*

Remark 3 *The normalized download cost in Theorem 1 is the same as the normalized download cost for the classical PIR problem [12] 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. *Noting that the download cost in [12] monotonically increases in the number of messages, 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.*

Remark 4 *The optimal prefetching strategy exploits the entire cache memory of the user as the capacity expression is monotonically increasing in* M.

Remark 5 *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 [32] 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 [32], 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 [32]. 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 (11). We extend the techniques presented in [12], [32] to the PIR problem with partially known private side information.

For the prefetching vector $\mathbf{m} = (m_1, \ldots, m_N)$ satisfying (4), we note that satisfying the memory size constraint with equality leads to a valid lower bound on (11). 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.

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,*

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

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 [12, Lemma 5] if $\tilde{M} = 0$, i.e., the user has no partially known private side information. Lemma 1 differs from its counterpart in [31, Lemma 1] in two aspects, namely, the left hand side is $D(r) - L(1-r)$ in [31] as the user requests to download the uncached bits only, and the bound in [31, 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 [32].

Proof: We start with the right hand side of (14),

$$\begin{aligned} &I\left(W_{\tilde{M}+2:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+1,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1}\right) \\ &\quad = I\left(W_{\tilde{M}+2:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}, W_{\tilde{M}+1} | \mathcal{W}_{\mathbb{H}}\right) \\ &\qquad - I\left(W_{\tilde{M}+2:K}; W_{\tilde{M}+1} | \mathcal{W}_{\mathbb{H}}\right). \end{aligned} \tag{15}$$

For the first term on the right hand side of (15), we have

$$
\begin{aligned}
&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}]},W_{\tilde{M}+1}|\mathcal{W}_{\mathbb{H}}\right)\\
&=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}}\right)\\
&\quad+I\left(W_{\tilde{M}+2:K};W_{\tilde{M}+1}|\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]},A_{1:N}^{[\tilde{M}+1,\mathbb{H}]},\mathcal{W}_{\mathbb{H}}\right) \qquad (16)\\
&\overset{(9)}{=}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}}\right)+o(L) \qquad (17)\\
&\overset{(5),(6)}{=}I\left(W_{\tilde{M}+2:K};A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)+o(L) \qquad (18)\\
&=H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)\\
&\quad-H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]},W_{\tilde{M}+2:K}\right)+o(L) \qquad (19)\\
&\overset{(9)}{=}H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)\\
&\quad-H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]},W_{\tilde{M}+1}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]},W_{\tilde{M}+2:K}\right)\\
&\quad+o(L) \qquad (20)\\
&\le H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)\\
&\quad-H\left(W_{\tilde{M}+1}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]},W_{\tilde{M}+2:K}\right)+o(L) \qquad (21)\\
&\overset{(5),(6)}{=}H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)\\
&\quad-H\left(W_{\tilde{M}+1}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+2:K}\right)+o(L) \qquad (22)\\
&=H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)-L+o(L) \qquad (23)\\
&\le H\left(A_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)-L+o(L) \qquad (24)\\
&\le D-L+o(L), \qquad (25)
\end{aligned}
$$

where (17), (20) follow from the decodability of $W_{\tilde{M}+1}$ given $\left(\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]},A_{1:N}^{[\tilde{M}+1,\mathbb{H}]},\mathcal{W}_{\mathbb{H}}\right)$, (18) follows from the independence of $W_{\tilde{M}+2:K}$ and $\left(\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)$, (22) follows from the independence of $W_{\tilde{M}+1}$ and $\left(\mathbb{H},Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}\right)$, and (25) follows from the independence bound.

For the second term on the right hand side of (15), we have

$$
\begin{aligned}
&I\left(W_{\tilde{M}+2:K};W_{\tilde{M}+1}|\mathcal{W}_{\mathbb{H}}\right)\\
&=H\left(W_{\tilde{M}+1}|\mathcal{W}_{\mathbb{H}}\right)-H\left(W_{\tilde{M}+1}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+2:K}\right) \qquad (26)\\
&=L-L=0. \qquad (27)
\end{aligned}
$$

Combining (15), (25), and (27) yields (14). ■

In the following lemma, we prove an inductive relation for the mutual information term on the right hand side of (14).

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,*

$$
\begin{aligned}
&I\left(W_{k:K};\mathbb{H},Q_{1:N}^{[k-1,\mathbb{H}]},A_{1:N}^{[k-1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1}\right)\\
&\quad\ge\frac{1}{N}I\left(W_{k+1:K};\mathbb{H},Q_{1:N}^{[k,\mathbb{H}]},A_{1:N}^{[k,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k}\right)\\
&\qquad+\frac{L-o(L)}{N}. \qquad (28)
\end{aligned}
$$

Lemma 2 is a generalization of [12, Lemma 6] to our setting. The main difference between Lemma 2 and [32] 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 nth database.

Proof: We start with the left hand side of (28),

$$
\begin{aligned}
&I\left(W_{k:K};\mathbb{H},Q_{1:N}^{[k-1,\mathbb{H}]},A_{1:N}^{[k-1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1}\right)\\
&=\frac{1}{N}\times N\\
&\quad\times I\left(W_{k:K};\mathbb{H},Q_{1:N}^{[k-1,\mathbb{H}]},A_{1:N}^{[k-1,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1}\right) \qquad (29)\\
&\ge\frac{1}{N}\sum_{n=1}^{N}I\Big(W_{k:K};\mathbb{H}_n,Q_n^{[k-1,\mathbb{H}]},A_n^{[k-1,\mathbb{H}]}\\
&\qquad\qquad|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1}\Big) \qquad (30)\\
&\ge\frac{1}{N}\sum_{n=1}^{N}I\Big(W_{k:K};Q_n^{[k-1,\mathbb{H}]},A_n^{[k-1,\mathbb{H}]}\\
&\qquad\qquad|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1},\mathbb{H}_n\Big) \qquad (31)\\
&\overset{(7)}{=}\frac{1}{N}\sum_{n=1}^{N}I\Big(W_{k:K};Q_n^{[k,\mathbb{H}]},A_n^{[k,\mathbb{H}]}\\
&\qquad\qquad|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1},\mathbb{H}_n\Big) \qquad (32)\\
&\overset{(5),(6)}{=}\frac{1}{N}\sum_{n=1}^{N}I\left(W_{k:K};A_n^{[k,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1},\mathbb{H}_n,Q_n^{[k,\mathbb{H}]}\right) \qquad (33)\\
&\overset{(8)}{=}\frac{1}{N}\sum_{n=1}^{N}H\left(A_n^{[k,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1},\mathbb{H}_n,Q_n^{[k,\mathbb{H}]}\right) \qquad (34)\\
&\ge\frac{1}{N}\sum_{n=1}^{N}H\left(A_n^{[k,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1},\mathbb{H},Q_{1:N}^{[k,\mathbb{H}]},A_{1:n-1}^{[k,\mathbb{H}]}\right) \qquad (35)\\
&\overset{(8)}{=}\frac{1}{N}\sum_{n=1}^{N}I\Big(W_{k:K};A_n^{[k,\mathbb{H}]}\\
&\qquad\qquad|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1},\mathbb{H},Q_{1:N}^{[k,\mathbb{H}]},A_{1:n-1}^{[k,\mathbb{H}]}\Big) \qquad (36)\\
&=\frac{1}{N}I\left(W_{k:K};A_{1:N}^{[k,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1},\mathbb{H},Q_{1:N}^{[k,\mathbb{H}]}\right) \qquad (37)\\
&\overset{(5),(6)}{=}\frac{1}{N}I\left(W_{k:K};\mathbb{H},Q_{1:N}^{[k,\mathbb{H}]},A_{1:N}^{[k,\mathbb{H}]}|\mathcal{W}_{\mathbb{H}},W_{\tilde{M}+1:k-1}\right) \qquad (38)
\end{aligned}
$$

$$\overset{(9)}{=} \frac{1}{N} I\left(W_{k:K}; W_k, \mathbb{H}, Q_{1:N}^{[k,\mathbb{H}]}, A_{1:N}^{[k,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:k-1}\right) - \frac{o(L)}{N} \tag{39}$$

$$= \frac{1}{N} I\left(W_{k:K}; W_k | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:k-1}\right) + \frac{1}{N} I\left(W_{k:K}; \mathbb{H}, Q_{1:N}^{[k,\mathbb{H}]}, A_{1:N}^{[k,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:k}\right) - \frac{o(L)}{N} \tag{40}$$

$$= \frac{1}{N} I\left(W_{k+1:K}; \mathbb{H}, Q_{1:N}^{[k,\mathbb{H}]}, A_{1:N}^{[k,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:k}\right) + \frac{L - o(L)}{N}, \tag{41}$$

where (30) follows from the non-negativity of mutual information, (32) follows from the privacy constraint, (33) follows from the independence of the messages and the queries, (34), (36) follow from the fact that answer strings are deterministic functions of the messages and the queries, (35) follows from the fact that conditioning reduces entropy, (38) follows from the independence of $W_{k:K}$ and $\left(\mathbb{H}, Q_{1:N}^{[k,\mathbb{H}]}\right)$, (39) follows from the reliability constraint on W_k, and (41) follows from the independence of W_k and $(\mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:k-1})$. ■

Now, we are ready to derive the lower bound for arbitrary K, N, and $\tilde{M}$. This can be obtained by applying Lemma 1 and Lemma 2 successively.

Lemma 3 *For fixed N, K, and $\tilde{M} \leq M$, we have*

$$D \geq L\left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-\tilde{M}-1}}\right) - o(L). \tag{42}$$

Proof: We have

$$D \overset{(14)}{\geq} L + I\left(W_{\tilde{M}+2:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+1,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+1,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1}\right) - o(L) \tag{43}$$

$$\overset{(28)}{\geq} L + \frac{L}{N} + \frac{1}{N} I\left(W_{\tilde{M}+3:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+2,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+2,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:\tilde{M}+2}\right) - o(L) \tag{44}$$

$$\overset{(28)}{\geq} L + \frac{L}{N} + \frac{L}{N^2} + \frac{1}{N} I\left(W_{\tilde{M}+4:K}; \mathbb{H}, Q_{1:N}^{[\tilde{M}+3,\mathbb{H}]}, A_{1:N}^{[\tilde{M}+3,\mathbb{H}]} | \mathcal{W}_{\mathbb{H}}, W_{\tilde{M}+1:\tilde{M}+3}\right) - o(L) \tag{45}$$

$$\overset{(28)}{\geq} \ldots \tag{46}$$

$$\overset{(28)}{\geq} L\left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-\tilde{M}-1}}\right) - o(L), \tag{47}$$

where (43) follows from Lemma 1, (44)-(47) follow from applying Lemma 2 starting from $k = \tilde{M} + 2$ to $k = K$, which differs from [12] in terms of the starting point of the induction. ■

We conclude the converse proof by dividing by L and taking $L \to \infty$ in (42), to have

$$D^* \geq 1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-\tilde{M}-1}}. \tag{48}$$

Finally, we note that the right hand side of (48) is monotonically decreasing in $\tilde{M}$. Since $\tilde{M} \leq M$, the lowest lower bound is obtained by taking $\tilde{M} = M$, which yields the final converse bound,

$$D^* \geq 1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-M-1}}. \tag{49}$$

Remark 6 *We note that if* (49) *is tight, any prefetching strategy* $\mathbf{m}$ *such that* $\sum_{n=1}^{N} m_n < M$ *is strictly suboptimal. Furthermore, the lower bound in* (49) *is the same for all prefetching strategies* $\mathbf{m}$ *satisfying* $\sum_{n=1}^{N} m_n = M$*. In Section V, we show that this lower bound is tight.*

V. Achievability Proof

We first note that the achievability scheme proposed in [32] 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 [32] 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 [12], which protects the privacy of the desired message, i.e., in the first stage, the user designs the queries such that no information is leaked about which message out of the K messages is the desired one. 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 nth database (i.e., with the exception of the m_n messages that the nth database has provided). 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 [32] also protects the privacy of the side information. The scheme in [32] ensures that the queries do not reveal the identity of the M messages that are possessed by the user as side information. In our model, we note that we need to protect the privacy of $M - m_n$ messages from the nth database, as the remaining m_n messages are known to the nth database. Since the privacy constraint imposed on the side information in our model is less restricted than [32], using the scheme in [32] satisfies the privacy constraint of the side information in our case as well. That is, the nth database cannot infer which other $M - m_n$ messages the user holds. The PIR scheme in [32] achieves the normalized download cost in Theorem 1. The PIR scheme in [32] 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 [12], [32]. To protect the privacy of the partially known private side information and the privacy of the desired message, similar to [12], 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 [32]. 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 [12], [32] is that since the nth 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 nth database. This effectively reduces the number of messages that the scheme in [32] needs to operate on to $K - m_n$ messages in contrast to K in [32]. 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 (12) is achievable by this scheme. This scheme reduces the message size requirement from $L = N^K$ in [32] to $L = N^{K-\frac{M}{N}}$ here, simplifying the achievable scheme.

A. Motivating Examples

1) $N = 2$ Databases, $K = 4$ Messages, and $M = 2$ Cached Messages: Assume that each message is of size 8 symbols. We use a_i, b_i, c_i and d_i, for $i = 1, \dots, 8$, to denote the symbols of messages W_1, W_2, W_3 and W_4, 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 (12) 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
$a_3 + b_2$	$a_5 + b_1$
$a_4 + d_2$	$a_6 + c_2$
$b_3 + d_3$	$b_4 + c_3$
$a_7 + b_4 + d_4$	$a_8 + b_3 + c_4$

$\mathcal{W}_{\mathbb{H}_1}$=$\{W_3\}$	$\mathcal{W}_{\mathbb{H}_2}$=$\{W_4\}$

2) $N = 2$ Databases, $K = 5$ Messages, and $M = 2$ Cached Messages: Assume that each message is of size 16 symbols. We use a_i, b_i, c_i, d_i and e_i, for $i = 1, \dots, 16$, to denote the symbols of messages W_1, W_2, W_3, W_4, and W_5, respectively. In this example, in the prefetching phase, the user caches message W_4 from database 1, and message W_5 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 II. In Table II, the user queries 15 symbols. Since the user knows e_1 from the cached message W_5, in order to use the partially known private side information, the user in fact queries the MDS coded version of the 15 symbols. By using these 15 symbols as the systematic part, we can use a $(29, 15)$ MDS code. By downloading the 14 parity symbols, the user can reconstruct the whole 15 symbols. Therefore, the normalized download cost for our achievability scheme is $\frac{14+14}{16} = \frac{7}{4}$, which matches the lower bound in (12) for this case.

For database 1, the query table in Table II induces the same distribution on the messages W_1, W_2, W_3 and W_5. Therefore, we guarantee the privacy of the desired message. The reliability constraint can also be verified. Note that b_2, c_2 are downloaded from database 2, and e_2 is downloaded in the prefetching phase. Therefore, a_3, a_4 and a_5 are decodable. By getting $b_6 + d_3$ from database 2, the user can get b_6 due to the private side information W_4. Similarly, c_6 is also decodable. Therefore, the user can decode a_{10} from $a_{10} + b_6 + e_5$ and a_{11} from $a_{11} + c_6 + e_6$. By getting $b_8 + c_8 + d_7$ from database 2, the user can get $b_8 + c_8$ due to the private side information W_4. Therefore, the user can decode a_{15} from $a_{15} + b_8 + c_8 + e_8$. Similar arguments follow for database 2.

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 (12), 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 jth symbol of the permuted message W_i. Suppose the user wishes to retrieve W_θ

TABLE II

QUERY TABLE FOR $K = 5$, $N = 2$, $M = 2$

DB1	DB2
a_1	a_2
b_1	b_2
c_1	c_2
e_1	d_1
$a_3 + b_2$	$a_6 + b_1$
$a_4 + c_2$	$a_7 + c_1$
$a_5 + e_2$	$a_8 + d_2$
$b_3 + c_3$	$b_5 + c_5$
$b_4 + e_3$	$b_6 + d_3$
$c_4 + e_4$	$c_6 + d_4$
$a_9 + b_5 + c_5$	$a_{12} + b_3 + c_3$
$a_{10} + b_6 + e_5$	$a_{13} + b_4 + d_5$
$a_{11} + c_6 + e_6$	$a_{14} + c_4 + d_6$
$b_7 + c_7 + e_7$	$b_8 + c_8 + d_7$
$a_{15} + b_8 + c_8 + e_8$	$a_{16} + b_7 + c_7 + d_8$

$\mathcal{W}_{\mathbb{H}_1}=\{W_4\}$	$\mathcal{W}_{\mathbb{H}_2}=\{W_5\}$

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 rth 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,

$$\begin{aligned} p &= \binom{K-m}{1} + \binom{K-m}{2}(N-1) + \dots \\ &\quad + \binom{K-m}{K-m}(N-1)^{K-m-1} && (50) \\ &= \frac{1}{N-1}\Bigg[\binom{K-m}{1}(N-1) + \binom{K-m}{2}(N-1)^2 + \dots \\ &\quad + \binom{K-m}{K-m}(N-1)^{K-m}\Bigg] && (51) \\ &= \frac{1}{N-1}\left(N^{K-m} - 1\right), && (52) \end{aligned}$$

where $\binom{K-m}{r}$ in (50) 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,

$$\begin{aligned} q &= \binom{(N-1)m}{1} + \binom{(N-1)m}{2}(N-1) + \dots \\ &\quad + \binom{(N-1)m}{(N-1)m}(N-1)^{(N-1)m-1} && (53) \\ &= \frac{1}{N-1}\Bigg[\binom{(N-1)m}{1}(N-1) + \dots \\ &\quad + \binom{(N-1)m}{(N-1)m}(N-1)^{(N-1)m}\Bigg] && (54) \\ &= \frac{1}{N-1}\left(N^{(N-1)m} - 1\right), && (55) \end{aligned}$$

where $\binom{(N-1)m}{r}$ in (53) 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,

$$\begin{aligned} L &= N\Bigg[\binom{K-m-1}{0} + \binom{K-m-1}{1}(N-1) + \dots \\ &\quad + \binom{K-m-1}{K-m-1}(N-1)^{K-m-1}\Bigg] && (56) \\ &= N \times N^{K-m-1} = N^{K-m}, && (57) \end{aligned}$$

where $\binom{K-m-1}{r-1}$ in (56) 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,

$$\frac{D}{L} = \frac{N(p-q)}{L} \qquad (58)$$

$$= \frac{\frac{N}{N-1}\left(N^{K-m}-1\right) - \frac{N}{N-1}\left(N^{(N-1)m}-1\right)}{N^{K-m}} \tag{59}$$

$$= \frac{N}{N-1} \times \frac{N^{K-m} - N^{(N-1)m}}{N^{K-m}} \tag{60}$$

$$= \frac{1}{1-\frac{1}{N}} \times \left[1-\left(\frac{1}{N}\right)^{K-M}\right], \tag{61}$$

which matches the lower bound in (12).

Remark 7 *Note that although our achievable scheme and the scheme in [32] are both using MDS coding to exploit the available side information, the field size requirements for realizing the MDS codes are different. For the scheme of [32], a* $(2\tilde{p}-\tilde{q}, \tilde{p})$ *MDS code is used, where* $\tilde{p} = \frac{1}{N-1}(N^K - 1)$ *and* $\tilde{q} = \frac{1}{N-1}(N^M - 1)$*. This requires larger field size than the* $(2p-q, p)$ *MDS code used in our scheme (if* $\frac{M}{N} \in \mathbb{N}$*), since* $2\tilde{p}-\tilde{q} > (2p-q)$*.*

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 nth 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 [29] and no knowledge in [30]–[32]. 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 nth 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 [32] 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. 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 [32]. 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,” *J. ACM*, vol. 45, no. 6, pp. 965–981, 1998.

[2] W. Gasarch, “A survey on private information retrieval,” in *Proc. Bull. EATCS*, 2004, pp. 72–107.

[3] C. Cachin, S. Micali, and M. Stadler, “Computationally private information retrieval with polylogarithmic communication,” in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn.*, 1999, pp. 402–414.

[4] R. Ostrovsky and W. Skeith III, “A survey of single-database private information retrieval: Techniques and applications,” in *Proc. Int. Workshop Public Key Cryptogr.*, 2007, pp. 393–411.

[5] S. Yekhanin, “Private information retrieval,” *Commun. ACM*, vol. 53, no. 4, pp. 68–73, 2010.

[6] N. B. Shah, K. V. Rashmi, and K. Ramchandran, “One extra bit of download ensures perfectly private information retrieval,” in *Proc. IEEE ISIT*, Jun. 2014, pp. 856–860.

[7] G. Fanti and K. Ramchandran, “Efficient private information retrieval over unsynchronized databases,” *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 7, pp. 1229–1239, Oct. 2015.

[8] T. Chan, S. Ho, and H. Yamamoto, “Private information retrieval for coded storage,” in *Proc. IEEE ISIT*, Jun. 2015, pp. 2742–2846.

[9] A. Fazeli, A. Vardy, and E. Yaakobi, “Codes for distributed PIR with low storage overhead,” in *Proc. IEEE ISIT*, Jun. 2015, pp. 2852–2856.

[10] R. Tajeddine and S. El Rouayheb, “Private information retrieval from MDS coded data in distributed storage systems,” in *Proc. IEEE ISIT*, Jul. 2016, pp. 7081–7093.

[11] H. Sun and S. A. Jafar, “The capacity of private information retrieval,” in *Proc. IEEE Globecom*, Dec. 2016, pp. 1–6.

[12] 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.

[13] H. Sun and S. A. Jafar, “Blind interference alignment for private information retrieval,” in *IEEE ISIT*, Jul. 2016, pp. 560–564.

[14] 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.

[15] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb, “Private information retrieval schemes for coded data with arbitrary collusion patterns,” in *Proc. IEEE ISIT*, Jun. 2017, pp. 1908–1912.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Q. Wang and M. Skoglund, “Symmetric private information retrieval for MDS coded distributed storage,” 2016, *arXiv:1610.04530*. [Online]. Available: https://arxiv.org/abs/1610.04530

[23] 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.

[24] 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.

[25] Y. Zhang and G. Ge, “A general private information retrieval scheme for MDS coded databases with colluding servers,” 2017, *arXiv:1704.06785*. [Online]. Available: https://arxiv.org/abs/1704.06785

[26] Y. Zhang and G. Ge, “Multi-file private information retrieval from MDS coded databases with colluding servers,” 2017, *arXiv:1705.03186*. [Online]. Available: https://arxiv.org/abs/1705.03186

[27] Q. Wang and M. Skoglund, “Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers,” 2017, *arXiv:1708.05673*. [Online]. Available: https://arxiv.org/abs/1708.05673

[28] Q. Wang and M. Skoglund, “Secure symmetric private information retrieval from colluding databases with adversaries,” in *Proc. IEEE Allerton*, Jun. 2017, pp. 1–5.

[29] R. Tandon, “The capacity of cache aided private information retrieval,” in *Proc. IEEE Allerton*, Oct. 2017, pp. 1–6.

[30] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, “Private information retrieval with side information,” 2017, *arXiv:1709.00112*. [Online]. Available: https://arxiv.org/abs/1709.00112

[31] 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.

[32] Z. Chen, Z. Wang, and S. A. Jafar, “The capacity of private information retrieval with private side information,” 2017, *arXiv:1806.01253*. [Online]. Available: https://arxiv.org/abs/1806.01253

[33] M. Karmoose, L. Song, M. Cardone, and C. Fragouli, "Private broadcasting: An index coding approach," 2017, *arXiv:1701.04958*. [Online]. Available: https://arxiv.org/abs/1701.04958

[34] M. Karmoose, L. Song, M. Cardone, and C. Fragouli, "Preserving privacy while broadcasting: *k*-limited-access schemes," 2017, *arXiv:1705.08437*. [Online]. Available: https://arxiv.org/abs/1705.08437

Yi-Peng Wei (S'15) received his B.Sc. in Electrical Engineering from National Tsing Hua University, Taiwan, in 2009, M.Sc. in Graduate Institute of Communication Engineering from National Taiwan University, Taiwan, in 2012, and Ph.D. degree in electrical engineering from the University of Maryland at College Park, MD, USA, in 2019 with his Ph.D. thesis on private information retrieval with side information. In 2019, he joined Google as a software engineer.

Karim Banawan (S'13–M'18) received the B.Sc. and M.Sc. degrees, with highest honors, in electrical engineering from Alexandria University, Alexandria, Egypt, in 2008, 2012, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park, MD, USA, in 2017 and 2018, respectively, with his Ph.D. thesis on private information retrieval and security in networks. He was the recipient of the Distinguished Dissertation Fellowship from the Department of Electrical and Computer Engineering, at the University of Maryland College Park, for his Ph.D. thesis work. 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 and private information retrieval.

Sennur Ulukus (S'90–M'98–SM'15–F'16) is the Anthony Ephremides Professor in Information Sciences and Systems in the Department of Electrical and Computer Engineering at the University of Maryland at 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. She received her Ph.D. degree in Electrical and Computer Engineering from Wireless Information Network Laboratory (WINLAB), Rutgers University, and B.S. and M.S. degrees in Electrical and Electronics Engineering from Bilkent University. 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, distributed coded computation, energy harvesting communications, physical layer security, and wireless energy and information transfer.

Dr. Ulukus is a fellow of the IEEE, and a Distinguished Scholar-Teacher of 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, an 2005 NSF CAREER Award, the 2010-2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Outstanding Teaching Award. She is a Distinguished Lecturer of the IEEE Information Theory Society for 2018-2019. She is on the Editorial Board of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING since 2016. She was an Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS SERIES ON GREEN COMMUNICATIONS AND NETWORKING (2015-2016), IEEE TRANSACTIONS ON INFORMATION THEORY (2007-2010), and IEEE TRANSACTIONS ON COMMUNICATIONS (2003-2007). She was a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (2015 and 2008), *Journal of Communications and Networks* (2012), and IEEE TRANSACTIONS ON INFORMATION THEORY (2011). She is a TPC co-chair of 2019 ITW, 2017 IEEE ISIT, 2016 IEEE Globecom, 2014 IEEE PIMRC, and 2011 IEEE CTW.