Symmetric Private Information Retrieval with User-Side Common Randomness

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

Sennur Ulukus  
zhusheng@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 $(d, \rho_S, \rho_U)$, where $d$ is the download cost, $\rho_S$ is the amount of shared database (server) common randomness, and $\rho_U$ is the amount of available user-side common randomness. We show that with a suitable amount of $\rho_U$, this new SPIR achieves the capacity of conventional PIR. As a corollary, single-database SPIR becomes feasible. Further, the presence of user-side $\rho_U$ reduces the amount of required server-side $\rho_S$.

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 $C_{\text{PIR}} = \frac{1 - \frac{1}{N}}{1 - \frac{1}{N}}$ and $C_{\text{SPIR}} = 1 - \frac{1}{N}$. First, $C_{\text{SPIR}}$ is smaller than $C_{\text{PIR}}$, 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 $C_{\text{PIR}} = \frac{1}{N}$ 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 $[5], [6]$. 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., $[13]–[51]$.

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

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.,

\[
H(W_k) = L, \quad k \in [K]
\]

\[
H(W_{1:K}) = H(W_1) + \cdots + H(W_K) = KL
\]

As in [4], we use a random variable \( F \) to denote the randomness in the retrieval strategy implemented by the user. Due to the user privacy constraint, the realization of \( F \) 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 \( R_S \); we will call this server-side common randomness. Before the retrieval process starts, the user obtains a partial knowledge of \( R_S \) and all the random variables in the server, we obtain the following equivalent expression for user privacy for all potential query realizations \( q \),

\[
\text{[user privacy]} \quad P(Q_n^{[k,R_U]} = q) = P(Q_n^{[k',R_U]} = q)
\]

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

\[
\text{[database privacy]} \quad I(W_k; F, A_{1:N}^{[k,R_U]}, R_U) = 0
\]

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,

\[
I(R_S \setminus R_U; F, A_{1:N}^{[k,R_U]}, W_k, R_U) = 0
\]

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 \( \rho_S \), and the normalized user-side common randomness as \( \rho_U \) as,

\[
d = \frac{D}{L}, \quad \rho_S = \frac{H(R_S)}{L}, \quad \rho_U = \frac{H(R_U)}{L}
\]

where \( L \) is the message length. Our goal in this paper is to determine the largest region for simultaneously achievable triples \( (d, \rho_S, \rho_U) \) 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 \( (d, \rho_S, \rho_U) \).

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

\[
d \geq 1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}
\]

\[
\rho_S - \rho_U \geq \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}
\]

\[
\frac{N - 1}{N} - d + \rho_U \geq 1
\]

\[
\frac{N}{N - 1} \rho_U + N \rho_S \geq \frac{N}{N - 1}
\]

**Remark 1** The right hand side of (13) is the optimum normalized download cost of classical PIR, \( d_{\text{PIR}} \) [3]. Thus, (12) states that \( d \geq d_{\text{PIR}} \). When \( \rho_U = 0 \), i.e., when there is no user-side common randomness, (15) becomes \( d \geq d_{\text{SPIR}} \), where \( d_{\text{SPIR}} = \frac{N - 1}{N - 1} \) is the optimum normalized download cost of classical SPIR [3]. Note that \( d_{\text{SPIR}} > d_{\text{PIR}} \) for all \( N \). Therefore, when \( \rho_U = 0 \), (12) is binding, (13) is loose, and we have \( d = d_{\text{SPIR}} \). Theorem 1 implies that with appropriate \( \rho_U \), e.g., with \( \rho_U = \frac{1}{N} \), both (12) and (13) can be made binding, at which time the new SPIR download cost achieves \( d = d_{\text{PIR}} \).

**Remark 2** When \( \rho_U = 0 \), Theorem 1 reduces to the capacity region of classical SPIR [4], as in this case, (15) gives \( d \geq \frac{N}{N - 1} \). (16) gives \( \rho_S \geq \frac{N}{N - 1} \), and (13) and (14) are non-binding.
Remark 3 The gap between $\rho_S$ and $\rho_U$ 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., $\mathcal{R}_S \setminus \mathcal{R}_U$, is utilized to hide the undesired messages.

Remark 4 From (15), 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 $d = d_{\text{SPIR}} = 2$ is achieved by $\rho_S = 1$ [4]. In Theorem 1 $d = 2$ can be achieved by $\rho_S = \frac{1}{2}$ 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

$$d \geq K \quad (17)$$

$$\rho_S - \rho_U \geq K - 1 \quad (18)$$

$$\rho_U \geq 1 \quad (19)$$

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], [7], 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 $\rho_S$ and $\rho_U$ 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

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

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

| $\mathcal{R}_U$ | Desired message: $W_1$ | Desired message: $W_2$ |
|----------------|------------------------|------------------------|
| $S_1$          | $W_1 + S_1$            | $W_2 + S_1$            |
| $W_2 + S_2$    | $W_3 + S_2$            | $W_1 + S_2$            |
| $W_3 + S_3$    | $W_1 + S_3$            | $W_3 + S_3$            |
| $S_2$          | $W_1 + S_2$            | $W_2 + S_2$            |
| $W_2 + S_3$    | $W_3 + S_3$            | $W_1 + S_3$            |
| $W_3 + S_1$    | $W_1 + S_1$            | $W_3 + S_1$            |
| $S_3$          | $W_1 + S_3$            | $W_2 + S_3$            |
| $W_2 + S_1$    | $W_3 + S_1$            | $W_1 + S_1$            |
| $W_3 + S_2$    | $W_1 + S_2$            | $W_2 + S_1$            |

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.

#### 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 $4! \cdot 3 = 288$ 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.

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

| $\mathcal{R}_U$ | Desired message: $W_1$ | Desired message: $W_2$ |
|----------------|------------------------|------------------------|
| $DB1$          | $a_1 + S_1$            | $a_2 + S_1$            |
| $b_1 + S_2$    | $b_2 + S_1$            | $b_1 + S_2$            |
| $a_3 + S_3$    | $a_4 + b_3$            | $b_3 + a_2 + S_3$      |
| $a_5 + S_5$    | $a_6 + S_6$            | $b_1 + S_1$            |
| $b_1 + S_5$    | $b_2 + S_3$            | $a_1 + S_2$            |
| $a_3 + S_4$    | $a_4 + b_2 + S_4$      | $b_1 + a_3 + S_3$      |
| $a_5 + S_2$    | $a_6 + S_6$            | $b_1 + S_1$            |
| $b_1 + S_3$    | $b_2 + S_4$            | $a_1 + S_2$            |
| $a_3 + S_4$    | $a_4 + b_2 + S_4$      | $b_1 + a_3 + S_3$      |
| $a_5 + S_2$    | $a_6 + S_6$            | $b_1 + S_1$            |
| $b_1 + S_3$    | $b_2 + S_4$            | $a_1 + S_2$            |
| $a_3 + S_4$    | $a_4 + b_2 + S_4$      | $b_1 + a_3 + S_3$      |
| $a_5 + S_2$    | $a_6 + S_6$            | $b_1 + S_1$            |
| $b_1 + S_3$    | $b_2 + S_4$            | $a_1 + S_2$            |
| $a_3 + S_4$    | $a_4 + b_2 + S_4$      | $b_1 + a_3 + S_3$      |
| $a_5 + S_2$    | $a_6 + S_6$            | $b_1 + S_1$            |
| $b_1 + S_3$    | $b_2 + S_4$            | $a_1 + S_2$            |
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 3-6, 9 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: $R_S$ and $R_U$ with respect to techniques in [3], and $R_U$ with respect to techniques in [4].

Lemma 1

$$I(W_{2;K}; Q_{1;N}^{[1,R_U]}, A_{1;N}^{[1,R_U]}, R_S | W_1) \leq D - L$$  (21)

Lemma 2

$$I(W_{K;K}; Q_{1;N}^{[k-1,R_U]}, A_{1;N}^{[k-1,R_U]}, R_S | W_{1:k-1}) \geq \frac{1}{N} I(W_{k+1;K}; Q_{1;N}^{[k,R_U]}, A_{1;N}^{[k,R_U]}, R_S | W_{1:k}) + \frac{L}{N}$$  (22)

Lemma 3 (Minimal download cost $d$)

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

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

$$I(W_{2;K}; Q_{1;N}^{[1,R_U]}, A_{1;N}^{[1,R_U]}, R_S | W_1) \geq \left( \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \right) L$$  (24)

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

Lemma 4 (Minimal difference between $\rho_S$ and $\rho_U$)

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

Proof: From 24, we have the following relation,

$$H(W_{2;K}; Q_{1;N}^{[1,R_U]}, A_{1;N}^{[1,R_U]}, W_1, R_S) \leq (K - 1) L - \left( \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}} \right) L$$  (26)

Next, we have the following upper bound,

$$I(W_{2;K}; R_S \setminus R_U | Q_{1;N}^{[1,R_U]}, A_{1;N}^{[1,R_U]}, W_1, R_U) \leq H(R_S \setminus R_U | Q_{1;N}^{[1,R_U]}, A_{1;N}^{[1,R_U]}, W_1, R_U)$$  (27)

and the following lower bound,

$$I(W_{2;K}; R_S \setminus R_U | Q_{1;N}^{[1,R_U]}, A_{1;N}^{[1,R_U]}, W_1, R_U) \geq (K - 1) L - H(W_{2;K}; Q_{1;N}^{[1,R_U]}, A_{1;N}^{[1,R_U]}, W_1, R_S)$$  (29)

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

Combining 28 and 30 yields the desired result. 

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

$$H(A_n^{[k,R_U]} | Q_n^{[k,R_U]}, W_k, R_U) \geq H(A_n^{[k',R_U]} | Q_n^{[k',R_U]}, W_k, R_U) - H(R_U)$$  (31)

Lemma 6

$$H(A_n^{[k,R_U]} | Q_n^{[k,R_U]}, R_U) = H(A_n^{[k',R_U]} | Q_n^{[k',R_U]}, R_U')$$  (32)

Lemma 7

$$H(A_n^{[k,R_U]} | F, Q_n^{[k,R_U]}, W_k, R_U) = H(A_n^{[k,R_U]} | Q_n^{[k,R_U]}, W_k, R_U)$$  (33)

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

$$H(A_n^{[k',R_U]} | Q_n^{[k',R_U]}, R_U) = H(A_n^{[k',R_U]} | Q_n^{[k',R_U]}, W_k, R_U')$$  (34)

Lemma 9 (Minimal bound for $d$ and $\rho_U$)

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

Proof: Starting from the message length assumption 1, 

$$L = H(W_k) \geq H(W_k | F, R_U)$$  (36)

$$H(W_k | F, R_U) = H(W_k | F, A_{1;N}^{[k,R_U]}, R_U)$$  (37)

$$= I(W_k; A_{1;N}^{[k,R_U]} | F, R_U)$$  (38)

$$= H(A_{1;N}^{[k,R_U]} | F, R_U) - H(A_{1;N}^{[k,R_U]} | F, W_k, R_U)$$  (39)

$$\leq H(A_{1;N}^{[k,R_U]} | F, R_U) - H(A_{1;N}^{[k,R_U]} | F, Q_n^{[k,R_U]}, W_k, R_U)$$  (40)

$$\leq H(A_{1;N}^{[k,R_U]} | F, R_U) - H(A_{1;N}^{[k,R_U]} | Q_n^{[k,R_U]}, W_k, R_U')$$  (41)

$$+ H(R_U)$$  (42)

$$= H(A_{1;N}^{[k,R_U]} | F, R_U) - H(A_{1;N}^{[k,R_U]} | Q_n^{[k,R_U]}, W_k, R_U')$$  (43)

$$+ H(R_U)$$  (44)
where (66) follows from the steps between (41)-(44) by Lemma 10 (Minimal bound for $\rho_U$ and $\rho_S$)

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

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

$$0 = I(W_k; F, A_{1:N}^{[k, R_U]}, R_U)$$

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

$$NL \leq NH(A_{1:N}^{[k, R_U]}|F, R_U) - \sum_{n=1}^{N} H(A_n^{[k, R_U]}|F, R_U) + NH(R_U) \leq (N-1)H(A_{1:N}^{[k, R_U]}|F, R_U) + NH(R_U) \leq (N-1)D + NH(R_U)$$

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 subtracted 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 II without $S_3$.

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 $|R_S|$) to each common randomness index (e.g., $S_1$ becomes $S_2$ in Table II). Repeat it $|R_S|$ times such that each query cell has a different seed common randomness index.

4) Query cell determination: The user has $|R_U|$ 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 $|R_S|$ is $1 + \cdots + N^{K-1}$ and the total amount of user-side common randomness required $|R_U|$ is 1.
