Private Set Intersection Using Multi-Message Symmetric Private Information Retrieval

Zhusheng Wang[1], Karim Banawan[2], and Sennur Ulukus[1]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
[2]Electrical Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Abstract*—We study the problem of private set intersection (PSI). In PSI, there are two entities, each storing a set $\mathcal{P}_i$, whose elements are picked from a finite set $\mathbb{S}_K$, on N_i replicated and non-colluding databases. It is required to determine the set intersection $\mathcal{P}_1 \cap \mathcal{P}_2$ without leaking any information about the remaining elements to the other entity. We first show that the PSI problem can be recast as a multi-message symmetric private information retrieval (MM-SPIR) problem. Next, as a stand-alone result, we show that the exact capacity of MM-SPIR is $C_{MM-SPIR} = 1 - \frac{1}{N}$ when $P \leq K - 1$, if the common randomness S satisfies $H(S) \geq \frac{P}{N-1}$ per desired symbol. This result implies that there is no gain for MM-SPIR over successive single-message SPIR. We present a novel capacity-achieving scheme which builds seamlessly over the multi-message PIR (MM-PIR) scheme. Based on this capacity result for the MM-SPIR problem, we show that the optimal download cost for the PSI problem is given by $\min\left\{\left\lceil\frac{P_1 N_2}{N_2-1}\right\rceil, \left\lceil\frac{P_2 N_1}{N_1-1}\right\rceil\right\}$, where P_i is the cardinality of the set $\mathcal{P}_i$.

I. Introduction

Private set intersection (PSI) refers to the problem of determining the common elements in two sets without leaking any further information about the remaining elements in the sets. PSI has been a major research topic in the field of cryptography starting with [1]. As a practical motivation for PSI, consider an airline company which has a list of its customers, and a law enforcement agency which has a list of suspected terrorists. The airline company and the law enforcement agency wish to determine the intersection of their respective lists without the airline company revealing the rest of its customers and the law enforcement agency revealing the rest of the suspects in its list (see also [2], [3]).

Since the entities in PSI want to *privately retrieve* the elements that belong to the intersection, private information retrieval (PIR) can be a building block for the PSI problem [4]. Nevertheless, it is needed to keep the remaining elements of the sets secret from the other entity. This gives rise naturally to the problem of symmetric PIR (SPIR), which was originally introduced in [5]. Recently, Sun and Jafar reformulated the problems of PIR and SPIR from an information-theoretic point of view, and determined the fundamental limits of both of these problems, in [6] and [7], respectively. Subsequently, the fundamental limits of many interesting variants of PIR and SPIR have been considered, see for example [8]–[51].

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

To use SPIR to implement PSI, the ith entity needs to privately check the presence of each element in its set at the other entity. Hence, the ith entity needs to retrieve *multiple messages* from the other entity, where the messages correspond to the *incidences* of each element of the set. This establishes the connection between PSI and *multi-message* SPIR (MM-SPIR). The MM-SPIR problem is interesting on its own right and has remained an open problem until this work. The papers that are most closely related to our work are the ones that focus on *symmetry* and *multi-message* aspects of PIR as in [7], [11]–[15], [20], [29]–[31]. None of these works considers the interplay between the data privacy constraint and the joint retrieval of multiple messages, as needed in MM-SPIR.

In this paper, first focusing on MM-SPIR as a stand-alone problem, we derive its capacity to be $C_{SM-PIR} = C_{MM-PIR} = 1 - \frac{1}{N}$. We show that the databases need to share a random variable S such that $H(S) \geq \frac{P}{N-1}$ per desired symbol. This implies that, unlike MM-PIR, there is no gain from jointly retrieving the P messages. For the extreme case $P = K$, full capacity is attained without the need for any common randomness. Further, we propose a novel capacity-achieving scheme for $1 \leq P \leq K - 1$. The query structure of the scheme resembles its counterpart in [20]. Our scheme is surprisingly optimal for all P and K in contrast to the scheme in [20] which is proved to be optimal only if P is at least half of K or K/P is an integer. By plugging $P = 1$, our scheme serves as an alternative capacity-achieving scheme for the SM-SPIR scheme in [7]. As an added advantage, our scheme extends seamlessly the MM-PIR scheme to satisfy the database privacy constraint without changing the query structure. Hence, by operating such a scheme the databases can support SPIR and PIR simultaneously.

We ultimately consider the PSI problem. There are two entities E_1 and E_2. The entity E_i has a set (list) $\mathcal{P}_i$, whose elements are picked from a finite set $\mathbb{S}_K$ and has a cardinality P_i. The set $\mathcal{P}_i$ is stored on N_i non-colluding and replicated databases. It is required to compute the intersection $\mathcal{P}_1 \cap \mathcal{P}_2$ without leaking information about $\mathcal{P}_1 \setminus \mathcal{P}_2$ or $\mathcal{P}_2 \setminus \mathcal{P}_1$ with the minimum download cost. We first show that this problem can be recast as an MM-SPIR problem, where a user needs to retrieve P messages from a library containing K messages. In this MM-SPIR problem, messages correspond to *incidences* of elements in these sets with respect to the field elements. The incidence vector is a binary vector of length K that stores a 1 in the position of the jth element of the field if this field element is in $\mathcal{P}_i$. This transforms each set into a library of K binary messages (of length 1 bit each). Therefore, in transforming the PSI problem into an MM-SPIR problem, two restrictions arise: First, the message size is 1 bit. Second, in our formulation, we restrict the set generation model to the case where the resulting messages are independent. Following these constructions, entity E_i performs MM-SPIR of the messages corresponding to its set $\mathcal{P}_i$ within the databases of the other entity. By decoding these messages, the intersection $\mathcal{P}_1 \cap \mathcal{P}_2$ is determined without leaking any information about $\mathcal{P}_1 \setminus \mathcal{P}_2$ or $\mathcal{P}_2 \setminus \mathcal{P}_1$. We show that the optimum download cost for the PSI problem is $\min\left\{\left\lceil \frac{P_1 N_2}{N_2-1}\right\rceil, \left\lceil \frac{P_2 N_1}{N_1-1}\right\rceil\right\}$. The linear scaling of the download cost appears in the problem of determining the set intersection even without any privacy constraints.We only provide sketches of the proofs here due to space limitations; proof details, examples and figures can be found in [52].

II. PSI: Problem Formulation

Consider a setting with two entities E_1 and E_2 each storing a set $\mathcal{P}_i$, $i = 1, 2$. For each element of the finite set $\mathbb{S}_K$, the entity E_i adds this element to its set $\mathcal{P}_i$ independently from the remaining field elements with probability q_i. We focus on the case of $q_i = \frac{1}{2}$ for $i = 1, 2$. After generation of the set $\mathcal{P}_i$, the cardinality of $\mathcal{P}_i \subseteq \mathbb{S}_K^{P_i}$ is denoted by $|\mathcal{P}_i| = P_i$, and is public knowledge. The entity E_i stores $\mathcal{P}_i$ in a replicated fashion on N_i non-colluding databases.

The entities E_1 and E_2 want to compute the intersection $\mathcal{P}_1 \cap \mathcal{P}_2$ privately. To that end, E_1 sends the query $Q_{n_2}^{[\mathcal{P}_1]}$ to the n_2th database (which is associated to E_2) for all $n_2 \in [N_2]$, where $[N_2]$ (and also $[1 : N_2]$) denotes integers from 1 to N_2. Since E_1 does not know $\mathcal{P}_2$ in advance, it generates the queries $Q_{1:N_2}^{[\mathcal{P}_1]} = \left\{Q_{n_2}^{[\mathcal{P}_1]} : n_2 \in [N_2]\right\}$ independently from $\mathcal{P}_2$, hence,

$$I(Q_{1:N_2}^{[\mathcal{P}_1]}; \mathcal{P}_2) = 0 \tag{1}$$

The databases associated with E_2 respond with answer strings $A_{1:N_2}^{[\mathcal{P}_1]}$. The n_2th answer string $A_{n_2}^{[\mathcal{P}_1]}$ is a deterministic function of the set $\mathcal{P}_2$ and the query $Q_{n_2}^{[\mathcal{P}_1]}$, thus,

$$H(A_{n_2}^{[\mathcal{P}_1]} | Q_{n_2}^{[\mathcal{P}_1]}, \mathcal{P}_2) = 0, \quad n_2 \in [N_2] \tag{2}$$

E_1 should be able to reliably compute the intersection $\mathcal{P}_1 \cap \mathcal{P}_2$ based on the collected answer strings $A_{1:N_2}^{[\mathcal{P}_1]}$, i.e.,

$$\text{[PSI reliability]} \qquad H(\mathcal{P}_1 \cap \mathcal{P}_2 | Q_{1:N_2}^{[\mathcal{P}_1]}, A_{1:N_2}^{[\mathcal{P}_1]}) = 0 \tag{3}$$

For privacy, first, the queries sent by E_1 should not leak any information about $\mathcal{P}_1$, i.e., any individual database associated with E_2 learns nothing about $\mathcal{P}_1$ from the queries,

$$[E_1 \text{ privacy}] \qquad I(\mathcal{P}_1; Q_{n_2}^{[\mathcal{P}_1]}) = 0, \quad n_2 \in [N_2] \tag{4}$$

Second, E_1 should not learn anything further than $\mathcal{P}_1 \cap \mathcal{P}_2$ from E_2 based on the collected answer strings. Thus, E_1 should learn nothing about the information contained in the subset $(\mathcal{P}_2 \setminus \mathcal{P}_1) \cup \overline{(\mathcal{P}_1 \cup \mathcal{P}_2)} = \bar{\mathcal{P}}_1$ of E_2,

$$[E_2 \text{ privacy}] \qquad I(\bar{\mathcal{P}}_1; A_{1:N_2}^{[\mathcal{P}_1]}) = 0 \tag{5}$$

An achievable PSI scheme should satisfy the PSI reliability constraint (3), the E_1 privacy constraint (4), and the E_2 privacy constraint (5). The efficiency of the scheme is measured by the total number of downloaded bits by one of the entities E_1 or E_2 in order to compute $\mathcal{P}_1 \cap \mathcal{P}_2$, denoted by D,

$$D = \sum_{n_2=1}^{N_2} H(A_{n_2}^{[\mathcal{P}_1]}) \tag{6}$$

The optimal download cost is $D^* = \inf D$ over all achievable PSI schemes.

III. From PSI to MM-SPIR

We show that PSI can be reduced to an MM-SPIR problem, if the entities allow storing their sets in a specific searchable format. This transformation has the same flavor as [53] and [35]. This enables PIR, which assumes that the user knows the position of the desired file in the databases. Define the incidence vector $X_i \in \mathbb{F}_2^K$ as a binary vector of size K associated with the set $\mathcal{P}_i$. Denote the jth element of the incidence vector X_i by $X_i(j)$ where $X_i(j) = 1$ if $j \in \mathcal{P}_i$ for all $j \in \mathbb{S}_K$. Hence, $X_i(j)$ is an i.i.d. random variable for all $j \in \mathbb{S}_K$ such that $X_i(j) \sim \text{Ber}(q_i)$. The entity E_i constructs the incidence vector X_i corresponding to the set $\mathcal{P}_i$ and replicates X_i at all of its N_i associated databases. The PSI determination process is performed over X_1 or X_2.

Assume that E_1 initiates the PSI process. E_1 does not know M, the size of the intersection, in advance. The only information E_1 has is $\mathcal{P}_1$. Consequently, E_1 wants to verify the existence of each element of $\mathcal{P}_1$ in $\mathcal{P}_2$ to deduce $\mathcal{P}_1 \cap \mathcal{P}_2$. Thus, E_1 needs to jointly and reliably download the bits $W_{\mathcal{P}_1} = \{X_2(j) : j \in \mathcal{P}_1\}$ by sending N_2 queries to the databases associated with E_2 and collecting the corresponding answer strings. Hence, the reliability constraint is,

$$H(W_{\mathcal{P}_1} | Q_{1:N_2}^{[\mathcal{P}_1]}, A_{1:N_2}^{[\mathcal{P}_1]}) = 0 \tag{7}$$

Since E_1 is searching for the existence of all elements of $\mathcal{P}_1$ in $\mathcal{P}_2$ without leaking any information about $\mathcal{P}_1$ to any individual database associated with E_2, the E_1 privacy constraint in (4) dictates,

$$I(\mathcal{P}_1; Q_{n_2}^{[\mathcal{P}_1]}) = 0, \quad n_2 \in [N_2] \tag{8}$$

This is the privacy constraint in the MM-PIR problem [20].

To ensure the E_2 privacy constraint, the answers from E_2 databases should not leak any information contained in the subset $\bar{\mathcal{P}}_1$ of E_2, which is equivalent to not leaking any information about $W_{\bar{\mathcal{P}}_1}$,

$$I(W_{\bar{\mathcal{P}}_1}; A_{1:N_2}^{[\mathcal{P}_1]}) = 0 \tag{9}$$

This is exactly the database privacy constraint in MM-SPIR.

Consequently, the PSI problem reduces to MM-SPIR with i.i.d. messages of length 1 bit each, if the entities E_1 and E_2 are allowed to construct the corresponding incidence vectors for the original sets. In Section V, we derive in detail the capacity of the MM-SPIR problem, which in turn gives the most efficient information-theoretic PSI scheme.

IV. Main Result

Our main result provides the optimal download cost for the PSI problem under the assumptions in Sections II and III.

Theorem 1 *In the PSI problem, if the elements of the sets are added independently with probability $q_i = \frac{1}{2}$ from a finite set of size K, and if the set $\mathcal{P}_1$ where $|\mathcal{P}_1| = P_1$ is stored among N_1 databases and the set $\mathcal{P}_2$ where $|\mathcal{P}_2| = P_2$ is stored among N_2 databases, then the optimal download cost is,*

$$D^* = \min\left\{\left\lceil \frac{P_1 N_2}{N_2 - 1}\right\rceil, \left\lceil \frac{P_2 N_1}{N_1 - 1}\right\rceil\right\} \tag{10}$$

The proof of Theorem 1 is a direct consequence of the capacity result for MM-SPIR presented in Section V. We note that compared with the best-known result for PSI under computational guarantees [2], assuming that $P_1 \geq P_2$, and $N_1 = N_2$, our result $O(P_2)$ is only linear in the size of the smaller set in contrast to $O(P_2 \log P_1)$ in [2]. The linear scalability of our scheme matches the linear scalability of the best-known set intersection algorithms without any privacy constraints. Our result is private in information-theoretic (absolute) sense and does not need any assumptions about the computational powers of the entities. Furthermore, the achievable scheme is fairly simple and easy to implement compared to the fully homomorphic encryption needed in [2]. The only drawback of our approach is that it needs multiple non-colluding databases (N_1 or N_2 needs to be strictly larger than 1), otherwise, our scheme is infeasible.

V. MM-SPIR as a Stand-Alone Problem

In this section, we consider the MM-SPIR problem. We present the problem in a stand-alone format, i.e., we present a formal problem description in Section V-A, followed by the main result in Section V-B, the converse in Section V-C, and a novel achievability in Section V-D.

A. MM-SPIR: Formal Problem Description

There are N non-colluding databases each storing K i.i.d. messages. Each message is composed of L i.i.d. and uniformly chosen symbols from a sufficiently large finite field, i.e., $H(W_k) = L$ for $k \in [K]$ and $H(W_{1:K}) = KL$.

In MM-SPIR, the user needs to retrieve $W_{\mathcal{P}}$, with the desired message set $\mathcal{P}$ having size $|\mathcal{P}| = P$. Following [7], let $\mathcal{F}$ denote the private randomness possessed by the user to satisfy the user privacy constraint. A necessary common randomness S must be shared among the N databases to satisfy the database privacy constraint. S and $\mathcal{F}$ are generated independently, and independent of the message set $W_{1:K}$ without knowing the desired index set $\mathcal{P}$. Then,

$$H(\mathcal{F}, S, \mathcal{P}, W_{1:K}) = H(\mathcal{F}) + H(S) + H(\mathcal{P}) + H(W_{1:K}) \tag{11}$$

To perform MM-SPIR, a user generates a query $Q_n^{[\mathcal{P}]}$ and sends it to the nth database. Hence, the queries $Q_{1:N}^{[\mathcal{P}]}$ are deterministic functions of $\mathcal{F}$, i.e.,

$$H(Q_1^{[\mathcal{P}]}, Q_2^{[\mathcal{P}]}, \cdots, Q_N^{[\mathcal{P}]} | \mathcal{F}) = 0, \quad \forall \mathcal{P} \tag{12}$$

From (11) and (12), the queries are independent of $W_{1:K}$, i.e.,

$$I(Q_{1:N}^{[\mathcal{P}]}; W_{1:K}) = 0 \tag{13}$$

After receiving a query from the user, each database truthfully generates an answer string based on the messages and the common randomness, hence,

$$H(A_n^{[\mathcal{P}]} | Q_n^{[\mathcal{P}]}, W_{1:K}, S) = 0, \quad \forall n, \forall \mathcal{P} \tag{14}$$

After collecting all the answer strings from the N databases, the user should be able to decode $W_{\mathcal{P}}$ reliably, therefore,

$$\text{[reliability]} \quad H(W_{\mathcal{P}} | A_{1:N}^{[\mathcal{P}]}, Q_{1:N}^{[\mathcal{P}]}, \mathcal{F}) = 0 \quad \forall \mathcal{P} \tag{15}$$

The user privacy constraint can be written as,

$$\text{[user privacy]} \quad I(\mathcal{P}; Q_n^{[\mathcal{P}]}, A_n^{[\mathcal{P}]}, W_{1:K}, S) = 0, \ \forall n, \mathcal{P} \tag{16}$$

In order to protect the databases' privacy, the user should learn nothing about $W_{\bar{\mathcal{P}}}$ which is the complement of $W_{\mathcal{P}}$,

$$\text{[database privacy]} \quad I(W_{\bar{\mathcal{P}}}; Q_{1:N}^{[\mathcal{P}]}, A_{1:N}^{[\mathcal{P}]}, \mathcal{F}) = 0, \quad \forall \mathcal{P} \tag{17}$$

An achievable MM-SPIR scheme is a scheme that satisfies the MM-SPIR reliability constraint (15), the user privacy constraint (16), and the database privacy constraint (17). Following the definition of the sum retrieval rate of $W_{\mathcal{P}}$ in [20], we define the sum retrieval rate of MM-SPIR as,

$$R_{MM-SPIR} = \frac{H(W_{\mathcal{P}})}{\sum_{n=1}^{N} H(A_n^{[\mathcal{P}]})} = \frac{PL}{\sum_{n=1}^{N} H(A_n^{[\mathcal{P}]})} \tag{18}$$

The sum capacity of MM-SPIR, $C_{MM-SPIR}$, is the supremum of the sum retrieval rates over all achievable schemes.

B. MM-SPIR: Main Results

Theorem 2 *The MM-SPIR capacity for $N \geq 2$, $K \geq 2$, and $P \leq K$, is given by,*

$$C_{MM-SPIR} = \begin{cases} 1, & P = K \\ 1 - \frac{1}{N}, & 1 \leq P \leq K-1, \ H(S) \geq \frac{PL}{N-1} \\ 0, & \textit{otherwise} \end{cases} \tag{19}$$

We note that the result implies that the capacity of MM-SPIR is exactly the same as the capacity of SM-SPIR [7]. Hence, there is no gain from joint retrieval in comparison to successive single-message SPIR [7]. This in contrast to the gain in MM-PIR [20] in comparison to successive single-message PIR [6]. MM-SPIR capacity expression in Theorem 2 inherits all of the structural remarks from [7]. Furthermore, for the extreme case of $P = K$, the SPIR capacity is 1 without using any common randomness. This is due to the fact that the user privacy and the database privacy constraints are trivially satisfied, and hence the user can simply download all of the messages from one of the databases without using any common randomness.

C. MM-SPIR: Converse Proof

To prove the converse of Theorem 2, we first need the following lemmas. Lemmas 1 are 2 are direct extensions to [7, Lemmas 1 and 2] to the setting of MM-SPIR.

Lemma 1 (Symmetry) $\forall n,\ \forall \mathcal{P}_1 \neq \mathcal{P}_2$ *s.t.* $|\mathcal{P}_1| = |\mathcal{P}_2|$

$$H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, Q_n^{[\mathcal{P}_1]}) = H(A_n^{[\mathcal{P}_2]}|W_{\mathcal{P}_1}, Q_n^{[\mathcal{P}_2]}) \quad (20)$$
$$H(A_n^{[\mathcal{P}_1]}|Q_n^{[\mathcal{P}_1]}) = H(A_n^{[\mathcal{P}_2]}|Q_n^{[\mathcal{P}_2]}) \quad (21)$$

Lemma 2 (Effect of conditioning on user's randomness)

$$H(A_n^{[\mathcal{P}]}|W_{\mathcal{P}}, \mathcal{F}, Q_n^{[\mathcal{P}]}) = H(A_n^{[\mathcal{P}]}|W_{\mathcal{P}}, Q_n^{[\mathcal{P}]}),\ \forall n, \forall \mathcal{P} \quad (22)$$

Next, we need Lemma 3, which is an existence proof for index sets with specific properties. This technical lemma is needed in the proofs of upcoming two lemmas, Lemma 4 and Lemma 5. First, we give the definitions of relevant index sets $\mathcal{P}_a$, $\mathcal{P}_b$, $\mathcal{P}_c$, $\mathcal{P}_d$, and an element i_m. Given $\mathcal{P}_1$ and $\mathcal{P}_2$, we divide $\mathcal{P}_1$ into two disjoint partitions $\mathcal{P}_a$ and $\mathcal{P}_b$ (i.e., $\mathcal{P}_a \cup \mathcal{P}_b = \mathcal{P}_1$ and $\mathcal{P}_a \cap \mathcal{P}_b = \emptyset$), where $\mathcal{P}_a \subseteq \mathcal{P}_2$ (i.e., $\mathcal{P}_1 \cap \mathcal{P}_2 = \mathcal{P}_a$), $\mathcal{P}_b \subseteq \bar{\mathcal{P}}_2$. Suppose $|\mathcal{P}_a| = M \in [1 : P-1]$. Note that since $\mathcal{P}_1 \neq \mathcal{P}_2$, we cannot have $M = P$. We assume that $\mathcal{P}_a = \{i_1, \cdots, i_M\}$ for clarity of presentation. Given an arbitrary number $m \in [1 : M]$, we define a new index set $\mathcal{P}_c = \{i_1, \cdots, i_m\}$ which consists of exactly the first m elements in the index set $\mathcal{P}_a$. Let i_m be the last element from the index set $\mathcal{P}_c$. We obtain a new index set $\mathcal{P}_d = \{i_1, \cdots, i_{m-1}\}$ after removing this element. That means $\mathcal{P}_c = \mathcal{P}_d \cup \{i_m\}$.

Lemma 3 *For* $K \geq 3$, $1 \leq P \leq K-1$, *given index sets* $\mathcal{P}_1$, $\mathcal{P}_2$ *such that* $|\mathcal{P}_i| = P$ *for* $i = 1, 2$ *and* $\mathcal{P}_1 \neq \mathcal{P}_2$, *we can construct an index set* $\mathcal{P}_3$ *such that,*

i) $\mathcal{P}_3 \neq \mathcal{P}_1$ *and* $\mathcal{P}_3 \neq \mathcal{P}_2$,
ii) $|\mathcal{P}_3| = P$, *and*
iii) $\mathcal{P}_3$ *includes* $\mathcal{P}_b \cup \mathcal{P}_d$ *but does not include the common element* i_m *in* $\mathcal{P}_1 \cap \mathcal{P}_2$.

Next, we need the following lemma. Lemma 4 states that revealing any individual answer given the messages $(W_{\mathcal{P}_b}, W_{\mathcal{P}_d})$ does not leak any information about the message W_{i_m}.

Lemma 4 (Message leakage within any answer string) *When* $1 \leq P \leq K-1$ *and* $M \geq 1$, *for arbitrary* $m \in [1 : M]$,

$$\begin{aligned} H(W_{i_m}|W_{\mathcal{P}_b}, W_{\mathcal{P}_d}, &A_n^{[\mathcal{P}_2]}, Q_n^{[\mathcal{P}_2]}) \\ &= H(W_{i_m}|W_{\mathcal{P}_b}, W_{\mathcal{P}_d}, Q_n^{[\mathcal{P}_2]}) \quad (23) \end{aligned}$$

Finally, we prove that conditioning on an undesired message set does not decrease the uncertainty on any answer string.

Lemma 5 (Conditioning on an undesired message set)

$$\begin{aligned} H(A_n^{[\mathcal{P}_2]}|W_{\mathcal{P}_1}, Q_n^{[\mathcal{P}_2]}) &= H(A_n^{[\mathcal{P}_2]}|Q_n^{[\mathcal{P}_2]}), \\ &\forall n,\ \forall \mathcal{P}_1, \mathcal{P}_2 \ \textit{s.t.}\ \mathcal{P}_1 \neq \mathcal{P}_2, |\mathcal{P}_1| = |\mathcal{P}_2| \quad (24) \end{aligned}$$

Now, we are ready to construct the main body of the converse proof for MM-SPIR, as well as the minimal entropy of common randomness. Since we dealt with the inter-relations between message subsets in the previous lemmas and reached similar conclusions to those in SM-SPIR [7], the main body of the converse proof will be similar in structure to SM-SPIR.

The proof for $R \leq C_{MM-SPIR}$:

$$\begin{aligned}
PL &= H(W_{\mathcal{P}_1}) && (25)\\
&\overset{(11)}{=} H(W_{\mathcal{P}_1}|\mathcal{F}) && (26)\\
&\overset{(15)}{=} H(W_{\mathcal{P}_1}|\mathcal{F}) - H(W_{\mathcal{P}_1}|A_{1:N}^{[\mathcal{P}_1]}, \mathcal{F}) && (27)\\
&= I(W_{\mathcal{P}_1}; A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) && (28)\\
&= H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_{1:N}^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}) && (29)\\
&\overset{(12)}{=} H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_{1:N}^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}, Q_n^{[\mathcal{P}_1]}) && (30)\\
&\leq H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}, Q_n^{[\mathcal{P}_1]}) && (31)\\
&\overset{(22)}{=} H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, Q_n^{[\mathcal{P}_1]}) && (32)\\
&\overset{(20)}{=} H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_n^{[\mathcal{P}_2]}|W_{\mathcal{P}_1}, Q_n^{[\mathcal{P}_2]}) && (33)\\
&\overset{(24)}{=} H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_n^{[\mathcal{P}_2]}|Q_n^{[\mathcal{P}_2]}) && (34)\\
&\overset{(21)}{=} H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_n^{[\mathcal{P}_1]}|Q_n^{[\mathcal{P}_1]}) && (35)\\
&\leq H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_n^{[\mathcal{P}_1]}|Q_n^{[\mathcal{P}_1]}, \mathcal{F}) && (36)\\
&\overset{(12)}{=} H(A_{1:N}^{[\mathcal{P}_1]}|\mathcal{F}) - H(A_n^{[\mathcal{P}_1]}|\mathcal{F}) && (37)
\end{aligned}$$

By summing (37) up for all $n \in [1 : N]$ and letting $\mathcal{P}$ denote the general desired index set, we obtain,

$$\begin{aligned}
NPL &\leq NH(A_{1:N}^{[\mathcal{P}]}|\mathcal{F}) - \sum_{n=1}^{N} H(A_n^{[\mathcal{P}]}|\mathcal{F}) && (38)\\
&\leq NH(A_{1:N}^{[\mathcal{P}]}|\mathcal{F}) - H(A_{1:N}^{[\mathcal{P}]}|\mathcal{F}) && (39)\\
&= (N-1)H(A_{1:N}^{[\mathcal{P}]}|\mathcal{F}) && (40)\\
&\leq (N-1)\sum_{n=1}^{N} H(A_n^{[\mathcal{P}]}|\mathcal{F}) && (41)\\
&\leq (N-1)\sum_{n=1}^{N} H(A_n^{[\mathcal{P}]}) && (42)
\end{aligned}$$

which leads to the desired converse result on the retrieval rate,

$$R = \frac{PL}{\sum_{n=1}^{N} H(A_n^{[\mathcal{P}]})} \leq \frac{N-1}{N} = 1 - \frac{1}{N} \quad (43)$$

The proof for $H(S) \geq \frac{PL}{N-1}$:

$$\begin{aligned}
0 &\overset{(17)}{=} I(W_{\bar{\mathcal{P}}_1}; A_{1:N}^{[\mathcal{P}_1]}, Q_{1:N}^{[\mathcal{P}_1]}, \mathcal{F}) && (44)\\
&\geq I(W_{\bar{\mathcal{P}}_1}; A_{1:N}^{[\mathcal{P}_1]}, \mathcal{F}) && (45)\\
&= I(W_{\bar{\mathcal{P}}_1}; A_{1:N}^{[\mathcal{P}_1]}, W_{\mathcal{P}_1}, \mathcal{F}) && (46)\\
&\overset{(15)}{=} I(W_{\bar{\mathcal{P}}_1}; A_{1:N}^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}) && (47)\\
&\geq I(W_{\bar{\mathcal{P}}_1}; A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}) && (48)\\
&= H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}) - H(A_n^{[\mathcal{P}_1]}|W_{1:K}, \mathcal{F}) && (49)\\
&\overset{(12),(14)}{=} H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}) - H(A_n^{[\mathcal{P}_1]}|W_{1:K}, \mathcal{F})\\
&\quad + H(A_n^{[\mathcal{P}_1]}|W_{1:K}, \mathcal{F}, S) && (50)\\
&= H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}) - I(S; A_n^{[\mathcal{P}_1]}|W_{1:K}, \mathcal{F}) && (51)\\
&= H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1}, \mathcal{F}) - H(S|W_{1:K}, \mathcal{F})\\
&\quad + H(S|A_n^{[\mathcal{P}_1]}, W_{1:K}, \mathcal{F}) && (52)
\end{aligned}$$

$$\stackrel{(11)}{=} H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1},\mathcal{F})-H(S)+H(S|A_n^{[\mathcal{P}_1]},W_{1:K},\mathcal{F}) \tag{53}$$

$$\geq H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1},\mathcal{F}) - H(S) \tag{54}$$

$$\stackrel{(12)}{=} H(A_n^{[\mathcal{P}_1]}|W_{\mathcal{P}_1},\mathcal{F},Q_n^{[\mathcal{P}_1]}) - H(S) \tag{55}$$

$$= H(A_n^{[\mathcal{P}_1]}|Q_n^{[\mathcal{P}_1]}) - H(S) \tag{56}$$

where (56) follows from the steps between (32)-(35) by applying Lemma 1, 2 and 5 again.

By summing (56) up for all $n \in [1:N]$ and letting $\mathcal{P}$ denote the general desired index set again, we obtain,

$$0 \geq \sum_{n=1}^{N} H(A_n^{[\mathcal{P}]}|Q_n^{[\mathcal{P}]}) - NH(S) \tag{57}$$

$$\geq H(A_{1:N}^{[\mathcal{P}]}|Q_n^{[\mathcal{P}]}) - NH(S) \tag{58}$$

$$\geq H(A_{1:N}^{[\mathcal{P}]}|Q_n^{[\mathcal{P}]},\mathcal{F}) - NH(S) \tag{59}$$

$$= H(A_{1:N}^{[\mathcal{P}]}|\mathcal{F}) - NH(S) \tag{60}$$

$$\geq \frac{N}{N-1}PL - NH(S) \tag{61}$$

where (60) follows from (12) and (61) follows from (40), which leads to a lower bound for the minimal required entropy of common randomness S,

$$H(S) \geq \frac{PL}{N-1} \tag{62}$$

D. MM-SPIR: Achievability Proof

Since the MM-SPIR capacity is the same as the SM-SPIR capacity, and the required common randomness is P times the required common randomness for SM-SPIR, we can use the achievable scheme in [7] successively P times in a row (by utilizing independent common randomness each time) to achieve the MM-SPIR capacity. Although the query structure for the capacity-achieving scheme for SPIR in [7] is quite simple, it is fundamentally different than the query structure for the capacity-achieving scheme for PIR in [6]. This means that user/databases should execute different query structures for different database privacy levels. In this paper, by combining ideas for achievability from [20] and [13], we propose an alternative capacity-achieving scheme for MM-SPIR for any P. Our achievability scheme enables us to switch between MM-PIR and MM-SPIR seamlessly, and therefore support different database privacy levels, as the basic query structures are similar.

For convenience, we use the *k-sum* notation in [6], [20]. A k-sum is a sum of k symbols from k different messages. Thus, a k-sum symbol appears only in round k. We denote the number of stages in round k by α_k, which was originally introduced in [20]. In addition, we use ν to denote the number of repetitions of the scheme in [20] we need before we start assigning common randomness symbols.

Our achievability scheme is primarily based on the one in [20], with the addition of downloading and/or mixing common randomness variables into symbol downloads appropriately. We note that, here we extend the *near-optimal* algorithm in [20], which was originally proposed for $P \leq \frac{K}{2}$, to the case of $P \geq \frac{K}{2}$, and therefore, use it for all $1 \leq P \leq K-1$. Our achievability scheme comprises the following steps:

1) *Initial MM-PIR Query Generation:* Generate an initial query table strictly following the near-optimal procedure in [20] for arbitrary K, P and N.
2) *Repetition:* Repeat Step 1 for a total of ν times. The purpose of the repetition is to *i)* get an integer number of common randomness generated at each database by a symmetric algorithm, and *ii)* get equal number of symbols downloaded from each desired message. Let ν_0 be the smallest integer such that $\frac{(N-1)^{K-P}N\nu_0}{P}$ is an integer. Similarly, for $1 \leq k \leq \min\{P, K-P\}$, let ν_k be the smallest integer such that $\frac{\binom{P}{k}\alpha_k\nu_k}{N-1}$ is an integer. Choose ν as the lowest common multiple of all ν_k.
3) *Common Randomness Assignment:*
    a) In round 1, assign $\frac{\nu P\alpha_1}{N-1}$ independent common randomness symbols to each database, and download them. At each database, mix every 1-sum symbol containing a desired message symbol with an arbitrary common randomness symbol already downloaded from another database, making sure that every 1-sum symbol at each database is mixed with a different common randomness symbol. Mix all other 1-sum symbols not containing a desired symbol with a new common randomness symbol which is not downloaded by the user.
    b) In round k ($k \geq 2$), assign $\frac{\nu\binom{P}{k}\alpha_k}{N-1}$ independent common randomness symbols to each database, and download them. At each database: Mix every k-sum symbol containing only desired message symbols with an arbitrary common randomness symbol already downloaded from another database. Mix every k-sum symbol containing p desired message symbols ($1 \leq p \leq k-1$) with the common randomness symbol from the $(k-p)$-sum symbol having the same $k-p$ undesired message symbols downloaded at any other database. Mix every k-sum symbol not containing any desired message symbols with a new common randomness symbol which is not downloaded by the user.
    c) Repeat Step 3b until k reaches K. Note that if $\alpha_k = 0$, nothing is done.

VI. Putting Everything Together

Finally, we map our MM-SPIR result in Theorem 2 back to the PSI problem to obtain Theorem 1. Recall that, in the PSI problem, by generating the sets $\mathcal{P}_1$ and $\mathcal{P}_2$ by i.i.d. drawing the elements, we obtain i.i.d. messages in the MM-SPIR problem. Further, by choosing $q_i = \frac{1}{2}$, we obtain uniformly distributed messages, with message size $L = 1$. Therefore, the PSI problem is equivalent to an MM-SPIR problem with $L = 1$. We extend our MM-SPIR results for a finite message size L straightforwardly, see [52, Theorem 3]. Now, using this with $L = 1$, we obtain the result of this paper in Theorem 1.

References

[1] M. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. In *EUROCRYPT*. Springer, 2004.

[2] H. Chen, K. Laine, and P. Rindal. Fast private set intersection from homomorphic encryption. In *ACM SIGSAC CCS*, 2017.

[3] E. De Cristofaro and G. Tsudik. Practical private set intersection protocols with linear complexity. In *International Conference on Financial Cryptography and Data Security*. Springer, January 2010.

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

[5] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin. Protecting data privacy in private information retrieval schemes. In *ACM STOC*, May 1998.

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

[7] H. Sun and S. A. Jafar. The capacity of symmetric private information retrieval. *IEEE Trans. on Info. Theory*, 65(1):322–329, January 2019.

[8] H. Sun and S. A. Jafar. The capacity of robust private information retrieval with colluding databases. *IEEE Trans. on Info. Theory*, 64(4):2361–2370, April 2018.

[9] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb. Private information retrieval schemes for coded data with arbitrary collusion patterns. In *IEEE ISIT*, June 2017.

[10] R. Bitar and S. El Rouayheb. Staircase-PIR: Universally robust private information retrieval. In *IEEE ITW*, November 2018.

[11] Q. Wang and M. Skoglund. Symmetric private information retrieval for MDS coded distributed storage. In *IEEE ICC*, May 2017.

[12] Q. Wang and M. Skoglund. Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers. In *IEEE ITW*, November 2017.

[13] Q. Wang, H. Sun, and M. Skoglund. Symmetric private information retrieval with mismatched coded messages and randomness. In *IEEE ISIT*, July 2019.

[14] Q. Wang and M. Skoglund. Secure symmetric private information retrieval from colluding databases with adversaries. In *Allerton Conference*, October 2017.

[15] T. Guo, R. Zhou, and C. Tian. On the information leakage in private information retrieval systems. *IEEE Trans. on Info. Forensics and Security*, 15:2999–3012, 2020.

[16] K. Banawan and S. Ulukus. The capacity of private information retrieval from coded databases. *IEEE Trans. on Info. Theory*, 64(3):1945–1956, March 2018.

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

[18] S. Kumar, H.-Y. Lin, E. Rosnes, and A. G. i Amat. Achieving maximum distance separable private information retrieval capacity with linear codes. *IEEE Trans. on Info. Theory*, 65(7):4243–4273, July 2019.

[19] H. Sun and S. A. Jafar. Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al. *IEEE Trans. on Info. Theory*, 64(2):1000–1022, February 2018.

[20] K. Banawan and S. Ulukus. Multi-message private information retrieval: Capacity results and near-optimal schemes. *IEEE Trans. on Info. Theory*, 64(10):6842–6862, October 2018.

[21] K. Banawan and S. Ulukus. The capacity of private information retrieval from Byzantine and colluding databases. *IEEE Trans. on Info. Theory*, 65(2):1206–1219, February 2019.

[22] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti. Private information retrieval from coded storage systems with colluding, Byzantine, and unresponsive servers. *IEEE Trans. on Info. Theory*, 65(6):3898–3906, June 2019.

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

[24] Y.-P. Wei, K. Banawan, and S. Ulukus. Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching. *IEEE Trans. on Info. Theory*, 65(5):3215–3232, May 2019.

[25] Y.-P. Wei, K. Banawan, and S. Ulukus. Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits. *IEEE JSAC*, 36(6):1126–1139, June 2018.

[26] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson. Private information retrieval with side information: The single server case. In *Allerton Conference*, October 2017.

[27] Z. Chen, Z. Wang, and S. Jafar. The capacity of T-private information retrieval with private side information. *IEEE Trans. on Info. Theory*. doi: 10.1109/TIT.2020.2977919.

[28] Y.-P. Wei, K. Banawan, and S. Ulukus. The capacity of private information retrieval with partially known private side information. *IEEE Trans. on Info. Theory*, 65(12):8222–8231, December 2019.

[29] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali. Multi-message private information retrieval with private side information. In *IEEE ITW*, November 2018.

[30] A. Heidarzadeh, B. Garcia, S. Kadhe, S. E. Rouayheb, and A. Sprintson. On the capacity of single-server multi-message private information retrieval with side information. In *Allerton Conference*, October 2018.

[31] S. Li and M. Gastpar. Single-server multi-message private information retrieval with side information. In *Allerton Conference*, October 2018.

[32] Y.-P. Wei and S. Ulukus. The capacity of private information retrieval with private side information under storage constraints. *IEEE Trans. on Info. Theory*, 66(4):2023–2031, April 2020.

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

[34] M. Mirmohseni and M. A. Maddah-Ali. Private function retrieval. In *IWCIT*, April 2018.

[35] Z. Chen, Z. Wang, and S. Jafar. The asymptotic capacity of private search. *IEEE Trans. on Info. Theory*. doi: 10.1109/TIT.2020.2977082.

[36] M. A. Attia, D. Kumar, and R. Tandon. The capacity of private information retrieval from uncoded storage constrained databases. Available at arXiv:1805.04104.

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

[38] C. Tian. On the storage cost of private information retrieval. Available at arXiv:1910.11973.

[39] Y.-P. Wei, B. Arasli, K. Banawan, and S. Ulukus. The capacity of private information retrieval from decentralized uncoded caching databases. *Information*, 10, December 2019.

[40] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus. The capacity of private information retrieval from heterogeneous uncoded caching databases. *IEEE Trans. on Info. Theory*. doi: 10.1109/TIT.2020.2964762.

[41] K. Banawan and S. Ulukus. Private information retrieval from non-replicated databases. In *IEEE ISIT*, July 2019.

[42] K. Banawan and S. Ulukus. Private information retrieval through wiretap channel II: Privacy meets security. *IEEE Trans. on Info. Theory*. doi: 10.1109/TIT.2020.2977058.

[43] Q. Wang, H. Sun, and M. Skoglund. The capacity of private information retrieval with eavesdroppers. *IEEE Trans. on Info. Theory*, 65(5):3198–3214, May 2019.

[44] H. Yang, W. Shin, and J. Lee. Private information retrieval for secure distributed storage systems. *IEEE Trans. on Info. Forensics and Security*, 13(12):2953–2964, December 2018.

[45] Z. Jia, H. Sun, and S. Jafar. Cross subspace alignment and the asymptotic capacity of X-secure T-private information retrieval. *IEEE Trans. on Info. Theory*, 65(9):5783–5798, September 2019.

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

[47] R. Zhou, C. Tian, H. Sun, and T. Liu. Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size. *IEEE Trans. on Info. Theory*. doi: 10.1109/TIT.2020.2977073.

[48] K. Banawan and S. Ulukus. Asymmetry hurts: Private information retrieval under asymmetric-traffic constraints. *IEEE Trans. on Info. Theory*, 65(11):7628–7645, November 2019.

[49] K. Banawan and S. Ulukus. Noisy private information retrieval: On separability of channel coding and information retrieval. *IEEE Trans. on Info. Theory*, 65(12):8232–8249, December 2019.

[50] R. G. L. D'Oliveira and S. El Rouayheb. One-shot PIR: Refinement and lifting. *IEEE Trans. on Info. Theory*, 66(4):2443–2455, April 2020.

[51] R. Tajeddine, A. Wachter-Zeh, and C. Hollanti. Private information retrieval over random linear networks. *IEEE Trans. on Info. Forensics and Security*, 15:790–799, 2020.

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

[53] B. Chor, N. Gilboa, and M. Naor. Private information retrieval by keywords. *IACR Cryptology ePrint Archive*, 1997.