Private Information Retrieval with Side Information: the Single Server Case

Swanand Kadhe, Brenden Garcia, Anoosheh Heidarzadeh, Salim El Rouayheb, and Alex Sprintson

***Abstract*— We study the problem of Private Information Retrieval (PIR) in the presence of prior side information. The problem setup includes a database of K independent messages possibly replicated on several servers, and a user that needs to retrieve one of these messages. In addition, the user has some prior side information in the form of a subset of M messages, not containing the desired message and unknown to the servers. This problem is motivated by practical settings in which the user can obtain side information opportunistically from other users or has previously downloaded some messages using classical PIR schemes. The objective of the user is to retrieve the required message without revealing its identity while minimizing the amount of data downloaded from the server.**

We focus on achieving information-theoretic privacy in two scenarios: (i) the user wants to protect jointly its demand and side information; (ii) the user wants to protect only the information about its demand, but not the side information. To highlight the role of side information, we focus on the case of a single server. We prove that, in the first scenario, the minimum download cost is $K-M$ messages, and in the second scenario, it is $\lceil \frac{K}{M+1} \rceil$ messages. This is a significant improvement compared to the minimum cost of K messages in the setting where the user has no side information. Our proof techniques use a reduction from the PIR with side information problem to an index coding problem. We leverage this reduction to prove converse results, as well as to design achievability schemes.

I. Introduction

Consider the following Private Information Retrieval (PIR) setting first studied in [1], [2]: a user wishes to privately download a message belonging to a database with copies stored on a single or multiple remote servers, without revealing which message it is requesting. In a straightforward PIR scheme, the user would download all the messages in the database. However, this scheme may not be feasible due to the its high communication cost. In the case of a single server (i.e., there is only one copy of the database), it can be shown that downloading the whole database is necessary to achieve perfect privacy in an information-theoretic sense. If computational (cryptographic) privacy is desired, then PIR schemes with lower communication overhead do exist [3], [4], but they do not offer information-theoretic privacy guarantees and usually have high computational complexity. In contrast, in this paper, we design and analyze schemes that achieve information-theoretic privacy.

Interestingly, more efficient PIR schemes, achieving perfect information-theoretic privacy, can be constructed when the database is replicated on multiple servers with restriction on the servers' collusion. This replication-based model has been the one that is predominantly studied in the PIR literature, with breakthrough results in the past few years (e.g., [5]–[10]). Recently, there has been a renewed interest in PIR for the case in which the data is stored on the servers using erasure codes, which result in better storage overhead compared to the traditional replication techniques [11]–[18].

In this paper, we study the PIR problem when the user has prior side information about the database. In particular, we assume that the user already has a random subset of the database messages that is unknown to the server(s)[1]. This side information could have been obtained in several ways. For example, the user could have obtained these messages opportunistically from other users in its network, overheard them from a wireless broadcast channel, or downloaded them previously through classical PIR schemes. The next example illustrates how this side information could be leveraged to devise efficient PIR schemes. In particular, the following example shows that perfect information-theoretic privacy can be achieved with a single server case without having to download the entire database.

Example 1 (single-server PIR with side information)**.** *Consider a remote server that has a database formed of an even number of binary messages denoted by $X_1, \ldots, X_K$ of equal length. A user wants to download one of these messages from the server without revealing to the server which one. Moreover, the user has one message as side information chosen uniformly at random among all the other messages and unknown to the server. We propose two PIR schemes that leverage the side information and compare them to the straightforward scheme that downloads all the K messages.*

1) Maximum Distance Separable (MDS) PIR scheme. *This scheme downloads $K-1$ messages. The user sends to the server the number of messages in its side information (one in this example). The server responds by coding all the messages using a $(2K-1, K)$ systematic MDS code and sending the $K-1$ parity symbols of the code. Therefore, the user can always*

Swanand Kadhe, Brenden Garcia, Anoosheh Heidarzadeh, and Alex Sprintson are with the Department of Electrical and Computer Engineering at Texas A&M University, USA; emails:{swanand.kadhe, brendengarcia, anoosheh, spalex}@tamu.edu.

Salim El Rouayheb is with ECE Department at Rutgers University, email: sye8@soe.rutgers.edu. Part of this work was done while he was with the ECE department at the Illinois Institute of Technology.

The work of S. El Rouayheb was supported in part by NSF Grant CCF 1652867 and ARL Grant W911NF-17-1-0032.

[1]We assume that this side information subset does not contain the desired message. Otherwise, the problem is degenerate.

decode all the messages using its side information and the coded messages received from the server.

2) Partition and Code PIR scheme. *This scheme downloads $K/2$ messages. Suppose the message the user wants is X_W and the one in its side information is X_S for some $W, S \in \{1, \ldots, K\}$, $W \neq S$. The user chooses a random partition of $\{1, \ldots, K\}$ formed only of sets of size 2 and containing $\{W, S\}$, and sends indices of all pairs in the partition to the server. The server sends back the XOR of the messages indexed by each set. For example, suppose $W = 1$ and $S = 2$, i.e, the user wants X_1 and has X_2 as side information. The user chooses a random partition $\{\{i_1, i_2\}, \{i_3, i_4\}, \ldots, \{i_{K-1}, i_K\}\}$ and sends it to the server. The partition is chosen such that $\{1, 2\}$ is one of the sets in the partition (i.e., $i_j = 1$ and $i_{j+1} = 2$ for some $j \in \{1, 3, \ldots, K-1\}$. The server responds with $X_{i_1} + X_{i_2}, \ldots, X_{i_{K-1}} + X_{i_K}$. The user can always decode because it always receives $X_W + X_S$. Intuitively, perfect privacy is achieved here because the index of the desired message can be in any set of the partition. Moreover, in each set it could be either one of messages in the set, since the server does not know the index of the side information.* ■

We will show later that the two schemes above are optimal but achieve different privacy constraints. The MDS PIR scheme protects both the indices of the desired message and that of the side information, whereas the Partition and Code PIR scheme protects only the former.

A. Our Contributions

We consider the PIR with side information problem as illustrated in Example 1. A user wishes to download a message from a set of K messages that belong to a database stored on a single remote server.

The user has a random subset of M messages as side information. The identity of the messages in this subset is unknown to the server. We focus on PIR schemes that achieve information-theoretic privacy. Our goal is to minimize the download cost, which dominates the total communication cost (download plus upload) for large message sizes. Under this setting, we distinguish between two types of privacy constraints:

(i) hiding the identity of both the requested message and the side information from the server;
(ii) hiding only the identity of the desired message.

The latter, and less stringent, privacy constraint is justified when the side information is obtained opportunistically given that it is random and assumed to be independent of the user's request. In the case in which the side information messages were obtained previously through PIR, this constraint implies that the identity of these messages may be leaked to the server(s). However, this type of privacy can still be relevant when privacy is only desired for a certain duration of time, i.e., when the user is ambivalent about protecting the identity of messages downloaded as long as it has happened far enough in the past.

We focus on the single server scenario as the canonical case to understand the role of side information in PIR. We characterize the capacity of PIR with side information in the case of a single server for the two privacy constraints mentioned above. We show that when protecting both the side information and the request, the minimum download rate[2] for PIR is $(K - M)^{-1}$, and this can be achieved by a generalization of the MDS PIR scheme in Example 1. Moreover, we show that when only the privacy of the demand is required, the minimum download rate is $\lceil \frac{K}{M+1} \rceil^{-1}$, and this can be achieved by a generalization of the Partition and Code PIR scheme in Example 1. We present achievability and converse proofs that use a reduction to an Index Coding problem.

B. Related Work

The initial work on PIR in [1], [2] and in the literature that followed focused on designing PIR schemes for replicated data that have efficient communication cost accounting for both the size of the user queries and the servers' responses. PIR schemes with communication cost that is subpolynomial in the number of messages were devised in [9] and [19]. Information-theoretic bounds on the download rate (servers' responses) and achievable schemes were devised in [5] and [6]. Recently, there has been a growing body of work studying PIR for coded data motivated by lower overhead of codes [11]–[18], [20], [21].

The role of side information in improving PIR schemes has so far received little attention in the literature. The closest work to ours is the concurrent work of Tandon [22] in which the capacity of cache-aided PIR is characterized. The main difference with the model in [22] is our assumption that the indices of the side information messages are unknown to the servers, as is the case in the scenarios mentioned above. This lack of knowledge at the servers can be leveraged to reduce the communication cost of PIR even in the case of a single server. We also restrict our study to side information that is subset of the data, whereas the cache model in [22] allows any function of the data. Another related line of work is that of private broadcasting by Karmoose et al. [23], which considers the index coding setting with multiple users with side information and a single server. Here too, the server does know the content of the side information at the users. Moreover, the privacy constraint is to protect the request and side information of a user from the other users through a carefully designed encoding matrix. In contrast, the goal of our scheme is to protect the identity of the requested data from the server. We also note that the case in which the side information is unknown at the server was also considered in the index coding literature under the name of blind index

[2]The download rate is defined as the inverse of the normalized download cost i.e., the ratio of message length to the total size of downloaded information. (in bits)

coding [24]. However, the goal there was to minimize the broadcast rate without privacy constraints.

II. Problem Formulation and Main Results

For a positive integer K, denote $\{1,\dots,K\}$ by $[K]$. For a set $\{X_1,\dots,X_K\}$ and a subset $S \subset [K]$, let $X_S = \{X_j : j \in S\}$. For a subset $S \subset [K]$, let $\mathbf{1}_S$ denote the characteristic vector of the set S, which is a binary vector of length K such that, for all $j \in [K]$, its j-th entry is 1 if $j \in S$, otherwise it is 0. Let $\mathbb{F}_q$ denote the finite field of order q.

We assume that the database consists of a set of K messages $X = \{X_1,\dots,X_K\}$, with each message being independently and uniformly distributed over $\mathbb{F}_{2^t}$ (i.e., each message X_j is t bits long). We also assume that there are $N \geq 1$ non-colluding servers which store identical copies of the K messages.

A user is interested in downloading a message X_W for some $W \in [K]$. We refer to W as the *demand index* and X_W as the demand. The user has the knowledge of a subset X_S of the messages for some $S \subset [K]$, $|S| = M$, $M < K$. We refer to S as the *side information index set* and X_S as the *side information*.

Let $\mathbf{W}$ and $\mathbf{S}$ denote the random variables corresponding to the demand index and the side information index set. We restrict our attention to the class of distributions $p_{\mathbf{W}}(\cdot)$ of $\mathbf{W}$ such that $p_{\mathbf{W}}(W) > 0$ for every $W \in [K]$.

An important distribution of $\mathbf{W}$ and $\mathbf{S}$ that we focus on in this work is as follows. Let the demand index W be distributed uniformly over $[K]$, i.e.,

$$p_{\mathbf{W}}(W) = \frac{1}{K}, \tag{1}$$

for all $W \in [K]$. Further, let the side information index set S have the following conditional distribution given W:

$$p_{\mathbf{S}|\mathbf{W}}(S|W) = \begin{cases} \frac{1}{\binom{K-1}{M}}, & \text{if } W \notin S \text{ and } |S| = M, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We note that this implies the following joint distribution on $(\mathbf{W}, \mathbf{S})$:

$$p_{\mathbf{W},\mathbf{S}}(W,S) = \begin{cases} \frac{1}{(K-M)\binom{K}{M}}, & W \notin S, |S| = M, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We assume that the servers do not know the side information realization at the user and only know the *a priori* distributions $p_{\mathbf{W}}(W)$ and $p_{\mathbf{S}|\mathbf{W}}(S|W)$.

To download the message X_W given the side information X_S, the user sends a query $Q^{[W,S]}$ from an alphabet $\mathcal{Q}$ to the server. The server responds to the query it receives with an answer $A^{[W,S]}$ over an alphabet $\mathcal{A}$. We refer to the set of query and answer as the *PIR with side information (PIR-SI) scheme*. Our focus in this paper is on non-interactive (single round) schemes. A PIR-SI scheme should satisfy the following requirements.

1. The query $Q^{[W,S]}$ to the server is a (potentially stochastic) function of W, S, and X_S. We assume that the answer from the server is a deterministic function of the query and the messages, i.e.,
$$H\left(A^{[W,S]} \mid Q^{[W,S]}, X_1, X_2, \cdots, X_K\right) = 0, \tag{4}$$
for all $W \in [K]$ and $S \subseteq [K] \setminus \{W\}$.
2. From the answer $A^{[W,S]}$ and the side information X_S, the user should be able to decode the desired message X_W, i.e.,
$$H\left(X_W \mid A^{[W,S]}, X_S\right) = 0, \tag{5}$$
for all $W \in [K]$, $S \subseteq [K] \setminus \{W\}$.
3. The PIR-SI scheme should guarantee privacy for the user by ensuring one of the following two conditions, referred to as W-privacy and (W,S)-privacy as defined below.

Definition 1. *W-**privacy**: The server cannot infer any information about the demand index from the query, answer, and messages i.e., we have*

$$I\left(\mathbf{W}\,;\, Q^{[\mathbf{W},\mathbf{S}]}, A^{[\mathbf{W},\mathbf{S}]}, X_1, X_2, \cdots, X_K\right) = 0. \tag{6}$$

Definition 2. *(W,S)-**privacy**: The server cannot infer any information about the demand index as well as the side information index set from the query, answer, and messages i.e., we have*

$$I\left(\mathbf{W},\mathbf{S}\,;\, Q^{[\mathbf{W},\mathbf{S}]}, A^{[\mathbf{W},\mathbf{S}]}, X_1, X_2, \cdots, X_K\right) = 0. \tag{7}$$

We refer to a PIR-SI scheme preserving W-privacy or (W,S)-privacy as W-PIR-SI or (W,S)-PIR-SI scheme, respectively.

The *rate* of a a PIR-SI scheme is defined as the ratio of the message length (t bits) to the total length of the answer (in bits) as follows:[3]

$$R = \frac{t}{H\left(A^{[W,S]}\right)}. \tag{8}$$

The *capacity* of W-PIR-SI or (W,S)-PIR-SI problem, respectively denoted by C_W or $C_{W,S}$, is defined as the supremum of rates over all W-PIR-SI or (W,S)-PIR-SI schemes for a given N, K, and M, respectively.

III. Main Results

We summarize our main results for single server case in Theorems 1 and 2, which characterize the capacity W-PIR-SI and (W,S)-PIR-SI, respectively.

Theorem 1. *For the W-PIR-SI problem with $N = 1$ server, K messages, and side information size M, when the demand index $\mathbf{W}$ and the side information index set $\mathbf{S}$ are jointly distributed according to* (3)*, the capacity is*

$$C_W = \left\lceil \frac{K}{M+1} \right\rceil^{-1}. \tag{9}$$

[3]Note that the download rate dominates the total communication rate for large enough messages.

Our proof for Theorem 1 is based on two parts. We prove the converse in Section IV-B for any joint distribution of $(\mathbf{W}, \mathbf{S})$. Then, we construct an achievability scheme in Section IV-C for the distribution given in (3).

Theorem 2. *For the (W, S)-PIR-SI problem with $N = 1$ server storing K messages and for any arbitrary joint distribution of the demand index $\mathbf{W}$ and the side information index set $\mathbf{S}$ such that the size of $\mathbf{S}$ is equal to M, the capacity is*

$$C_{W,S} = (K - M)^{-1}. \tag{10}$$

First, we show that the capacity $C_{W,S}$ of the (W, S)-PIR-SI problem with $N = 1$ server, K messages, and size information size M is upper bounded by $(K - M)^{-1}$ for any joint distribution of the side information index set and the demand index (see Section V-A). Further, we construct a scheme based on maximum distance separable (MDS) codes, which achieves this bound for any arbitrary joint distribution of $(\mathbf{W}, \mathbf{S})$ such that the size of $\mathbf{S}$ is equal to M (see Section V-B).

IV. W-Privacy Problem

Our converse proofs for Theorems 1 and 2 in the single-server case use the following simple yet powerful observation.

Proposition 1. *Let $A^{[W,S]}$ be an answer from the server that satisfies the decodability requirement* (5) *and the W-privacy requirement* (6)*. Then, the following two conditions hold:*

1) *For each message $X_i, i = 1, \ldots, K$, there exists a subset $X_{S_i} \subseteq \{X_1, \cdots, X_K\} \setminus X_i$, with $|X_{S_i}| = M$, and a decoding function D_i satisfying $D_i\left(A^{[W,S]}, X_{S_i}\right) = X_i$.*
2) *There exists a function D_W such that $D_W\left(A^{[W,S]}, X_S\right) = X_W$.*

Proof. The first condition is implied by the W-privacy requirement. Indeed, if this was not the case, then the server would know that message X_i is not one of the messages requested by the user which, in turn, would violate the W-privacy condition (6). Note that the first condition holds under the assumption that $\mathbf{W}$ has a distribution such that $p_{\mathbf{W}}(W) > 0 \ \forall W \in [K]$.

The second condition is implied by the decodability requirement. ■

The above proposition enables us to show a relation of the PIR-SI problem with an instance of index coding with side information problem [25]–[27]. We begin with briefly reviewing the index coding problem.

A. Index Coding problem

Consider a server with K messages $X_1, \cdots, X_K$ of length t with $X_j \in \{0,1\}^t$. Consider L clients $R_1, \cdots, R_L$, $L \geq K$, where for each i, R_i is interested in one message, denoted by $X_{f(i)} \in \{X_i\}$, and knows some subset $X_{S_i} \subset \{X_i\}$ of the other messages, referred to as its side information.

An index code of length ℓ for this setting is a set of codewords in $\{0,1\}^\ell$ together with an encoding function $E : \{0,1\}^{tK} \to \{0,1\}^\ell$, and a set of L decoding functions $D_1, \cdots, D_L$ such that $D_i\left(E\left(X_1, \cdots, X_K\right), X_{S_i}\right) = X_{f(i)}$ for all $i \in [L]$ and $[X_1, \cdots, X_K] \in \{0,1\}^{tK}$. We refer to $E\left(X_1, \cdots, X_K\right)$ as a *solution* to the instance of the index coding problem.

When $L = K$ and every client requires a distinct message, the side information of all the clients can be represented by a simple directed graph $G = (V, E)$, where $V = \{1, 2, \cdots, K\}$ with the vertex i corresponding to the message X_i, and there is an arc $(i, j) \in E$ if $j \in S_i$. We denote the out-neighbors of a vertex i as $\mathcal{N}(i)$.

For a given instance of the index coding problem, the minimum encoding length ℓ as a function of message-length t is denoted as β_t, and the *broadcast rate* is defined as in [28], [29]

$$\beta = \inf_t \frac{\beta_t}{t} \tag{11}$$

B. Converse for Theorem 1

The key step of the converse is to show that for any scheme that satisfies the W-privacy constraint (6), the answer from the server must be a solution to an instance of the index coding problem that satisfies certain requirements as specified in the following lemma.

Lemma 1. *For a W-PIR-SI scheme, for any demand index W and side information index set S, the answer $A^{[W,S]}$ from the server must be a solution to an instance of the index coding problem that satisfies the following requirements:*

1) *The instance has the messages $X_1, \cdots, X_K$;*
2) *There are K clients such that each client wants to decode a distinct message from $X_1, \cdots, X_K$, and possesses a side information that includes M messages;*
3) *The client that wants X_W has the side information set X_S; for each other client the side information set has M arbitrary messages from $X_1, \cdots, X_K$.*

Proof. The sets X_{S_i} mentioned in Proposition 1 can be used to construct the following instance of the Index Coding problem. The instance has the message set $X_1, \cdots, X_K$ and K clients $\{R_1, \cdots, R_k\}$ such that:

- Client R_W requires packet X_W and has the side information set X_S;
- Each other client $R_i,\ i \neq W$ requires X_i and has side information set X_{S_i}.

It is easy to verify that the instance satisfies all the conditions stated in the lemma and that $A^{[W,S]}$ is the feasible index code for this instance. ■

Note that Lemma 1 shows that the answer $A^{[W,S]}$ from the server must be a solution to an instance of the index coding problem in which the out-degree of every vertex in the corresponding side information graph G is equal to M. Next, we lower bound the broadcast rate for an index coding problem with side information graph G such that out-degree of every vertex in G is M as follows.

Lemma 2. *Let G be a directed graph on K vertices such that each vertex has out-degree M. Then, the broadcast rate of the corresponding instance of the index coding problem is lower bounded by $\lceil \frac{K}{M+1} \rceil$.*

Proof. For any side information graph G, the broadcast rate β is lower bounded by the size of the maximum acyclic induced subgraph (MAIS) of G, denoted as $MAIS(G)$ [28], [30].

We show that for any graph G that satisfies the conditions of the lemma (i.e., the out-degree of each of the K vertices of G is M) it holds that

$$MAIS(G) \geq \left\lceil \frac{K}{M+1} \right\rceil.$$

Specifically, we build an acyclic subgraph of G induced by set Z through the following procedure:

Step 1. Set $Z = \emptyset$ and a candidate set of vertices $V' = V$;
Step 2. Add an arbitrary vertex $i \in V'$ into Z, i.e., $Z = Z \cup \{i\}$;
Step 3. Set $V' = V' \setminus (\mathcal{N}(i) \cup \{i\})$;
Step 4. There are two cases:
  Case 1: If $V' \neq \emptyset$, then repeat Steps 2-4.
  Case 2: If $V' = \emptyset$, then terminate the procedure and return Z.

It is easy to see that the vertices in set Z returned by the procedure induce an acyclic subgraph of G. If the vertices are ordered in the order they are added to Z, then there can only be an edge (i, j) if j was added to Z before i. This implies that the subgraph induced by Z cannot contain a cycle.

Further, note that the set Z contains at least $\lceil \frac{K}{M+1} \rceil$ vertices. At each removal step, there are at most $M + 1$ vertices removed from V. Thus, the procedure iterates at least $\lceil \frac{K}{M+1} \rceil$ times, and in each iteration we add one vertex to Z. This implies that the size of Z is at least $\lceil \frac{K}{M+1} \rceil$. ■

Corollary 1 (Converse of Theorem 1)**.** *For the W-PIR-SI problem with single server, K messages, and side information size M, the capacity is at most $\left\lceil \frac{K}{M+1} \right\rceil^{-1}$.*

Proof. Lemmas 1 and 2 imply that the length of the answer $A^{[W,S]}$ is at least $t \cdot \left\lceil \frac{K}{M+1} \right\rceil$ bits for any given W and S. Then, by (8), it follows that $R \leq \left\lceil \frac{K}{M+1} \right\rceil^{-1}$. ■

C. Achievability for Theorem 1

In this section, we propose a W-PIR-SI scheme for $N = 1$ server, K messages, and side information size M, which achieves the rate $\left\lceil \frac{K}{M+1} \right\rceil^{-1}$. Recall that we assume that the distribution of the demand index W and the conditional distribution of the side information index set S given W are given respectively in (1) and (2). We describe the proposed scheme, referred to as the *Partition and Code* PIR scheme, in the following.

Partition and Code PIR Scheme: Given K, M, W, and S, denote $g \triangleq \left\lceil \frac{K}{M+1} \right\rceil$. The scheme consists of the following three steps.

Step 1. The user creates a partition of the K messages into g sets. For the ease of understanding, we describe the special case of $(M + 1) \mid K$ first.

(a) Special case of $(M + 1) \mid K$: Denote $P_1 \triangleq W \cup S$. The user randomly partitions the set of messages $[K] \setminus P_1$ into $g - 1$ sets, each of size $M + 1$, denoted as $P_2, \ldots, P_g$.

(b) General case: Let $P_1, \ldots, P_g$ be a collection of g empty sets. Note that, although empty at the beginning, once constructed, the sets $P_1, \ldots, P_{g-1}$ will be of size $M+1$, and the set P_g will be of size $K-(g-1)(M+1)$. The user begins by assigning probabilities to the sets according to their sizes: the sets $P_1, \ldots, P_{g-1}$ are each assigned a probability $\frac{M+1}{K}$, and the set P_g is assigned a probability $\frac{K-(g-1)(M+1)}{K}$. Then, the user chooses a set randomly according to the assigned probabilities of the sets.

If the chosen set is a set $P \in \{P_1, \ldots, P_{g-1}\}$, then the user fills the set P with the demand index W and the side information index set S of the user. Next, it fills the remaining sets choosing one index at a time from the set of indices of the remaining messages uniformly at random until all the message indices are filled.

If the chosen set is the set P_g, then it fill P_g with the demand index W, and fill the remaining $K - (g - 1)(M + 1) - 1$ places in the set P_g with randomly chosen elements from the side information index set S. (Note that once P_g is filled, it is possible that not all of the indices in the side information index set S are placed in the set.) Next, fill the remaining sets by choosing one index at a time from the set of indices of the unplaced packets uniformly at random until all packet indices are placed.

Step 2. The user sends to the server a uniform random permutation of the partition $\{P_1, \cdots, P_g\}$, ie., it sends $\{P_1, \cdots, P_g\}$ in a random order.

Step 3. The server computes the answer $A^{[W,S]}$ as a set of g inner products given by $A^{[W,S]} = \{A_{P_1}, \ldots, A_{P_g}\}$, where $A_P = [X_1, \ldots, X_K] \cdot \mathbf{1}_P$ for all $P \in \{P_1, \ldots, P_g\}$.

Upon receiving the answer from the server, the user decodes X_W by subtracting off the contributions of its side information X_S from A_P for some $P \in \{P_1, \ldots, P_g\}$ such that $W \in P$.

Example 2. *Assume that $K = 8$ and $M = 2$. Assume that the user demands the message X_2 and has two messages X_4 and X_6 as side information, i.e., $W = 2$ and $S = \{4, 6\}$. Following the Partition and Code PIR scheme, the user labels three sets as P_1, P_2, and P_3, and assigns probability $\frac{3}{8}$ to each of the two sets P_1 and P_2, and assigns probability $\frac{2}{8}$ to the set P_3. Next, the user chooses one of these sets at random according to the assigned probabilities. Assume the user has chosen the set P_3. The user then places 2 into the set P_3, and chooses another element from $\{4, 6\}$ uniformly at random to place in P_3 as well. Say the user chooses 6 from the set $\{4, 6\}$, then the set P_3 becomes $P_3 = \{2, 6\}$. Then the*

user fills the other sets P_1 and P_2 randomly to exhaust the elements from $\{1,2,3,5,7,8\}$. Say the user chooses $P_1 = \{1,7,8\}$ and $P_2 = \{3,4,5\}$. Then the user sends to the server a random permutation of $\{\mathbf{1}_{P_1}, \mathbf{1}_{P_2}, \mathbf{1}_{P_3}\}$ as the query $Q^{[2,\{4,6\}]}$. The server sends three coded packets back to the user: $Y_1 = X_1 + X_7 + X_8$, $Y_2 = X_3 + X_4 + X_5$, and $Y_3 = X_2 + X_6$. The user can decode for X_2 by computing $X_2 = Y_3 - X_6$. From the server's perspective the user's demand is in either $\{1,7,8\}$ or $\{3,4,5\}$ with probability $\frac{3}{8}$ each, or in $\{2,6\}$ with probability $\frac{2}{8}$. The probability P_1 (or P_2) contains W is $\frac{1}{3}$, and the probability that P_3 contains W is $\frac{1}{2}$. In either case, it follows that $\mathbb{P}(\mathbf{W} = W|Q^{[1,\{2,3\}]}) = \frac{1}{8} = p_{\mathbf{W}}(W)$.

In the following, we show that the Partition and Code PIR scheme satisfies the W-privacy requirement for the setting in which the user's demand index W and side information index set S (given W) are distributed according to (1) and (2), respectively.

Lemma 3 (Achievability of Theorem 1)**.** *Consider the scenario of a W-PIR-SI problem in which:*

- *The server has packets $\{X_1, X_2, ..., X_K\}$;*
- *There is one user with $|W| = 1, |S| = M$ such that $0 \leq M \leq K-1$;*
- *The demand index W and the side information index set S (given the demand index W) follow the distributions given in* (1) *and* (2)*, respectively.*

In this scenario, the Partition and Code PIR scheme satisfies the W-privacy, and has rate $R = \left\lceil \frac{K}{M+1} \right\rceil^{-1}$.

Proof. To show that the Partition and Code PIR scheme satisfies the W-privacy, it suffices to show that

$$\mathbb{P}(\mathbf{W} = W|Q^{[W,S]}) = p_{\mathbf{W}}(W).$$

We consider two cases as follows:

(i) W is in one of the sets in $\{P_1, \ldots, P_{g-1}\}$. In this case, for every $i \in [g-1]$, we have

$$\begin{aligned}\mathbb{P}(\mathbf{W} \in P_i|Q^{[W,S]}) &= \sum_{j \in P_i} \mathbb{P}(\mathbf{W} = j|Q^{[W,S]}) \\ &= \frac{M+1}{K},\end{aligned}$$

and

$$\mathbb{P}(\mathbf{W} = W|\mathbf{W} \in P_i, Q^{[W,S]}) = \frac{1}{M+1}.$$

(ii) W is the set P_g. In this case,

$$\mathbb{P}(\mathbf{W} \in P_g|Q^{[W,S]}) = \frac{K-(g-1)(M+1)}{K},$$

and

$$\mathbb{P}(\mathbf{W} = W|\mathbf{W} \in P_g, Q^{[W,S]}) = \frac{1}{K-(g-1)(M+1)}.$$

Thus, we have

$$\begin{aligned}&\mathbb{P}(\mathbf{W} = W|Q^{[W,S]}) \\ &\quad= \sum_{i=1}^{g} \mathbb{P}(\mathbf{W} = W|\mathbf{W} \in P_i, Q^{[W,S]})\mathbb{P}(\mathbf{W} \in P_i|Q^{[W,S]}) \\ &\quad= \frac{1}{K}.\end{aligned}$$

To compute the rate of the scheme, note that

$$\begin{aligned}H(A^{[W,S]}) &= H([A_{P_1}, A_{P_2}, \ldots, A_{P_g}]) \\ &= \sum_{P \in \{P_1, P_2, \ldots, P_g\}} H(A_P) \\ &= t \times g,\end{aligned}$$

where the equalities follow since the messages X_j's (and hence the answers A_P's) are independently and uniformly distributed. Thus, the Partition and Code PIR scheme has rate

$$R = \frac{t}{t \times g} = \frac{1}{g} = \frac{M+1}{K}.$$

■

V. (W,S)-Privacy Problem

In this section we consider (W,S)-privacy in the PIR-SI problem. We show the proof of the converse and the achievability for Theorem 2 through a reduction to an index coding instance and an MDS coding scheme, respectively.

A. Converse for Theorem 2

When protecting the demand index and the side information index set of the user, the privacy constraint becomes

$$I(\mathbf{W}, \mathbf{S}; Q^{[\mathbf{W},\mathbf{S}]}, A^{[\mathbf{W},\mathbf{S}]}, X_1, X_2, ..., X_K) = 0.$$

For this case, a lower bound of $K - M$ on the number of transmissions can be shown. The proof of the converse in this case shows a necessary condition for privacy and a class of index coding problems that satisfy the necessary condition; and obtains a lower bound on the number of transmissions needed to solve the index coding problem that belong to this class.

Lemma 4. *For a (W,S)-PIR-SI scheme, for any demand index W and side information index set S, the answer $A^{[W,S]}$ from the server must be a solution to an instance of the index coding problem that satisfies the following requirements:*

1) *The instance has the message set $X_1, \cdots, X_K$;*
2) *There are $L = (K-M)\binom{K}{M}$ clients such that each client wants to decode one message, and possesses a side information set that includes M other messages;*
3) *The client that wants X_W has the side information set X_S; for each $i \in [K], i \neq W$, for each $S_i \subset [K] \setminus \{i\}$ such that $|S_i| = M$, there exists a client that demands X_i and possesses X_{S_i} as its side information.*

Proof. Given a demand index W and a side information index set S, let $A^{[W,S]}$ be an answer from the server that

satisfies the decodability requirement (5) and the (W, S)-privacy requirement (7). First, we note that the decodability requirement implies that there exists a function $D_{W,S}$ such that $D_{W,S}\left(A^{[W,S]}, X_S\right) = X_W$. Second, we note that the (W, S)-privacy requirement implies that for each message X_i and every set $S_i \subseteq [K]\backslash\{i\}$ of size M, there exists a function D_{i,S_i} satisfying $D_{i,S_i}\left(A^{[W,S]}, X_{S_i}\right) = X_i$. Otherwise, for a particular $\{i, S_i\}$, the server would know that the user cannot possess X_{S_i} and demand X_i, which, in turn, would violate the (W, S)-privacy requirement (7).

Now, consider an instance of the index coding problem satisfying the conditions stated in the lemma. Since decoding functions exists for each client as argued above, $A^{[W,S]}$ is a feasible index code for this instance. ■

Next, we give a lower bound on the broadcast rate for an instance satisfying the conditions in Lemma 4.

Lemma 5. *For any instance of the index coding problem satisfying the conditions specified in Lemma 4, the broadcast rate is at least $K - M$.*

Proof. Let J denote an instance of the index coding problem satisfying the conditions in Lemma 4. Let J' be an instance of the index coding problem with the K messages $X_1, \cdots, X_K$ and $K - M$ clients. Each client has the side information X_S and wants to decode one distinct message from $\{X_1, \cdots, X_K\} \setminus X_S$. Clearly, a solution to instance J is also a solution to instance J'. Since the messages are independent, the broadcast rate for J' is at least $K - M$, which completes the proof. ■

Corollary 2 (Converse of Theorem 2)**.** *For the (W, S)-PIR-SI problem with $N = 1$ server, K messages, and side information size M, the capacity is at most $(K - M)^{-1}$.*

Proof. Lemmas 4 and 5 imply that the length of the answer $A^{[W,S]}$ is at least $(K - M)t$ for any given W and S. Thus, by using (8), it follows that $R \leq (K - M)^{-1}$. ■

B. Achievability for Theorem 2

In this section, we give a (W, S)-PIR-SI scheme based on a maximum distance separable (MDS) code that achieves the rate of $1/(K - M)$. We assume that $t \geq \log_2(2K - M)$.

MDS PIR Scheme: Given a demand index W and a side information index set S of size M, the user queries the server to send the $K - M$ parity symbols of a systematic $(2K - M, K)$ MDS code over the finite field $\mathbb{F}_{2^t}$. We assume that $t \geq \log_2(2K - M)$, or equivalently, $2^t \geq 2K - M$. Thus, it is possible construct a $(2K - M, K)$ MDS code over $\mathbb{F}_{2^t}$. The answer $A^{[W,S]}$ from the server consists of the $K - M$ parity symbols.

Lemma 6 (Achievability of Theorem 2)**.** *The MDS PIR scheme satisfies the decodability condition in* (5) *and the (W, S)-privacy condition in* (7)*, and it has the rate of $R = (K - M)^{-1}$.*

Proof. (Sketch) For a $(2K - M, K)$ systematic MDS code, given the $K - M$ parity symbols and any M out of the K messages, the user can decode all of the remaining $K - M$ messages as the code is MDS. Thus, the user can recover its demanded message.

To ensure the (W, S)-privacy, note that the query and the answer are independent of the particular realization of demand index W and side information index set S, but only depend on the size M of the side information index set. As the server already knows the size of the side information index set, it does not get any other information about W and S from the query and the answer. Thus, the MDS PIR scheme satisfies the (W, S)-privacy requirement.

To compute the rate, note that for any W and S, the answer $A^{[W,S]}$ of the MDS PIR scheme consists of $K - M$ parity symbols of a $(2K - M, K)$ systematic MDS code over $\mathbb{F}_{2^t}$. For an MDS code, any parity symbol is a linear combination of all the messages. Thus, as each message is distributed uniformly over $\mathbb{F}_{2^t}$ and the code operates over $\mathbb{F}_{2^t}$, every parity symbol is also uniformly distributed over $\mathbb{F}_{2^t}$. Further, since the messages are independent, the parity symbols are independent. Hence, we have $H(A^{[W,S]}) = (K - M)t$. Therefore, the rate of the MDS PIR scheme is $R = (K - M)^{-1}$. ■

VI. Conclusion

In this paper we considered the problem of Private Information Retrieval (PIR) with side information, in which the user has *a priori* a subset of the messages at the server obtained from other sources. The goal of the user is to download a message, which is not in its side information, from the server while satisfying a certain privacy constraints. We consider two privacy requirements: W-privacy in which the user wants to protect the identity its demand (i.e., which message it wants to download), and (W, S)-privacy in which the user wants to protect the identity of the demand and the side information jointly. We focus on the case of single server. We establish the PIR capacity for (W, S)-privacy for arbitrary distribution of the demand index W and the side information index set S. In the case of W-privacy, we establish the PIR capacity for the uniform distribution.

PIR with side information is an interesting research area that gives rise to many new research problems. In particular, as part of future work we plan to investigate W-privacy in the multi-server scenario. In addition, we plan to develop efficient schemes for settings in which multiple messages are requested by the client, as well as the settings with coded side information sets.

References

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *IEEE Symposium on Foundations of Computer Science*, 1995, pp. 41–50.

[2] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 965–981, 1998.

[3] E. Kushilevitz and R. Ostrovsky, "Replication is not needed: Single database, computationally-private information retrieval," in *IEEE Symposium on Foundations of Computer Science*, 1997, p. 364.

[4] C. Cachin, S. Micali, and M. Stadler, "Computationally private information retrieval with polylogarithmic communication," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 1999, pp. 402–414.

[5] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *arXiv preprint arXiv:1602.09134*, 2016.

[6] ——, "The capacity of robust private information retrieval with colluding databases," *arXiv preprint arXiv:1605.00635*, 2016.

[7] S. Yekhanin, "Private information retrieval," *Communications of the ACM*, vol. 53, no. 4, pp. 68–73, 2010.

[8] A. Beimel and Y. Ishai, "Information-theoretic private information retrieval: A unified construction," in *Automata, Languages and Programming*. Springer, 2001, pp. 912–926.

[9] A. Beimel, Y. Ishai, E. Kushilevitz, and J.-F. Raymond, "Breaking the $O(n^{1/(2k-1)})$ barrier for information-theoretic private information retrieval," in *43rd Annual IEEE Symposium on Foundations of Computer Science*, 2002, pp. 261–270.

[10] W. Gasarch, "A survey on private information retrieval," in *Bulletin of the EATCS*. Citeseer, 2004.

[11] N. Shah, K. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *2014 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2014, pp. 856–860.

[12] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," *arXiv preprint arXiv:1410.5489*, 2014.

[13] R. Tajeddine and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *2016 IEEE International Symposium on Information Theory Proceedings (ISIT)*, 2016.

[14] R. Tajeddine, S. El Rouayheb, "Private Information Retrieval from MDS Coded data in Distributed Storage Systems (extended version)," 2016, http://www.ece.iit.edu/ salim/PIRv2.pdf.

[15] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *arXiv preprint arXiv:1609.08138*, 2016.

[16] A. Fazeli, A. Vardy, and E. Yaakobi, "Pir with low storage overhead: Coding instead of replication," *arXiv preprint arXiv:1505.06241*, 2015.

[17] S. Blackburn and T. Etzion, "PIR array codes with optimal pir rate," *arXiv preprint arXiv:1607.00235*, 2016.

[18] R. Freij-Hollanti, O. Gnilke, C. Hollanti, and D. Karpuk, "Private information retrieval from coded databases with colluding servers," *arXiv preprint arXiv:1611.02062*, 2016.

[19] Z. Dvir and S. Gopi, "2-Server PIR with subpolynomial communication," *Journal of the ACM (JACM)*, vol. 63, no. 4, p. 39, 2016.

[20] R. Tajeddine and S. El Rouayheb, "Robust private information retrieval on coded data," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017.

[21] 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 IEEE International Symposium on Information Theory (ISIT)*, 2017.

[22] R. Tandon, "The capacity of cache aided private information retrieval," *ArXiv e-prints*, June 2017. [Online]. Available: https://arxiv.org/abs/1706.07035

[23] M. Karmoose, L. Song, M. Cardone, and C. Fragouli, "Private broadcasting: an index coding approach," *CoRR*, vol. abs/1701.04958, 2017. [Online]. Available: http://arxiv.org/abs/1701.04958

[24] D. T. Kao, M. A. Maddah-Ali, and A. S. Avestimehr, "Blind index coding," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2076–2097, 2017.

[25] Z. Bar-Yossef, Y. Birk, T. S. Jayram, and T. Kol, "Index coding with side information," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1479–1494, March 2011.

[26] M. Effros, S. El Rouayheb, and M. Langberg, "An equivalence between network coding and index coding," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2478–2487, 2015.

[27] S. El Rouayheb, A. Sprintson, and C. Georghiades, "On the index coding problem and its relation to network coding and matroid theory," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3187–3195, 2010.

[28] N. Alon, E. Lubetzky, U. Stav, A. Weinstein, and A. Hassidim, "Broadcasting with side information," in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, Oct 2008, pp. 823–832.

[29] A. Blasiak, R. Kleinberg, and E. Lubetzky, "Broadcasting with side information: Bounding and approximating the broadcast rate," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5811–5823, Sept 2013.

[30] F. Arbabjolfaei, "Index coding: Fundamental limits, coding schemes, and structural properties," in *PhD Thesis, UC San Diego*, 2017.