Symmetric Private Information Retrieval with User-Side Common Randomness

Zhusheng Wang   Sennur Ulukus
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
zhusheng@umd.edu   *ulukus@umd.edu*

***Abstract*—We consider the problem of symmetric private information retrieval (SPIR) with user-side common randomness. In SPIR, a user retrieves a message out of K messages from N non-colluding and replicated databases in such a way that no single database knows the retrieved message index (user privacy), and the user gets to know nothing further than the retrieved message (database privacy). SPIR has a capacity smaller than the PIR capacity which requires only user privacy, is infeasible in the case of a single database, and requires shared common randomness among the databases. We introduce a new variant of SPIR where the user is provided with a random subset of the shared database common randomness, which is unknown to the databases. We determine the exact capacity region of the triple (d, ρ_S, ρ_U), where d is the download cost, ρ_S is the amount of shared database (server) common randomness, and ρ_U is the amount of available user-side common randomness. We show that with a suitable amount of ρ_U, this new SPIR achieves the capacity of conventional PIR. As a corollary, single-database SPIR becomes feasible. Further, the presence of user-side ρ_U reduces the amount of required server-side ρ_S.**

I. Introduction

Private information retrieval (PIR) is a fundamental problem where a user downloads a message out of K possible messages stored in N non-colluding and replicated databases in such a way that no single database can know which message the user has downloaded [1]. This privacy requirement is referred to as *user privacy*. Symmetric PIR (SPIR) is an extended version of PIR where while downloading its desired message the user learns nothing about the remaining messages stored in the databases [2]. This is referred to as *database privacy*. While PIR can be achieved with no shared common randomness, it is well-known that information-theoretic SPIR is possible only when the databases share a certain minimum amount of common randomness that is unknown to the user.

The information-theoretic capacity of PIR and SPIR have been found in [3], [4] as $C_{\text{PIR}} = \frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^K}$ and $C_{\text{SPIR}} = 1 - \frac{1}{N}$. First, C_{SPIR} is smaller than C_{PIR}, since SPIR is a more constrained problem than PIR, as it requires not only user privacy but also database privacy. Second, single-database SPIR is infeasible as $C_{\text{SPIR}} = 0$ for $N = 1$, while single database PIR is feasible as $C_{\text{PIR}} = \frac{1}{K}$ for $N = 1$. Our goal in this paper is two-fold: To explore ways to increase SPIR capacity to the level of PIR capacity, and perhaps more importantly, to make single-database SPIR feasible.

Our motivation to focus on SPIR comes from constantly growing importance of privacy, not only the privacy of the retrieving user, but also the privacy of the databases, as the stored information in the databases may belong to other users. In addition, recent papers showed that other important privacy primitives, such as private set intersection (PSI) [5], [6], can be recast as versions of the SPIR problem, e.g., multi-message SPIR. Thus, here, we investigate ways to increase the SPIR capacity. Further, in practical applications, enforcing non-collusion could be difficult, as in some cases, all databases may naturally belong to the same entity, e.g., in various multi-party secure computation problems [5]–[12]. If all databases collude or belong to the same entity, the system essentially becomes a single-database system [13]. The single-database PIR problem has been studied under extended conditions, e.g., side-information [14]–[17]. Here, we investigate single-database SPIR under extended conditions, with the goal of making it feasible. Other important variations of PIR and SPIR problem have also been investigated; see e.g., [18]–[51].

In this paper, we introduce SPIR with user-side common randomness, which solves the above two issues. In this model, the user obtains a part of the common randomness shared by the databases. The databases know the size of the user-side common randomness, but they do not know what the user possesses exactly. One way to implement this is for the user to fetch a part of the common randomness from the databases uniformly randomly, i.e., without the user knowing what it will get and without the databases knowing what it got, except for its cardinality. That is, all subsets of a certain size are equally likely to be obtained by the user. Another practical implementation could be for an external helper to distribute common randomness to the user and the databases randomly.

For database-side (server-side) common randomness of amount ρ_S and user-side common randomness of amount ρ_U, we determine the exact capacity region of the triple (d, ρ_S, ρ_U), where d is the download cost which is the inverse of the capacity. We show that with a suitable ρ_U, SPIR capacity becomes equal to the conventional PIR capacity. For the single-database case, since the conventional PIR capacity is $\frac{1}{K}$, this implies that single-database SPIR with user-side common randomness is feasible. In addition, the presence of user-side ρ_U reduces the amount of required server-side ρ_S.

II. Problem Formulation

We consider a system of $N \geq 1$ non-colluding databases each storing the same set of $K \geq 2$ i.i.d. messages each of which consisting of L i.i.d. symbols uniformly selected from a sufficiently large finite field $\mathbb{F}_q$, i.e.,

$$H(W_k) = L, \quad k \in [K] \tag{1}$$

$$H(W_{1:K}) = H(W_1) + \cdots + H(W_K) = KL \tag{2}$$

As in [4], we use a random variable $\mathcal{F}$ to denote the randomness in the retrieval strategy implemented by the user. Due to the user privacy constraint, the realization of $\mathcal{F}$ is only known to the user, and is unknown to any of the databases. Due to the database privacy constraint, databases need to share some amount of common randomness $\mathcal{R}_S$; we will call this *server-side* common randomness. Before the retrieval process starts, the user obtains a partial knowledge of $\mathcal{R}_S$. We denote it by $\mathcal{R}_U$, and call it *user-side* common randomness. $\mathcal{R}_U$ is a subset of $\mathcal{R}_S$. The user-side common randomness $\mathcal{R}_U$ is unknown to the databases, i.e., it is only known to be equally distributed among all subsets of $\mathcal{R}_S$ with cardinality $|\mathcal{R}_U|$.

The message set $W_{1:K}$ stored in the databases is independent of the desired message index k, retrieval strategy randomness at the user $\mathcal{F}$ and all the common randomness,

$$I(W_{1:K}; k, \mathcal{F}, \mathcal{R}_S, \mathcal{R}_U) = 0, \quad \forall k,\ \forall \mathcal{R}_U \tag{3}$$

During the query generation stage, the user has no access to the message set $W_{1:K}$ in the databases and the common randomness difference $\mathcal{R}_S \backslash \mathcal{R}_U$,

$$I(Q_{1:N}^{[k,\mathcal{R}_U]}; W_{1:K}, \mathcal{R}_S \backslash \mathcal{R}_U) = 0, \quad \forall k,\ \forall \mathcal{R}_U \tag{4}$$

Using the desired message index and the user-side common randomness, the user generates a query for each database according to the retrieval strategy randomness $\mathcal{F}$. Hence, the queries $Q_n^{[k,\mathcal{R}_U]}, n \in [N]$ are deterministic functions of $\mathcal{F}$,

$$H(Q_1^{[k,\mathcal{R}_U]}, Q_2^{[k,\mathcal{R}_U]}, \ldots, Q_N^{[k,\mathcal{R}_U]} | \mathcal{F}) = 0 \quad \forall k,\ \forall \mathcal{R}_U \tag{5}$$

After receiving a query from the user, each database generates a truthful answer based on the stored message set and the server-side common randomness,

$$H(A_n^{[k,\mathcal{R}_U]} | Q_n^{[k,\mathcal{R}_U]}, W_{1:K}, \mathcal{R}_S) = 0, \quad \forall n,\ \forall k,\ \forall \mathcal{R}_U \tag{6}$$

After collecting all N answers from the databases, the user should be able to decode the desired messages W_k reliably,

$$\text{[reliability]} \quad H(W_k | \mathcal{F}, A_{1:N}^{[k,\mathcal{R}_U]}, \mathcal{R}_U) = 0, \quad \forall k,\ \forall \mathcal{R}_U \tag{7}$$

Due to the user privacy constraint, the query generated to retrieve the desired message should be statistically indistinguishable from other queries. Specifically, for all k, k', all n, and all user-side common randomness $\mathcal{R}_U$, there exists some $\mathcal{R}'_U$ with $H(\mathcal{R}'_U) = H(\mathcal{R}_U)$ such that,

$$\begin{aligned}\text{[user privacy]} \quad & (Q_n^{[k,\mathcal{R}_U]}, A_n^{[k,\mathcal{R}_U]}, W_{1:K}, \mathcal{R}_S) \\ & \sim\ (Q_n^{[k',\mathcal{R}'_U]}, A_n^{[k',\mathcal{R}'_U]}, W_{1:K}, \mathcal{R}_S)\end{aligned} \tag{8}$$

Furthermore as in [45], after factorizing the joint distribution of all the random variables in the server, we obtain the following equivalent expression for user privacy for all potential query realizations q,

$$\text{[user privacy]} \quad P(Q_n^{[k,\mathcal{R}_U]} = q) = P(Q_n^{[k',\mathcal{R}'_U]} = q) \tag{9}$$

Due to the database privacy constraint, the user should learn nothing about $W_{\bar{k}}$ which is the complement of W_k, i.e., $W_{\bar{k}} = \{W_1, \cdots, W_{k-1}, W_{k+1}, \cdots, W_K\}$,

$$\text{[database privacy]} \quad I(W_{\bar{k}}; \mathcal{F}, A_{1:N}^{[k,\mathcal{R}_U]}, \mathcal{R}_U) = 0 \tag{10}$$

Again due to the database privacy, the user should not gain any knowledge about the remaining common randomness in the server even after retrieving the desired message,

$$I(\mathcal{R}_S \backslash \mathcal{R}_U; \mathcal{F}, A_{1:N}^{[k,\mathcal{R}_U]}, W_k, \mathcal{R}_U) = 0 \tag{11}$$

An achievable SPIR scheme is a scheme that satisfies the reliability constraint (7), the user privacy constraint (8) and the database privacy constraint (10). As usual, the efficiency of the scheme is measured in terms of the maximal number of downloaded bits by the user from all the databases, denoted by D. We define the normalized download cost d, the normalized server-side common randomness ρ_S, and the normalized user-side common randomness as ρ_U as,

$$d = \frac{D}{L}, \quad \rho_S = \frac{H(\mathcal{R}_S)}{L}, \quad \rho_U = \frac{H(\mathcal{R}_U)}{L} \tag{12}$$

where L is the message length. Our goal in this paper is to determine the largest region for simultaneously achievable triples (d, ρ_S, ρ_U) over all valid retrieval schemes.

III. Main Results

We state the main result of our paper in the following theorem which is the *capacity region* for the triple (d, ρ_S, ρ_U).

Theorem 1 *With user-side common randomness, the multi-database SPIR capacity region for $N \geq 2$ and $K \geq 2$ is*

$$d \geq 1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \tag{13}$$

$$\rho_S - \rho_U \geq \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \tag{14}$$

$$\frac{N-1}{N} d + \rho_U \geq 1 \tag{15}$$

$$\frac{N}{N-1} \rho_U + N\rho_S \geq \frac{N}{N-1} \tag{16}$$

Remark 1 *The right hand side of (13) is the optimum normalized download cost of classical PIR, d_{PIR} [3]. Thus, (13) states that $d \geq d_{PIR}$. When $\rho_U = 0$, i.e., when there is no user-side common randomness, (15) becomes $d \geq d_{SPIR}$, where $d_{SPIR} = \frac{N}{N-1}$ is the optimum normalized download cost of classical SPIR [4]. Note that $d_{SPIR} > d_{PIR}$ for all N. Therefore, when $\rho_U = 0$, (15) is binding, (13) is loose, and we have $d = d_{SPIR}$. Theorem 1 implies that with appropriate ρ_U, e.g., with $\rho_U = \frac{1}{N^K}$, both (13) and (15) can be made binding, at which time the new SPIR download cost achieves $d = d_{PIR}$.*

Remark 2 *When $\rho_U = 0$, Theorem 1 reduces to the capacity of classical SPIR [4], as in this case, (15) gives $d \geq \frac{N}{N-1}$, (16) gives $\rho_S \geq \frac{1}{N-1}$, and (13) and (14) are non-binding.*

Remark 3 *The gap between ρ_S and ρ_U must be no smaller than a specific value as a function of N and K as given on the right hand side of (14). This comes from the server privacy constraint, where part of the common randomness, i.e., $\mathcal{R}_S \backslash \mathcal{R}_U$, is utilized to hide the undesired messages.*

Remark 4 *From (16), we observe that the existence of user-side common randomness can help reduce the required amount of server-side common randomness. For instance, for $N = 2$ databases and $K = 2$ messages, classical SPIR optimum download cost $d = d_{SPIR} = 2$ is achieved by $\rho_S = 1$ [4]. In Theorem 1, $d = 2$ can be achieved by $\rho_S = \frac{3}{4}$ with $\rho_U = \frac{1}{4}$.*

Corollary 1 *With user-side common randomness, the single-database SPIR capacity region for $N = 1$ and $K \geq 2$ is*

$$d \geq K \tag{17}$$

$$\rho_S - \rho_U \geq K - 1 \tag{18}$$

$$\rho_U \geq 1 \tag{19}$$

Remark 5 *It is well-known that, for $N = 1$, classical SPIR is not feasible [4]. With user-side common randomness, single-database SPIR becomes feasible.*

Remark 6 *The optimal normalized download cost for single-database PIR is $d = K$ [3], [14], which is achieved by downloading all messages from the server. One of the difficulties of single-database SPIR is that downloading all messages is not a valid SPIR scheme. Corollary 1 shows that single-database PIR capacity can be achieved for single-database SPIR by means of user-side common randomness.*

Remark 7 *The first two terms in Corollary 1 follow from the first two terms in Theorem 1. The third term in Corollary 1 follows from the last two terms in Theorem 1 by multiplying both sides of the fourth term in Theorem 1 by $N - 1$.*

Remark 8 *Like multi-database SPIR, in the single-database SPIR as well, the gap between ρ_S and ρ_U must be no smaller than a specific value as a function of K as given in (18) to avoid information leakage on undesired messages.*

IV. Motivating Example

Example 1 *We consider a single-database case $N = 1$, $K = 3$ and $L = 1$. We use W_1, W_2 and W_3 to denote the three messages. Our new achievable scheme is given in Table I.*

The reliability constraint follows from the fact that the user can always decode the desired message by using its own common randomness. The server privacy constraint follows from the fact that the undesired messages are always mixed with unknown common randomness. For the user-privacy constraint, we have for all $k, k' \in [3], k \neq k'$ and a randomly selected $\mathcal{R}_U \in \{S_1, S_2, S_3\}$ under a uniform distribution, there exists another different $\mathcal{R}'_U \in \{S_1, S_2, S_3\}$ such that,

$$P(Q^{[k,\mathcal{R}_U]} = q) = P(Q^{[k',\mathcal{R}'_U]} = q) = \frac{1}{3} \tag{20}$$

where $q \in \{[W_1 + S_1, W_2 + S_2, W_3 + S_3], [W_1 + S_2, W_2 + S_3, W_3 + S_1], [W_1 + S_3, W_2 + S_1, W_3 + S_2]\}$. Specifically from the point of view of the server, the same set of queries can be invoked for any desired message $W_i, i \in [3]$ with the same probability distribution. This scheme achieves $d = 3$, $\rho_U = 1$ and $\rho_S = 3$, which exactly matches the boundary of the SPIR capacity region for $N = 1$ and $K = 3$ in Corollary 1.

$\mathcal{R}_U$	desired message		
	W_1	W_2	W_3
S_1	$W_1 + S_1$ $W_2 + S_2$ $W_3 + S_3$	$W_2 + S_1$ $W_3 + S_2$ $W_1 + S_3$	$W_3 + S_1$ $W_1 + S_2$ $W_2 + S_3$
S_2	$W_1 + S_2$ $W_2 + S_3$ $W_3 + S_1$	$W_2 + S_2$ $W_3 + S_3$ $W_1 + S_1$	$W_3 + S_2$ $W_1 + S_3$ $W_2 + S_1$
S_3	$W_1 + S_3$ $W_2 + S_1$ $W_3 + S_2$	$W_2 + S_3$ $W_3 + S_1$ $W_1 + S_2$	$W_3 + S_3$ $W_1 + S_1$ $W_2 + S_2$

TABLE I: The query table for the case $N = 1$, $K = 3$.

Example 2 *We consider a multi-database case $N = 2$, $K = 2$ and $L = 4$. We use $[a_1, a_2, a_3, a_4]$ as a random uniform permutation of the symbols in the first message W_1, and independently, $[b_1, b_2, b_3, b_4]$ as another one for W_2. Due to message index permutations, each set of queries represents one of $4! \cdot 4 \cdot 3 = 288$ different possible permutations. We have two different sets for each $\mathcal{R}_U$ because of necessary common randomness permutations. Our new achievable scheme for one random message index selection is given in Table II.*

$\mathcal{R}_U$	Desired message: W_1		Desired message: W_2	
	DB1	DB2	DB1	DB2
S_1	$a_1 + S_1$ $b_1 + S_2$ $a_3 + b_2 + S_3$	$a_2 + S_1$ $b_2 + S_3$ $a_4 + b_1 + S_2$	$b_1 + S_1$ $a_1 + S_2$ $b_3 + a_2 + S_3$	$b_2 + S_1$ $a_2 + S_3$ $b_4 + a_1 + S_2$
	$a_1 + S_1$ $b_1 + S_3$ $a_3 + b_2 + S_2$	$a_2 + S_1$ $b_2 + S_2$ $a_4 + b_1 + S_3$	$b_1 + S_1$ $a_1 + S_3$ $b_3 + a_2 + S_2$	$b_2 + S_1$ $a_2 + S_2$ $b_4 + a_1 + S_3$
S_2	$a_1 + S_2$ $b_1 + S_3$ $a_3 + b_2 + S_1$	$a_2 + S_2$ $b_2 + S_1$ $a_4 + b_1 + S_3$	$b_1 + S_2$ $a_1 + S_3$ $b_3 + a_2 + S_1$	$b_2 + S_2$ $a_2 + S_1$ $b_4 + a_1 + S_3$
	$a_1 + S_2$ $b_1 + S_1$ $a_3 + b_2 + S_3$	$a_2 + S_2$ $b_2 + S_3$ $a_4 + b_1 + S_1$	$b_1 + S_2$ $a_1 + S_1$ $b_3 + a_2 + S_3$	$b_2 + S_2$ $a_2 + S_3$ $b_4 + a_1 + S_1$
S_3	$a_1 + S_3$ $b_1 + S_1$ $a_3 + b_2 + S_2$	$a_2 + S_3$ $b_2 + S_2$ $a_4 + b_1 + S_1$	$b_1 + S_3$ $a_1 + S_1$ $b_3 + a_2 + S_2$	$b_2 + S_3$ $a_2 + S_2$ $b_4 + a_1 + S_1$
	$a_1 + S_3$ $b_1 + S_2$ $a_3 + b_2 + S_1$	$a_2 + S_3$ $b_2 + S_1$ $a_4 + b_1 + S_2$	$b_1 + S_3$ $a_1 + S_2$ $b_3 + a_2 + S_1$	$b_2 + S_3$ $a_2 + S_1$ $b_4 + a_1 + S_2$

TABLE II: The query table for the case $N = 2$, $K = 2$.

Verification that this proposed scheme achieves the user privacy and the database privacy constraints is similar to the one in Example 1. This scheme achieves $d = 1.5$, $\rho_U = 0.25$ and $\rho_S = 0.75$. This is a corner point of the capacity region in Theorem 1 where all inequalities are satisfied with equality. The other corner point when $\rho_U = 0$ is achieved by the classical SPIR scheme in [4]. Any point on the line segment joining these two points can be achieved by time-sharing between these two schemes. Any other remaining point in Theorem 1 can be achieved by adding extra randomness in the user- and server-side simultaneously, or by increasing the server-side common randomness and the download cost.

V. Converse Proof

We provide a sketch of the converse proof of Theorem 1 here. The four inequalities in Theorem 1 are proved in Lemmas 3, 4, 9 and 10 below. Towards proving these four lemmas, we need Lemmas 1-2 and Lemmas 5-8 below. We note that Lemmas 1-2 extend [3, Lemmas 5-6], and Lemmas 5-8 extend [4, Eqns. (26), (27), (30), (39)]. These extensions are needed because we have two additional sets of random variables in our system model: $\mathcal{R}_S$ and $\mathcal{R}_U$ with respect to techniques in [3], and $\mathcal{R}_U$ with respect to techniques in [4].

Lemma 1

$$I(W_{2:K}; Q_{1:N}^{[1,\mathcal{R}_U]}, A_{1:N}^{[1,\mathcal{R}_U]}, \mathcal{R}_S | W_1) \leq D - L \tag{21}$$

Lemma 2

$$\begin{aligned} &I(W_{k:K}; Q_{1:N}^{[k-1,\mathcal{R}_U]}, A_{1:N}^{[k-1,\mathcal{R}_U]}, \mathcal{R}_S | W_{1:k-1}) \\ &\quad \geq \frac{1}{N} I(W_{k+1:K}; Q_{1:N}^{[k,\mathcal{R}'_U]}, A_{1:N}^{[k,\mathcal{R}'_U]}, \mathcal{R}_S | W_{1:k}) + \frac{L}{N} \end{aligned} \tag{22}$$

Lemma 3 (Minimal download cost d)

$$d \geq 1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \tag{23}$$

Proof: Following steps similar to [3, Eqns. (62)-(67)] for Lemma 2, we obtain

$$\begin{aligned} &I(W_{2:K}; Q_{1:N}^{[1,\mathcal{R}_U]}, A_{1:N}^{[1,\mathcal{R}_U]}, \mathcal{R}_S | W_1) \\ &\quad \geq \left(\frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \right) L \end{aligned} \tag{24}$$

Combining the upper bound in Lemma 1 and the lower bound in (24) completes the proof. ■

Lemma 4 (Minimal difference between ρ_S and ρ_U)

$$\rho_S - \rho_U \geq \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \tag{25}$$

Proof: From (24), we have the following relation,

$$\begin{aligned} &H(W_{2:K} | Q_{1:N}^{[1,\mathcal{R}_U]}, A_{1:N}^{[1,\mathcal{R}_U]}, W_1, \mathcal{R}_S) \\ &\quad \leq (K-1)L - \left(\frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \right) L \end{aligned} \tag{26}$$

Next, we have the following upper bound,

$$\begin{aligned} &I(W_{2:K}; \mathcal{R}_S \backslash \mathcal{R}_U | Q_{1:N}^{[1,\mathcal{R}_U]}, A_{1:N}^{[1,\mathcal{R}_U]}, W_1, \mathcal{R}_U) \\ &\quad \leq H(\mathcal{R}_S \backslash \mathcal{R}_U | Q_{1:N}^{[1,\mathcal{R}_U]}, A_{1:N}^{[1,\mathcal{R}_U]}, W_1, \mathcal{R}_U) \end{aligned} \tag{27}$$

$$\overset{(11)}{=} H(\mathcal{R}_S \backslash \mathcal{R}_U) = H(\mathcal{R}_S) - H(\mathcal{R}_U) \tag{28}$$

and the following lower bound,

$$\begin{aligned} &I(W_{2:K}; \mathcal{R}_S \backslash \mathcal{R}_U | Q_{1:N}^{[1,\mathcal{R}_U]}, A_{1:N}^{[1,\mathcal{R}_U]}, W_1, \mathcal{R}_U) \\ &\overset{(10),(7),(5)}{=} (K-1)L - H(W_{2:K} | Q_{1:N}^{[1,\mathcal{R}_U]}, A_{1:N}^{[1,\mathcal{R}_U]}, W_1, \mathcal{R}_S) \end{aligned} \tag{29}$$

$$\overset{(26)}{\geq} \left(\frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \right) L \tag{30}$$

Combining (28) and (30) yields the desired result. ■

Lemma 5 *For any $k' \neq k$,*

$$\begin{aligned} &H(A_n^{[k,\mathcal{R}_U]} | Q_n^{[k,\mathcal{R}_U]}, W_k, \mathcal{R}_U) \\ &\quad \geq H(A_n^{[k',\mathcal{R}'_U]} | Q_n^{[k',\mathcal{R}'_U]}, W_k, \mathcal{R}'_U) - H(\mathcal{R}_U) \end{aligned} \tag{31}$$

Lemma 6

$$H(A_n^{[k,\mathcal{R}_U]} | Q_n^{[k,\mathcal{R}_U]}, \mathcal{R}_U) = H(A_n^{[k',\mathcal{R}'_U]} | Q_n^{[k',\mathcal{R}'_U]}, \mathcal{R}'_U) \tag{32}$$

Lemma 7

$$\begin{aligned} &H(A_n^{[k,\mathcal{R}_U]} | \mathcal{F}, Q_n^{[k,\mathcal{R}_U]}, W_k, \mathcal{R}_U) \\ &\quad = H(A_n^{[k,\mathcal{R}_U]} | Q_n^{[k,\mathcal{R}_U]}, W_k, \mathcal{R}_U) \end{aligned} \tag{33}$$

Lemma 8 *For any $k' \neq k$,*

$$H(A_n^{[k',\mathcal{R}'_U]} | Q_n^{[k',\mathcal{R}'_U]}, \mathcal{R}'_U) = H(A_n^{[k',\mathcal{R}'_U]} | Q_n^{[k',\mathcal{R}'_U]}, W_k, \mathcal{R}'_U) \tag{34}$$

Lemma 9 (Minimal bound for d and ρ_U)

$$\frac{N-1}{N} d + \rho_U \geq 1 \tag{35}$$

Proof: Starting from the message length assumption (1),

$$\begin{aligned} L = H(W_k) &\overset{(3)}{=} H(W_k | \mathcal{F}, \mathcal{R}_U) && (36) \\ &\overset{(7)}{=} H(W_k | \mathcal{F}, \mathcal{R}_U) - H(W_k | \mathcal{F}, A_{1:N}^{[k,\mathcal{R}_U]}, \mathcal{R}_U) && (37) \\ &= I(W_k; A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, \mathcal{R}_U) && (38) \\ &= H(A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, \mathcal{R}_U) - H(A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, W_k, \mathcal{R}_U) && (39) \\ &\leq H(A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, \mathcal{R}_U) - H(A_n^{[k,\mathcal{R}_U]} | \mathcal{F}, Q_n^{[k,\mathcal{R}_U]}, W_k, \mathcal{R}_U) && (40) \\ &\overset{(33)}{=} H(A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, \mathcal{R}_U) - H(A_n^{[k,\mathcal{R}_U]} | Q_n^{[k,\mathcal{R}_U]}, W_k, \mathcal{R}_U) && (41) \\ &\overset{(31)}{\leq} H(A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, \mathcal{R}_U) - H(A_n^{[k',\mathcal{R}'_U]} | Q_n^{[k',\mathcal{R}'_U]}, W_k, \mathcal{R}'_U) \\ &\quad + H(\mathcal{R}_U) && (42) \\ &\overset{(34)}{=} H(A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, \mathcal{R}_U) - H(A_n^{[k',\mathcal{R}'_U]} | Q_n^{[k',\mathcal{R}'_U]}, \mathcal{R}'_U) \\ &\quad + H(\mathcal{R}_U) && (43) \\ &\overset{(32)}{=} H(A_{1:N}^{[k,\mathcal{R}_U]} | \mathcal{F}, \mathcal{R}_U) - H(A_n^{[k,\mathcal{R}_U]} | Q_n^{[k,\mathcal{R}_U]}, \mathcal{R}_U) \\ &\quad + H(\mathcal{R}_U) && (44) \end{aligned}$$

$$\overset{(5)}{\leq} H(A_{1:N}^{[k,\mathcal{R}_U]}|\mathcal{F},\mathcal{R}_U) - H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},\mathcal{R}_U) + H(\mathcal{R}_U) \tag{45}$$

By summing (45) over all $n \in [1:N]$, we obtain the following relationship, which completes the proof,

$$\begin{aligned}
NL \leq\ & NH(A_{1:N}^{[k,\mathcal{R}_U]}|\mathcal{F},\mathcal{R}_U) - \sum_{n=1}^{N} H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},\mathcal{R}_U) \\
& + NH(\mathcal{R}_U) && (46) \\
\leq\ & (N-1)H(A_{1:N}^{[k,\mathcal{R}_U]}|\mathcal{F},\mathcal{R}_U) + NH(\mathcal{R}_U) && (47) \\
\leq\ & (N-1)\sum_{n=1}^{N} H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},\mathcal{R}_U) + NH(\mathcal{R}_U) && (48) \\
\leq\ & (N-1)D + NH(\mathcal{R}_U) && (49)
\end{aligned}$$

■

Lemma 10 (Minimal bound for ρ_U and ρ_S)

$$\frac{N}{N-1}\rho_U + N\rho_S \geq \frac{N}{N-1} \tag{50}$$

Proof: Starting with the database privacy constraint (10),

$$\begin{aligned}
0 &= I(W_{\bar{k}};\mathcal{F},A_{1:N}^{[k,\mathcal{R}_U]},\mathcal{R}_U) && (51) \\
&\overset{(3)}{=} I(W_{\bar{k}};A_{1:N}^{[k,\mathcal{R}_U]},\mathcal{R}_U|\mathcal{F}) && (52) \\
&\overset{(7)}{=} I(W_{\bar{k}};A_{1:N}^{[k,\mathcal{R}_U]},\mathcal{R}_U|\mathcal{F}) + I(W_{\bar{k}};W_k|\mathcal{F},A_{1:N}^{[k,\mathcal{R}_U]},\mathcal{R}_U) && (53) \\
&= I(W_{\bar{k}};A_{1:N}^{[k,\mathcal{R}_U]},W_k,\mathcal{R}_U|\mathcal{F}) && (54) \\
&= I(W_{\bar{k}};A_{1:N}^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) + I(W_{\bar{k}};W_k,\mathcal{R}_U|\mathcal{F}) && (55) \\
&\overset{(3)}{=} I(W_{\bar{k}};A_{1:N}^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) && (56) \\
&\geq I(W_{\bar{k}};A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) && (57) \\
&\overset{(6),(5)}{=} H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_{1:K},\mathcal{R}_U) \\
&\quad + H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_{1:K},\mathcal{R}_S) && (58) \\
&\geq H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_{1:K},\mathcal{R}_U) \\
&\quad + H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_{1:K},\mathcal{R}_S,\mathcal{R}_U) && (59) \\
&= H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - I(A_n^{[k,\mathcal{R}_U]};\mathcal{R}_S|\mathcal{F},W_{1:K},\mathcal{R}_U) && (60) \\
&= H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - H(\mathcal{R}_S|\mathcal{F},W_{1:K},\mathcal{R}_U) \\
&\quad + H(\mathcal{R}_S|\mathcal{F},A_n^{[k,\mathcal{R}_U]},W_{1:K},\mathcal{R}_U) && (61) \\
&\geq H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - H(\mathcal{R}_S|\mathcal{F},W_k,\mathcal{R}_U) && (62) \\
&= H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - H(\mathcal{R}_U,\mathcal{R}_S\backslash\mathcal{R}_U|\mathcal{F},W_k,\mathcal{R}_U) && (63) \\
&= H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - H(\mathcal{R}_S\backslash\mathcal{R}_U|\mathcal{F},W_k,\mathcal{R}_U) && (64) \\
&\overset{(11)}{=} H(A_n^{[k,\mathcal{R}_U]}|\mathcal{F},W_k,\mathcal{R}_U) - H(\mathcal{R}_S\backslash\mathcal{R}_U) && (65) \\
&= H(A_n^{[k,\mathcal{R}_U]}|Q_n^{[k,\mathcal{R}_U]},\mathcal{R}_U) - H(\mathcal{R}_S) && (66)
\end{aligned}$$

where (66) follows from the steps between (41)-(44) by applying Lemma 5 to Lemma 8 again.

By summing (66) over all $n \in [1:N]$, we obtain the following relationship, which completes the proof,

$$\begin{aligned}
0 &\geq \sum_{n=1}^{N} H(A_n^{[k,\mathcal{R}_U]}|Q_n^{[k,\mathcal{R}_U]},\mathcal{R}_U) - NH(\mathcal{R}_S) && (67) \\
&\geq H(A_{1:N}^{[1,\mathcal{R}_U]}|\mathcal{F},Q_n^{[1,\mathcal{R}_U]},\mathcal{R}_U) - NH(\mathcal{R}_S) && (68) \\
&\overset{(5),(47)}{\geq} \frac{N}{N-1}L - \frac{N}{N-1}H(\mathcal{R}_U) - NH(\mathcal{R}_S) && (69)
\end{aligned}$$

■

VI. Achievability Proof

Our achievability is based on the principle of converting a given PIR scheme into a valid SPIR scheme using the server-side and user-side common randomness in a manner that does not compromise the download cost; e.g., [52]. To that goal, the common randomness added to the desired symbols are substracted out as they are available at the user side, and the remaining common randomness unknown to the user protects the undesired messages. The challenge is to implement this for all possible user-side common randomness realizations which are unknown ahead of time. Steps of our achievable scheme:

1) *Initial PIR query generation:* For given N and K, generate an initial PIR query table for each desired message using the scheme in [3], e.g., Tables I-II without S_is.
2) *Server-side common randomness assignment:* For each desired message and each permutation of message index (e.g., 288 permutations in Example 2), mix all 1-sum symbols from the desired message across all the databases with the same new common randomness. We call it seed common randomness (e.g., S_1 in first three rows of Table II). Assign a new distinct common randomness to every 1-sum symbol from the undesired messages. For every k-sum symbol containing a desired message symbol, mix it with the common randomness from the $(k-1)$-sum symbol having the same $k-1$ undesired message symbols queried at another database. For every k-sum symbol not containing any desired message symbol, assign a new distinct common randomness. Repeat this until k reaches K. Next, we permute non-seed common randomness indices (e.g., 4th-6th rows of Table II). We call this whole modified query table a *query cell.*
3) *Server-side common randomness cycling:* While keeping each query cell, create a new one by adding 1 (mod $|\mathcal{R}_S|$) to each common randomness index (e.g., S_1 becomes S_2 in Table II). Repeat it $|\mathcal{R}_S|$ times such that each query cell has a different seed common randomness index.
4) *Query cell determination:* The user has $|\mathcal{R}_U|$ server-side common randomness. The user determines the query cell to be invoked, and selects a random permutation within that cell, by matching its user-side common randomness to the seed common randomness of the cell.

In this scheme, the message length L is N^K as in [3], the total amount of server-side common randomness required $|\mathcal{R}_S|$ is $1 + \cdots + N^{K-1}$ and the total amount of user-side common randomness required $|\mathcal{R}_U|$ is 1.

REFERENCES

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

[2] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin. Protecting data privacy in private information retrieval schemes. In *Thirtieth Annual ACM Symposium on Theory of Computing*, pages 151–160, May 1998.

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

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

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

[6] Z. Wang, K. Banawan, and S. Ulukus. Multi-party private set intersection: An information-theoretic approach. *IEEE Journal on Selected Areas in Information Theory*, 2(1):366–379, 2021.

[7] U. Feige, J. Killian, and M. Naor. A minimal model for secure computation. In *Proceedings of the twenty-sixth annual ACM Symposium on Theory of Computing*, pages 554–563, 1994.

[8] O. Goldreich. Secure multi-party computation. Manuscript. Preliminary version, 78, 1998.

[9] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: A review and open problems. In *Proceedings of the 2001 Workshop on New Security Paradigms*, page 13–22. ACM, 2001.

[10] M. J. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. In *Advances in Cryptology - EUROCRYPT 2004*, pages 1–19. Springer Berlin Heidelberg, 2004.

[11] L. Kissner and D. Song. Privacy-preserving set operations. In *Advances in Cryptology – CRYPTO 2005*, pages 241–257.

[12] Y. Zhao and H. Sun. Expand-and-randomize: An algebraic approach to secure computation. Available at arXiv: 2001.00539.

[13] 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.

[14] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson. Private information retrieval with side information. *IEEE Trans. on Info. Theory*, 66(4):2032–2043, April 2020.

[15] S. Kadhe, A.Heidarzadeh, A.Sprintson, and O. O. Koyluoglu. On an equivalence between single-server PIR with side information and locally recoverable codes. Available at arXiv:1907.00598.

[16] 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*, pages 180–187, October 2018.

[17] S. Li and M. Gastpar. Single-server multi-message private information retrieval with side information. In *Allerton Conference*, pages 173–179, October 2018.

[18] N. B. Shah, K. V. Rashmi, and K. Ramchandran. One extra bit of download ensures perfectly private information retrieval. In *IEEE ISIT*, pages 856–860, June 2014.

[19] S. Vithana, K. Banawan, and S. Ulukus. Semantic private information retrieval. Available at arXiv:2003.13667.

[20] 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.

[21] 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.

[22] 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*, 66(8):4904–4916, August 2020.

[23] Q. Wang and M. Skoglund. Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers. *IEEE Trans. on Info. Theory*, 65(8):5160–5175, August 2019.

[24] Q. Wang, H. Sun, and M. Skoglund. Symmetric private information retrieval with mismatched coded messages and randomness. In *IEEE ISIT*, pages 365–369, July 2019.

[25] 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.

[26] M. J. Siavoshani, S. P. Shariatpanahi, and M. A. Maddah-Ali. Private information retrieval for a multi-message scenario with private side information. In *IEEE Trans. on Communications*, 2021. Early Access.

[27] 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.

[28] Q. Wang and M. Skoglund. On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers. *IEEE Trans. on Info. Theory*, 65(5):3183–3197, May 2019.

[29] X. Yao, N. Liu, and W. Kang. The capacity of symmetric private information retrieval under arbitrary collusion and eavesdropping patterns. Available at arXiv:2010.08249.

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

[31] 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.

[32] 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.

[33] S. Kumar, A. G. i Amat, E. Rosnes, and L. Senigagliesi. Private information retrieval from a cellular network with caching at the edge. *IEEE Trans. on Communications*, 67(7):4900–4912, July 2019.

[34] 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.

[35] 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.

[36] 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.

[37] I. Samy, R. Tandon, and L. Lazos. On the capacity of leaky private information retrieval. In *IEEE ISIT*, pages 1262–1266, July 2019.

[38] K. Banawan and S. Ulukus. Private information retrieval through wiretap channel II: Privacy meets security. *IEEE Trans. on Info. Theory*, 66(7):4129–4149, July 2020.

[39] 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.

[40] 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.

[41] 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*, 66(6):3407–3416, June 2020.

[42] C. Tian, H. Sun, and J. Chen. Capacity-achieving private information retrieval codes with optimal message size and upload cost. *IEEE Trans. on Info. Theory*, 65(11):7613–7627, November 2019.

[43] Y. Zhou, Q. Wang, H. Sun, and S. Fu. The minimum upload cost of symmetric private information retrieval. In *IEEE ISIT*, pages 1030–1034, June 2020.

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

[45] C. Tian. On the storage cost of private information retrieval. *IEEE Trans. on Info. Theory*, 66(12):7539–7549, December 2020.

[46] K. Banawan and S. Ulukus. Private information retrieval from non-replicated databases. In *IEEE ISIT*, pages 1272–1276, July 2019.

[47] N. Raviv, I. Tamo, and E. Yaakobi. Private information retrieval in graph-based replication systems. *IEEE Trans. on Info. Theory*, 66(6):3590–3602, June 2020.

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

[50] 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.

[51] Z. Chen, Z. Wang, and S. A. Jafar. The asymptotic capacity of private search. *IEEE Trans. on Info. Theory*, 66(8):4709–4721, August 2020.

[52] M. Naor and B. Pinkas. Oblivious transfer and polynomial evaluation. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, page 245–254. ACM, 1999.