The Capacity of $T$-Private Information Retrieval with Private Side Information

Zhen Chen, Zhiying Wang, and Syed Ali Jafar

Abstract—We consider the problem of $T$-Private Information Retrieval with private side information (TPIR-PSI). In this problem, $N$ replicated databases store $K$ independent messages, and a user, equipped with a local cache that holds $M$ messages as side information, wishes to retrieve one of the other $K - M$ messages. The desired message index and the side information must remain jointly private even if any $T$ of the $N$ databases collude. We show that the capacity of TPIR-PSI is $C_{\text{TPIR}} = \Psi(T/N, K)$. The capacity of symmetric PIR (SPIR), where the user learns nothing about the database besides his desired message, was shown in [14] to be $C_{\text{SPIR}} = \Psi(1/N, \infty)$, and the capacity of STPIR, with both symmetric privacy and robustness against collusion among any $T$ servers, was characterized in [15] as $C_{\text{STPIR}} = \Psi(T/N, \infty)$. A number of other variants of PIR have also been investigated, such as PIR with MDS coded storage [12], multi-message PIR [16], multi-round PIR [17], secure PIR [18], and PIR with side information [19]–[29]. Especially relevant to this work is the problem of PIR with side information.

The recent focus on the capacity of PIR with side information started with the work on cache-aided PIR by Tandon [19], where the user has enough local cache memory to store a fraction $r$ of all messages as side information. In this model, the side information can be any function of the $K$ messages (subject to the storage constraint) and is globally known to both the user and all the databases. The capacity for this setting is characterized in [19] as $\Psi(1/N, K)/(1 - r)$.

Different from [19] which allows side information to be an arbitrary function of the messages, the side information in [20] (and in this paper) can only take the form of $M$ full messages cached by the user. Within this model there are several interesting variations depending on the constraints on the privacy of the side information.

- PIR-GSI, or PIR with global side information, implies that the side information is globally known.
- PIR-SI, i.e., PIR with (non-private) side information, corresponds to the case that the side information is not globally known, but the privacy of the side information need not be preserved.
- PIR-PSI, or PIR with private side information, refers to the setting where the joint privacy of both the desired message and the side information must be preserved. This is the focus of the paper.
- PIR-PSI, or PIR with separately private side information, refers to the setting where the privacy of the desired message and the privacy of side information must each be separately preserved (although their joint privacy need not be preserved). In Appendix A we provide some insights into the capacity of PIR-PSI.

Out of these four settings, PIR-GSI is rather trivial, and PIR-PSI has not been studied at all, perhaps because there is insufficient practical motivation for such an assumption. However, the remaining two variants, PIR-PSI and PIR-SI, have indeed drawn much attention, starting with the work of...
Kadhe et al. in [20].

For PIR-SI with a single database \((N = 1)\), Kadhe et al. showed in [20] that the capacity is \(\Psi(1/N, [K/(M + 1)])\). The single-database setting has seen rapid progress in various directions [23]–[29]. However, PIR-SI with multiple databases turns out to be considerably more challenging. In [20], Kadhe et al. provided an achievable scheme for PIR-SI with multiple databases \((N > 1)\), which achieves the rate \(\Psi(1/N, [K/(M + 1)])\). In spite of some progress in this direction [27], the capacity of PIR-SI generally remains open\(^1\) for multiple databases. In addition, the works in [21], [22] consider a different form of side information instead of full messages.

For PIR-PSI with a single database, Kadhe et al. found in [20] that the capacity is \((K - M)^{-1}\). The capacity of PIR-PSI with more than one database was left as an open problem in [20]. Remarkably, neither a general achievable scheme nor a converse was known in this case. It is this open problem that motivates this work.

The first contribution of this work is to show that the capacity of PIR-PSI is \(C_{\text{PIR-PSI}} = \Psi(1/N, K - M)\), for any arbitrary number of databases \(N\), thus settling this open problem. This allows us to completely order\(^2\) the four variants of PIR with side information that are listed above, in terms of their capacities as PIR-PSI \(\geq\) PIR-SPSI \(\geq\) PIR-PSI = PIR-GSI. Remarkably, all the inequalities can be strict for certain parameters.

As a generalization, we show that the capacity of TPIR-PSI, i.e., PIR where up to \(T\) databases may collude, is \(C_{\text{TPIR-PSI}} = \Psi(T/N, K - M)\). Evidently, the effect of private side information on capacity is the same as if the number of messages in TPIR was reduced from \(K\) to \(K - M\) [13]. Similar to the case with non-colluding databases, this is also the capacity if the side information is globally known to all databases as well.

As the second contribution of this work, we characterize the capacity of STPIR-PSI, i.e., PIR with private side information with symmetric privacy and robustness against any \(T\)-colluding servers. We show \(C_{\text{STPIR-PSI}} = \Psi(T/N, \infty)\), provided that the databases have access to common randomness (not available to the user) in the amount that is at least \(T/(N - T)\) bits per queried message bit. Otherwise, the capacity of STPIR-PSI is zero. Note that this is identical to the capacity of STPIR with no side information [15].

The remainder of this paper is organized as follows. Section II presents the problem statements. Section III presents the main results, i.e., the capacity characterizations of TPIR-PSI and STPIR-PSI. The proofs of the capacity results are presented in Section IV and Section V, and we conclude with Section VI.

**Notation:** We use bold font for random variables to distinguish them from deterministic variables, that are shown in normal font. For integers \(z_1 < z_2\), \([z_1 : z_2]\) represents the set \(\{z_1, z_1 + 1, \ldots, z_2\}\) and \((z_1 : z_2)\) represents the vector \((z_1, z_1 + 1, \ldots, z_2)\). The compact notation \([z]\) represents \([1 : z]\) for positive integer \(z\). For random variables \(W_i, i = 1, 2, \ldots, N\), and a set of positive integers \(S = \{s_1, s_2, \ldots, s_n\}\), where \(s_1 < s_2 < \cdots < s_n\), the notation \(W_S\) represents the vector \((W_{s_1}, W_{s_2}, \ldots, W_{s_n})\). For a matrix \(G\) and a vector \(S\), the notation \(G[S, :]\) represents the submatrix of \(G\) formed by retaining only the rows corresponding to the elements of the vector \(S\). For a matrix \(G\), its transpose is denoted as \(G^t\). \(F_q\) represents the finite field of size \(q\).

**II. Problem Statements**

**A. TPIR-PSI: T-Private Information Retrieval with Private Side Information**

The TPIR-PSI problem is parametrized by \((K, M, N, T)\). Consider \(K\) independent messages \(W_K = (W_1, \ldots, W_K)\), each containing \(L\) independent and uniform bits, i.e., their entropy satisfies

\[
H(W_1, \ldots, W_K) = H(W_1) + \cdots + H(W_K) \quad (1)
\]

\[
H(W_i) = \cdots = H(W_K) = L. \quad (2)
\]

There are \(N\) databases and each database stores all \(K\) messages \(W_1, \ldots, W_K\). A user is equipped with a local cache and has \(M (M < K)\) messages as side information. Let \(S = \{i_1, i_2, \ldots, i_M\}\) be \(M\) distinct indices chosen uniformly from \([K]\). These \(M\) cached messages are represented as \(W_S = (W_{i_1}, \ldots, W_{i_M})\). \(S\) is not known to the databases. A user wishes to retrieve \(W_{\Theta}\), where \(\Theta\) is a message index uniformly chosen from \([K]\) \(\setminus S\), as efficiently as possible, while revealing no information about \((\Theta, S)\) to any colluding subsets of up to \(T\) out of the \(N\) databases. Note the following independence,

\[
H(\Theta, S, W_1, \ldots, W_K) = H(\Theta, S) + \sum_{i=1}^{K} H(W_i). \quad (3)
\]

In order to retrieve \(W_{\Theta}\), the user generates \(N\) queries \(Q_{1}^{(\Theta, S)}, \ldots, Q_{N}^{(\Theta, S)}\) with the knowledge of \((\Theta, S, W_S)\). Since the queries are generated with no knowledge of the other \(K - M\) messages, the queries must be independent of them,

\[
I \left( \Theta, S, W_S, Q_{1}^{(\Theta, S)}, \ldots, Q_{N}^{(\Theta, S)}; W_K \mid S \right) = 0. \quad (4)
\]

The user sends query \(Q_{n}^{(\Theta, S)}\) to the \(n\)th database and in response, the \(n\)th database returns an answer \(A_{n}^{(\Theta, S)}\) which is a deterministic function of \(Q_{n}^{(\Theta, S)}\) and \(W_{[K]}\),

\[
H \left( A_{n}^{(\Theta, S)} \mid Q_{n}^{(\Theta, S)}, W_{1}, \ldots, W_{K} \right) = 0. \quad (5)
\]

Upon collecting the answers from all \(N\) databases, the user must be able to decode the desired message \(W_{\Theta}\) based on the queries and side information,

\[
H \left( W_{\Theta} \mid A_{[K]}^{(\Theta, S)}, Q_{[K]}^{(\Theta, S)}, W_{S}, S, \Theta \right) = 0. \quad (6)
\]

To satisfy the user-privacy constraint that any \(T\) colluding databases learn nothing about \((\Theta, S)\), the information available to any \(T\) databases (queries, answers and stored messages)
must be independent of \((\Theta, S)\). Let \(T\) be any subset of \([1 : N]\), of cardinality \(|T| = T\), \(Q^{(\Theta, S)}_T\) represents the vector of queries corresponding to \(Q^{(\Theta, S)}_n, n \in T\). \(A^{(\Theta, S)}_T\) is defined as the answer vector corresponding to \(A^{(\Theta, S)}_n, n \in T\). To satisfy the \(T\)-privacy requirement we must have \(\forall T \subset [1 : N], |T| = T\).

[User privacy] \(I\left(\Theta, S; Q^{(\Theta, S)}_T, A^{(\Theta, S)}_T, W_{[K]}\right) = 0\).  

A TPIR-PSI scheme is called feasible if it satisfies the correctness constraint (6) and the user-privacy constraint (7). For a feasible scheme, the TPIR-PSI rate indicates asymptotically how many bits of desired information are retrieved per downloaded bit, and is defined as follows.

\[
R_{\text{TPIR-PSI}} \triangleq \lim_{L \to \infty} \frac{L}{D},
\]

where \(D\) is the expected (over all \(\Theta, S, W_{[K]}\) and random queries) total number of bits downloaded by the user from all the databases. The capacity, \(C_{\text{TPIR-PSI}}\), is the supremum of \(R_{\text{TPIR-PSI}}\) over all feasible schemes.

B. STPIR-PSI: Symmetric T-Private Information Retrieval with Private Side Information

In symmetric \(T\)-colluding private information retrieval, an additional constraint is imposed: database privacy, which means that the user does not learn any information about \(W_{[K]}\) beyond the retrieved message, \(W_{\Theta}\), and the side information, \(W_S\). To facilitate database privacy, suppose the databases share a common random variable \(U\) that is not known to the user. It has been shown that without such common randomness, symmetric PIR is not feasible when there is more than one message [6], [14]. The common randomness is independent of the messages, the desired messages index, and the side information index, so that

\[
H(\Theta, S, W_1, \ldots, W_K, U) = H(\Theta, S) + \sum_{i=1}^{K} H(W_i) + H(U).
\]

The answering string \(A^{(\Theta, S)}_n\) is a deterministic function of \(Q^{(\Theta, S)}_n, W_{[K]}\) and common randomness \(U\).

\[
H\left(\Theta^{(\Theta, S)}_n W_{[K]} | Q^{(\Theta, S)}_n, W_1, \ldots, W_K, U\right) = 0.
\]

The correctness condition is the same as (6). The user-privacy condition is \(\forall T \subset [1 : N], |T| = T\).

[User privacy] \(I\left(\Theta, S; Q^{(\Theta, S)}_T, A^{(\Theta, S)}_T, W_{[K]}\right) = 0\).  

Database privacy requires that the user learns nothing about \(W_{[\Theta, S]} = W_{[K]}((\Theta, S))\), i.e., messages other than his desired message and the side information. Therefore,

\[
\{\text{DB privacy}\} I\left(\Theta, S; Q^{(\Theta, S)}_T, A^{(\Theta, S)}_T, W_{[K]}\right) = 0. \quad (12)
\]

An STPIR-PSI scheme is called feasible if it satisfies the correctness constraint (6), the user-privacy constraint (11) and the database-privacy constraint (12). For a feasible scheme, the STPIR-PSI rate indicates how many bits of desired information are retrieved per downloaded bit. The capacity, \(C_{\text{STPIR-PSI}}\), is the supremum of rates over all feasible STPIR-PSI schemes.

III. MAIN RESULTS

The following theorem presents our first result, the capacity of TPIR-PSI.

**Theorem 1.** For the TPIR-PSI problem with \(K\) messages, \(N\) databases and \(M (M < K)\) side information messages, the capacity is

\[
C_{\text{TPIR-PSI}} = \Psi(T/N, K-M) = \Psi(T/N), \quad (13)
\]

where \(\Psi(A, B) = (1 + A + A^2 + \cdots + A^{B-1})^{-1}\).

The following observations place Theorem 1 in perspective.

**Remark 1.** The expression \(C_{\text{TPIR-PSI}}\) equals the capacity of TPIR with \(K-M\) messages [13]. Evidently, the impact of private side information is equivalent to reducing the effective number of messages from \(K\) to \(K-M\).

**Remark 2.** Remarkably, the capacity expression in (13) matches the capacity for the setting where the side information is assumed to be globally known, i.e., if the \(M\) side information messages are globally known, then the capacity is also \(C_{\text{TPIR-GSI}} = \Psi(T/N, K-M)\). This can be seen as follows. The achievable scheme is the TPIR scheme of [13] after the cached messages are eliminated. To prove the converse by contradiction, suppose the capacity is greater than \(\Psi(T/N, K-M)\). Then there is a scheme \(\Pi\) that achieves a larger rate than \(\Psi(T/N, K-M)\) in the presence of the \(M\) globally known messages. Consider a TPIR problem with \(K-M\) messages and no side information. From [13] we know that its capacity is \(\Psi(T/N, K-M)\). It can be assumed that there are \(M\) globally known dummy messages. With this globally known side information, the user can use scheme \(\Pi\) to retrieve the desired message while achieving a rate larger than \(\Psi(T/N, K-M)\), thus exceeding the capacity of TPIR, i.e., creating a contradiction. Therefore, the capacity of TPIR with globally known side information is \(\Psi(T/N, K-M)\).

**Remark 3.** It is worthwhile to place the previous remark in perspective with the capacity results in [19], where it is also assumed that the side information is globally available. \(C_{\text{TPIR-GSI}}\) is in general less than the capacity expression found in [19]. The reason is that \(C_{\text{TPIR-GSI}}\) is the capacity for a setting where the side information can only be \(M\) full messages (excluding the desired one). However, in [19], the side information is allowed to be any function of all messages. The relaxed
messages, the capacity is \( \Psi(1/N, K)/\left(1 - M/N\right) \). It is easy to verify that \( C_{\text{TPR-GSI}} = \Psi(1/N, K - M)/\left(1 - M/K\right) \) when \( N \geq 2, K \geq 2, M \in [K - 1] \). Aside from this superficial distinction, it is notable that the essential insight in both settings is the same. The best strategy in the setting of [19] is to cache \( \frac{M}{K} \) portion of each message and use the protocol of the original PIR scheme [9] to download the uncached portion. What this means is that if the side information is globally known, then there is nothing better than removing the side information from the effective messages. The expression for \( C_{\text{TPR-GSI}} \) reflects the same insight — the role of globally known side information does not help improve the capacity.

Remark 4. Now we can completely order the four variants of PIR with side information, in terms of their capacities as PIR-GSI \( \geq \) PIR-PSI \( \geq \) PIR-PSI = PIR-GSI. Remarkably, all the inequalities can be strict for certain parameters. For example, as will be shown in Appendix A, suppose we have \( K = 6 \) messages stored at \( N = 1 \) database, and \( M = 2 \) of these messages are available to the user as side-information. Then for this example, the capacity of PIR-GSI is 1/2 while the capacity of PIR-PSI is no more than 1/3, so that PIR-GSI \( > \) PIR-PSI. Now suppose we have \( K = 6 \) messages stored at \( N = 1 \) database, and \( M = 1 \) of these messages is available to the user as side-information. Then for this example, the capacity of PIR-PSI is 1/3 while the capacity of PIR-PSI is only 1/5, so that PIR-PSI \( > \) PIR-PSI.

Our second result is the capacity of STPIR-PSI, presented in the following theorem.

**Theorem 2.** For the STPIR-PSI problem with \( K \geq 2 \) messages, \( N \) databases and \( M (M < K) \) side information messages, the capacity is

\[
C_{\text{STPIR-PSI}} = \begin{cases} 
1, & \text{if } M = K - 1, \\
1 - \frac{\ell}{N}, & \text{if } M < K - 1 \text{ and } \rho \geq \frac{T}{N-T}, \\
0, & \text{otherwise},
\end{cases}
\]

where \( \rho = \frac{H(M)}{L} \) is the amount of common randomness available to the databases, normalized by the message size.

The following observations are in order.

**Remark 5.** When there is only \( K = 1 \) message, or when there are \( M = K = 1 \) side information messages, the database-privacy constraint is satisfied trivially, so STPIR reduces to the PIR setting and the capacity is 1. Note that for symmetric PIR without side information, when \( K \geq 2 \), the common randomness is necessary for feasibility. However, for STPIR-PSI, if there are \( M = K - 1 \) side information messages, then common randomness is not needed.

**Remark 6.** When \( K \geq 2 \) and \( M < K - 1 \), then \( C_{\text{STPIR-PSI}} \) only depends on the number of databases \( N \), the colluding parameter \( T \), and the amount of common randomness. It is independent of the number of messages \( K \) and the number of side information messages \( M \).

**Remark 7.** The capacity of STPIR-PSI is strictly smaller than the capacity of PIR-PSI, which means that the additional requirement of preserving database privacy strictly penalizes the capacity. However, the penalty vanishes in the regime of large number of messages, i.e., \( C_{\text{TPR-PSI}} > C_{\text{STPIR-PSI}} \) for any finite \( K \) and \( C_{\text{TPR-PSI}} \rightarrow C_{\text{STPIR-PSI}} \) when \( K \rightarrow \infty \). This observation also holds for the case without side information.

**Remark 8.** \( C_{\text{STPIR-PSI}} \) is equal to the capacity of STPIR without side information, which is characterized in [30]. Furthermore, the capacity result remains the same even if the side information is globally known.

Thus, utilizing the private or globally known side information does not help improve the capacity.

**IV. PROOF OF THEOREM 1**

**A. Achievability**

The backbone of the achievable scheme for PIR-PSI with parameters \((K, M, N, T)\) is the achievable scheme of PIR [13]. We inherit the steps of the query structure construction and query specialization. The novel element of the achievable scheme is query redundancy removal based on the side information. To illustrate how this idea works, we present one toy example with \((K, M, N, T) = (3, 2, 3, 2)\), and then generalize it to arbitrary \((K, M, N, T)\).

1) **Example with** \((K, M, N, T) = (3, 2, 3, 2)\): Let us start with the case without side information \((K, M, N, T) = (3, 0, 3, 2)\), i.e., there are 3 messages, 3 databases and any 2 of them can collude. Following the construction of [13], let each message consist of \( L = N^K = 27 \) symbols from a finite field \( \mathbb{F}_q \) that is large enough so that a systematic (28,19) maximum distance separable (MDS) code exists. The MDS property implies that any 19 out of the 28 codeword symbols is sufficient to recover all 19 information symbols. A systematic code is a code in which the information symbols are(embedded in the codeword symbols) [31]. According to the query structure construction and query specialization for PIR [13], the messages \( W_1, W_2, W_3 \in \mathbb{F}_q^{27} \) are \( 27 \times 1 \) column vectors and let \( Y_1, Y_2, Y_3 \in \mathbb{F}_q^{27 \times 27} \text{ represent random matrices chosen privately by the user, independently and uniformly from all } 27 \times 27 \text{ full-rank matrices over } \mathbb{F}_q \). Let \( G_{e \times f} \) denote the generator matrix of an \((e, f)\) MDS code (e.g., a Reed Solomon code), for some integers \( e, f \). The generator matrices need not be systematic or random, and may be globally known. Define the \( 27 \times 1 \) column vectors \( a_{(1:27)}, b_{(1:27)}, c_{(1:27)} \in \mathbb{F}_q^{27} \) as follows.

\[
a_{(1:27)} = Y_1 W_1, \\
b_{(1:18)} = G_{18 \times 12} Y_2 [(1 : 12), :] W_2, \\
c_{(1:18)} = G_{18 \times 12} Y_3 [(1 : 12), :] W_3, \\
b_{(19:27)} = G_{9 \times 9} Y_4 [(13 : 18), :] W_2, \\
c_{(19:27)} = G_{9 \times 9} Y_5 [(13 : 18), :] W_3,
\]

where \( Y_d[(1 : 18), :] \) and \( Y_d[(1 : 18), :] \) are \( 18 \times 27 \) matrices comprised of the first 18 rows of \( Y_d \) and \( Y_d \), respectively.

\[a_{(1:27)} = Y_1 W_1,\]
\[b_{(1:18)} = G_{18 \times 12} Y_2 [(1 : 12), :] W_2,\]
\[c_{(1:18)} = G_{18 \times 12} Y_3 [(1 : 12), :] W_3,\]
\[b_{(19:27)} = G_{9 \times 9} Y_4 [(13 : 18), :] W_2,\]
\[c_{(19:27)} = G_{9 \times 9} Y_5 [(13 : 18), :] W_3,\]
Note that the same generator matrix $G_{18 \times 12}$ is used in (16) and (17), and the same generator matrix $G_{9 \times 6}$ is used in (18) and (19).

The downloaded symbols from each database are represented in Table I. We use $DB_i$ to represent the $i$th database. It correctly recovers the queried message and maintains user privacy even if 2 databases collude. The achieved rate is $R_{\text{tpr}} = 9/19$, namely, in this scheme the user recovers 9 desired symbols from a total of 19 downloads symbols from each database.

Now let us consider the case with side information $(K, M, N, T) = (3, 2, 3, 2)$, i.e., 2 of the messages are known to the user as side information. Assume the user knows $W_2$ and $W_3$ as side information and wishes to retrieve $W_1$. He does not need to download individual symbols of $W_2$, $W_3$, or the linear combinations comprised of only $W_2$, $W_3$ symbols, i.e., $b_i, c_i, 1 \leq i \leq 12$ and $b_j + c_j, 19 \leq j \leq 24$ in Table I. Therefore, 10 redundant symbols may be reduced from each database. Let us take the step of query redundancy removal.

The idea is that the user asks each database to encode the 19 original downloaded symbols with a systematic $(28, 19)$ MDS code and downloads only the 9 linear combinations corresponding to the non-systematic part, called parity symbols. Formally, let $G_{i,e,f}^s$ denote the generator matrix of a systematic $(e,f)$ MDS code. The generator matrix does not need to be random, and it may be globally known. For $i = 1, 2, 3$, denote by vector $X_i \in \mathbb{F}_q^{19}$ the symbols downloaded from $DB_i$ after the query structure construction and query specialization (symbols in the $DB_i$ column in Table I). The user asks each database to encode $X_i$ with a systematic $(28, 19)$ MDS code generator matrix $G_{28 \times 19}^s = [V_{19 \times 9} I_{19 \times 19}]$, where $I_{19 \times 19}$ is the identity matrix, and downloads only the 9 linear combinations corresponding to the parity part, $V_{19 \times 9} X_i$.

The correctness constraint is satisfied because of the property of MDS code and the correctness of the original TPIR scheme. Given $(b_j)_{j=0}^{12}, (c_j)_{j=13}^{24}$, $b_i + c_i, 1 \leq i \leq 12$ and $b_j + c_j, 19 \leq j \leq 24$ in Table I, the user is able to decode $X_1$, $X_2$ and $X_3$, which constitute the original TPIR scheme. The privacy is essentially inherited from the original PIR scheme and the fact that the MDS code is fixed a priori, i.e., it does not depend on $(\Theta, S)$. Thus, the rate achieved with private side information is $R_{\text{tpr},ps} = 27/27 = 1$ which gives a lower bound on the capacity.

2) Arbitrary $(K, M, N, T)$: Scheme description. For the sake of completeness, let us briefly introduce the original TPIR achievable scheme in [13]. In this scheme, the message is $L = N^K$ symbols from a large enough finite field $\mathbb{F}_q$, and the normalized total download is $1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-1}$. It has two key steps: 1) query structure construction and 2) query specialization.

1) Query Structure Construction: Construct the query structure. After this step, the query of each database is comprised of $K$ layers. Over the $k$th layer, the query symbols are in the form of sums of $k$ message symbols, each from one distinct message, called $k$-sum. There are $\binom{K}{k}$ possible “types” of $k$-sums and $(N - T)^{k-1}T^{K-k}$ distinct instances of each type of $k$-sum in $k$th layer. So, the total number of elements contained in layer $k$ is $\binom{K}{k}(N - T)^{k-1}T^{K-k}$. Therefore, the total number of symbols to be downloaded from each database is $\sum_{k=1}^{K} \binom{K}{k}(N - T)^{k-1}T^{K-k}$. This structure has two properties: symmetry across databases and message symmetry within the query from each database. Symmetry across databases means that the queries among the databases have the same structure (i.e., the same form of $k$-sums). Message symmetry implies that within the query of each database, any set of $M$ messages determines the same number of $k$-sums, $1 \leq k \leq M$.

2) Query Specialization: Map the message symbols to the symbols in the query structure. This step is to ensure the correctness and privacy.

Now we are ready to present the achievable scheme for arbitrary $(K, M, N, T)$. First do query structure construction and query specialization without considering the side information, and denote the scheme by $\Pi$. Then do query redundancy removal based on the side information. Due to symmetry across databases and message symmetry within the query from each database, regardless of the realization of side information, the number of queried symbols and the number of known symbols (based on the side information) in each query are constants. For each database, let $p_1$ denote the number of symbols to be downloaded in $\Pi$. Out of these $p_1$ symbols, let $p_2$ ($p_2 < p_1$) denote the number of user known symbols. Denote by vector $X_i \in \mathbb{F}_q^{19}$ the symbols downloaded from $DB_i$ in $\Pi$. For each database, use a systematic $(2p_1 - p_2, p_1)$ MDS code with generator matrix $G_{2p_1 - p_2 \times p_1}^s = [V_{19 \times (p_1 - p_2)} I_{p_1 \times p_1}]^\prime$ to encode the $p_1$ symbols into $2p_1 - p_2$ symbols, of which $p_1$ are systematic, and download only the $p_1 - p_2$ parity symbols, $V_{19 \times (p_1 - p_2)} X_i$. Note that the user does not need to know the realization of side information $S$ or $W_S$ in order to construct the queries. This is because the systematic MDS code in the query redundancy removal does not depend on $S$ or $W_S$. During the decoding, $S$ and $W_S$ are only used after the answers from the databases are collected. Therefore, the privacy of this TPIR-PSI scheme is inherited from the privacy of the original TPIR scheme. Correctness follows from the MDS property because in addition to the $p_1 - p_2$ downloaded symbols from $DB_i$, i.e., $V_{p_1 \times (p_1 - p_2)} X_i$, the user provides the $p_2$ symbols that he already knows, to obtain a total of $p_1$ symbols from the MDS code. Since any $p_1$ symbols from an MDS code suffice to recover the original $p_1$ symbols, the user recovers $X_i$. Then

$\text{Table I}
\begin{array}{|c|c|c|}
\hline
DB_1 & DB_2 & DB_3 \\
\hline
a_1, a_2, a_3, a_4 & a_5, a_6, a_7, a_8 & a_9, a_{10}, a_{11}, a_{12} \\
b_1, b_2, b_3, b_4 & b_5, b_6, b_7, b_8 & b_9, b_{10}, b_{11}, b_{12} \\
c_1, c_2, c_3, c_4 & c_5, c_6, c_7, c_8 & c_9, c_{10}, c_{11}, c_{12} \\
a_{13} + b_{13} & a_{15} + b_{15} & a_{21} + b_{17} \\
a_{14} + b_{14} & a_{16} + b_{16} & a_{22} + b_{18} \\
a_{17} + c_{13} & a_{19} + c_{15} & a_{23} + c_{17} \\
a_{18} + c_{14} & a_{20} + c_{16} & a_{24} + c_{18} \\
b_{19} + c_{19} & b_{21} + c_{21} & b_{23} + c_{23} \\
b_{20} + c_{20} & b_{22} + c_{22} & b_{24} + c_{24} \\
a_{25} + b_{25} + c_{25} & a_{26} + b_{26} + c_{26} & a_{27} + b_{27} + c_{27} \\
\hline
\end{array}
the correctness is inherited from the correctness of the original TPIR scheme. All that remains is to calculate the rate achieved by this scheme.

**Rate calculation.** Consider the scheme II, the total downloaded symbols from each database \( p_1 = \sum_{k=1}^{K} \binom{K}{k} (N - T)^{k-1} T^{K-k}. \) The next step is to calculate, out of these \( p_1 \) symbols, how many are already known to the user based on his side information. Suppose the user knows the \( M \) messages \( W_{i_1}, \ldots, W_{i_k}, \{i_1, \ldots, i_M\} \in [K] \) as side information beforehand. Thus the user knows all linear combinations that are comprised of symbols from these \( M \) messages. In terms of layer \( k \) \((k \leq M)\), the user knows all the instances of \( k \)-sum that contain only symbols \( W_{j_1}, W_{j_2}, \ldots, W_{j_k} \), where \( \{j_1, j_2, \ldots, j_k\} \subset \{i_1, \ldots, i_M\} \). So the total number of symbols known to the user corresponding to each database is \( p_2 = \sum_{k=1}^{M} \binom{M}{k} (N - T)^{k-1} T^{K-k}. \) Notice that \( p_1 \) can be simplified as,

\[
p_1 = \sum_{k=1}^{K} \binom{K}{k} (N - T)^{k-1} T^{K-k} = N^K - T.
\]

And \( p_2 \) can be simplified as,

\[
p_2 = \sum_{k=1}^{M} \binom{M}{k} (N - T)^{k-1} T^{K-k} = T^K - N^M - T^M - T.
\]

From each database the number of downloaded symbols of desired messages can be calculated as,

\[
m = \sum_{k=1}^{K} (N - T)^{k-1} T^{K-k} = N^{K-1}.
\]

Therefore, the rate achieved is

\[
R_{\text{TPIR-PSI}} = \frac{N m}{N(p_1 - p_2)} = \frac{N^{K-1} - T}{(N^K - T^K) - T^K - N^M - T^M} = \left(1 + \frac{T}{N} \right)^{-1}.
\]

This gives a lower bound on the capacity of TPIR-PSI, thus completing the proof of achievability for Theorem 1.

### B. Converse

Let \( S \) be a set whose elements are all possible realizations of \( S \), i.e., \( S = \{S \mid S \subset [K], |S| = M\} \). We will need the following lemmas.

**Lemma 1.** For all \( S_1 \in S, \theta \in [K] \setminus S_1, S_2 \subseteq [K] \setminus S_1 \), and \( T \subset [N], |T| = T \), given \( S = S_1, \Theta = \theta, A_T^\Theta, S_1 \leftrightarrow Q_T^\Theta, S_1 \leftrightarrow Q_T^\Theta, S_2 \) is a Markov chain.

**Proof.** In this proof, to be convenient, denote \( E_1 = S_1 \cup S_2 \) and \( E_2 = [K] \setminus (S_1 \cup S_2) \). It is equivalent to prove

\[
I \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S_1 \right) = 0.
\]

By the chain rule of mutual information,

\[
I \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S_1 \right)
= I \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S_1 \right) + I \left( W_{E_1} \mid A_T^\Theta, S_1 \right) - I \left( W_{E_1} \mid Q_T^\Theta, S_1 \right) - I \left( A_T^\Theta, S_1 \right).
\]

Therefore,

\[
I \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S_1 \right)
= I \left( W_{E_1} \mid A_T^\Theta, S_1 \right) - I \left( W_{E_1} \mid Q_T^\Theta, S_1 \right) - I \left( A_T^\Theta, S_1 \right)
= 0.
\]

Consider the first RHS mutual information term in (31),

\[
I \left( W_{E_1} \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S_1 \right) = I \left( W_{E_1} \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S_1 \right)
= 0,
\]

where (33) holds because of (1) and (4). The second RHS mutual information term in (31) satisfies

\[
I \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{[K] \setminus (S_1 \cup S_2)}, \Theta = \theta, S = S_1 \right) = 0
\]

because of (5). At last, the RHS negative mutual information term in (31) must also be zero because the LHS mutual information cannot be negative. Thus

\[
I \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S_1 \right) = 0.
\]

**Lemma 2.** For all \( S \in S, \theta, \theta' \in [K] \setminus S \), and \( T \subset [N], |T| = T \),

\[
H \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S \right) = H \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta', S = S \right), \quad \text{(34)}
\]

\[
H \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta, S = S \right) = H \left( A_T^\Theta, S_1 \mid Q_T^\Theta, S_1, W_{E_1}, \Theta = \theta', S = S \right). \quad \text{(35)}
\]
Proof. It follows from the user-privacy constraint (11) and the non-negativity of mutual information, that for all $S \in S$, $T \subset [N], |T| = T$

$$I\left(\Theta; Q^{|\Theta|, S}, A^{|\Theta|, S}, W_{[K]} \mid S = S\right) = 0, \quad (36)$$

which implies that $\forall \theta, \theta' \in [K] \setminus S$,

$$H\left(Q^{|\Theta|, S}, A^{|\Theta|, S}, W, W_s \mid \Theta = \theta, S = S\right) = H\left(Q^{|\Theta|, S}, A^{|\Theta|, S}, W, W_s \mid \Theta = \theta', S = S\right), \quad (37)$$

$$H\left(Q^{|\Theta|, S}, W, W_s \mid \Theta = \theta, S = S\right) = H\left(Q^{|\Theta|, S}, \Theta, S = S\right). \quad (38)$$

Subtracting (38) from (37) yields (34). Equation (35) is similarly obtained.

Before presenting the general converse, let us start with a simple example $(K, M, N, T) = (3, 1, 3, 2)$ for ease of exposition.

1) Converse for $(K, M, N, T) = (3, 1, 3, 2)$: The total download is bounded as,

$$D \geq H(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W, \Theta, S) \quad (39)$$

$$\geq \min_{\theta \in [K] \setminus S} H(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_s, \Theta = \theta, S = S). \quad (40)$$

We will derive a lower bound on the entropy in (40) that holds for all $(\theta, S)$.

For $(K, M, N, T) = (3, 1, 3, 2)$, without loss of generality suppose message $W_1$ is known as side information and $W_2$ is desired. Let $S = 1$. We bound the total download as,

$$D \geq H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_1, \Theta = 2, S = S\right) \quad (41)$$

$$\geq H\left(A^{|\Theta|, S} \mid W_2, Q^{|\Theta|, S}, W_1, \Theta = 2, S = S\right) \quad (42)$$

$$= H\left(W_2 \mid Q^{|\Theta|, S}, W_1, \Theta = 2, S = S\right) \quad (43)$$

where (44) holds because of (2), (4), the chain rule and non-negativity of entropy. Equation (45) holds due to Lemma 1. Equation (46) holds because of Lemma 2. Similarly,

$$D \geq L + H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_2, \Theta = 3, S = S\right), \quad (48)$$

$$D \geq L + H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_2, \Theta = 3, S = S\right). \quad (49)$$

Adding (47), (48), (49) and divided by 3 we have

$$D \geq L + \frac{1}{3} H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_2, \Theta = 3, S = S\right) \quad (50)$$

$$\geq L + \frac{1}{3} H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_2, \Theta = 3, S = S\right) \quad (51)$$

Here (51) follows from Han’s inequality, and (52) holds because from $(W_2, A^{|\Theta|, S}, Q^{|\Theta|, S}, \Theta = 3, S = S)$ one can recover $W_2$ with vanishing probability of error. Since the same argument holds for all realizations $(\Theta, S) = (\theta, S)$, this gives us the upper bound on the capacity of TPIR-PSI with $(K, M, N, T) = (3, 1, 3, 2)$ as $C_{\text{pr-psi}} \leq \frac{3}{5}$.

2) Converse for Arbitrary $(K, M, N, T)$: If $M = K - 1$, it is trivial that 1 is an upper bound, since any rates cannot be larger than 1. So let us assume that $M < K - 1$. For compact notation, let us define

$$D(K, S, \theta, V) \triangleq H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_{[V]}, \Theta = \theta, S = S\right).$$

Here $W_{[V]} = (W_1, W_2, \cdots, W_V)$ represents the messages that appear in the conditioning. Also, define an arbitrary $T \subset [N]$ with cardinality $|T| = T$ which represents the set of indices of colluding databases.

Without loss of generality, suppose messages $W_1, \cdots, W_M$ are known as side information and $W_{M+1}$ is desired. Then, we have

$$D(K, M, M + 1, M) = H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_{[M]}, \Theta = M + 1, S = [M]\right) \quad (56)$$

$$= H\left(W_\theta \mid Q^{|\Theta|, S}, W_{[M]}, \Theta = M + 1, S = [M]\right) \quad (57)$$

where equation (55) holds because of Lemma 1. Equation (56) holds because of Lemma 2. There are a total of $\binom{N}{M}$ such subsets $T$. Writing (57) for all such subsets, adding those inequalities and divided by $\binom{N}{M}$, we obtain

$$D(K, M, M + 1, M) \geq \frac{T}{N} H\left(A^{|\Theta|, S} \mid Q^{|\Theta|, S}, W_{[M+1]}, \Theta = M + 2, S = [M]\right).$$
where (58) follows from Han’s inequality. Proceeding along these lines, we have

\[
D(K, [M], M + 1, M) 
\geq L + \frac{T}{N} D(K, [M], M + 2, M + 1) 
\geq L + \frac{T}{N} \left( L + \frac{T}{N} D(K, [M], M + 3, M + 2) \right) 
\geq \cdots 
\geq L + \frac{T}{N} \left( L + \cdots + \frac{T}{N} \left( L + \frac{T}{N} D(K, [M], K, K - 1) \right) \right) 
\]

where \( D(K, [M], K, K - 1) \geq L \). Therefore,

\[
D(K, [M], M + 1, M) 
\geq L + \frac{T}{N} L + \cdots + \left( \frac{T}{N} \right)^{K-M-1} L 
= L \left( 1 + \frac{T}{N} + \cdots + \left( \frac{T}{N} \right)^{K-M-1} \right). 
\]

The above argument holds similarly for any \((\theta, S)\), and hence the upper bound on the rate of TPIR-PSI is

\[
R = \lim_{L \to \infty} \frac{L}{D} 
\leq \left( 1 + \frac{T}{N} + \frac{T}{N}^2 + \cdots + \left( \frac{T}{N} \right)^{K-M-1} \right)^{-1}. 
\]

Thus, the proof of converse for Theorem 1 is complete.

**Remark 9.** The converse can also be proved alternatively by a genie-aided approach using the capacity of TPIR-GSI of Remark 2 as follows. Starting from the TPIR-PSI problem, suppose we provide the indices of the side information \( S \) to all the databases, so the side information is now globally known and only the privacy of the desired message needs to be preserved. Any schemes for TPIR-PSI are applicable to this TPIR-GSI setting, because they preserve the privacy of the desired message index even after the side-information is revealed. This is because TPIR-PSI schemes satisfy \( I(\Theta_S, Q^{[\Theta]_S}, A^{[\Theta]_S}, W^{[K]}) = 0 \), which in turn implies that \( I(\Theta_S, Q^{[\Theta]_S}, A^{[\Theta]_S}, W^{[K]} | S) = 0 \). Therefore,

\[
C_{\text{TPIR-PSI}} \leq C_{\text{TPIR-GSI}} 
= \left( 1 + \frac{T}{N} + \frac{T}{N}^2 + \cdots + \left( \frac{T}{N} \right)^{K-M-1} \right)^{-1}. 
\]

**V. PROOF OF THEOREM 2**

**A. Achievability**

When \( M = K - 1 \), the user can download the sum of all the messages from one database and get the desired message with side information. The rate is 1, achieving the capacity. Note that in this case, common randomness among databases is not required. When \( M < K - 1 \), the achievable scheme can directly use the scheme of STPIR [14], [15], and the side information is simply not used.

**B. Converse**

When \( M = K - 1 \), it is obvious that 1 is an upper bound. When \( M < K - 1 \), we show that \( 1 - \frac{T}{N} \) is an upper bound.

**a) Proof of the bound \( R \leq 1 - \frac{T}{N} \)**: Let us start with an intuitive understanding of the upper bound, \( R \leq 1 - \frac{T}{N} \). Due to database privacy, given the side information, the answers from all \( N \) databases should be independent of the non-queried messages. At the same time, the answers from any \( T \) databases should contain no information about the queried message index since the user privacy must be preserved. Combining these two facts, given the side information, the answers from any \( T \) databases should contain no information about any individual message, whether desired or undesired. As a result, the useful information about the desired message must come from the remaining \( N - T \) databases. Thus, the download per database must be at least \( 1/(N - T) \) times the entropy of the desired message.

The formal proof is as follows. Since \( M < K - 1 \), for any \( S \in S \), there exist at least 2 messages that are not in the set \( S \). Any feasible STPIR-PSI scheme must satisfy the database-privacy constraint (12),

\[
0 = I(W^{[\Theta],[\Theta]_S}, Q^{[\Theta]_S} | W_S, S, \Theta) 
\]

Therefore, \( \forall T \subset [N], |T| = T, \forall S \in S, \) and for all distinct \( \theta, \theta' \in [K] \setminus S, \)

\[
0 = I(W^{[\Theta], [\Theta]_S}, Q^{[\Theta]_S} | W_S, \Theta = \theta, S = S) 
= I(W^{[\Theta], [\Theta]_S} | W_S, \Theta = \theta, S = S) 
+ I(W^{[\Theta], [\Theta]_S} | Q^{[\Theta]_S}, W_S, \Theta = \theta, S = S) 
= H(A^{[\Theta]_S} | Q^{[\Theta]_S}, W_S, \Theta = \theta, S = S) 
- H(A^{[\Theta]_S} | Q^{[\Theta]_S}, W_S, W_{T'}, \Theta = \theta, S = S) 
\leq H(A^{[\Theta]_S} | Q^{[\Theta]_S}, W_S, \Theta = \theta', S = S) 
- H(A^{[\Theta]_S} | Q^{[\Theta]_S}, W_S, W_{T'}, \Theta = \theta', S = S) \]

where (67) holds because \( T \) is a subset of \([N]\) and (69) holds due to (4). According to the correctness condition,

\[
L = H(W_{T'}) 
\]

\[
\leq I(W^{[\Theta], [\Theta]_S} | W_S, Q^{[\Theta]_S}, \Theta = \theta', S = S) 
\leq H(A^{[\Theta]_S} | W_S, Q^{[\Theta]_S}, \Theta = \theta', S = S) 
- H(A^{[\Theta]_S} | W_{T'}, W_S, Q^{[\Theta]_S}, \Theta = \theta', S = S) \]

\[
= H(A^{[\Theta]_S} | W_S, Q^{[\Theta]_S}, \Theta = \theta', S = S) 
= H(A^{[\Theta]_S} | W_S, Q^{[\Theta]_S}, \Theta = \theta', S = S) 
\]
where (74) follows due to Lemma 1. Writing (77) for all $U$ randomness

$$T \subset \Theta \in [K] \setminus S$, it is also true when averaged across them, so,

$$L \leq \frac{1}{N} \sum_{T} \frac{1}{N} \frac{H(A_{T}^{[S]} | W_{T}, Q_{T}^{[S]}, \Theta = \theta', S = S)}{H(A_{T}^{[S]} | W_{T}, Q_{T}^{[S]}, \Theta = \theta, S = S)}$$

where (79) is due to Han’s inequality. Since this inequality is true for all $S \in S, \theta' \in [K] \setminus S, \theta'$ holds because dropping conditioning does not reduce entropy. Therefore, $R = \lim_{L \to \infty} \frac{L}{T} \leq 1 - \frac{T}{N}$, and we have shown that the rate of any feasible STPIR-PSI scheme cannot be more than $1 - \frac{T}{N}$.

b) Proof of the bound $\rho \geq T/(N - T)$: Let us first explain the intuition behind this bound on the size of the common randomness $U$ that should be available to all databases but not to the user. We have already shown that the normalized size of the answer from any individual database must be at least $L/(N - T)$. Due to the user and database privacy constraints, the answers from any $T$ databases are independent of the messages. Therefore, to ensure database privacy, the amount of common randomness must be no smaller than the size of the answers from $T$ databases.

The formal proof is as follows. Suppose a feasible STPIR-PSI scheme exists that achieves a non-zero rate. Then we show that it must satisfy $\rho \geq T/(N - T)$. For $S = S \in S$ and $\Theta = \theta \in [K] \setminus S$, consider the answering strings $A_{1}^{[S]}, \ldots, A_{N}^{[S]}$ and the side information $W_{S}$, from which the user can retrieve $W_{\theta}$. According to the database-privacy constraint, we have

$$0 = I(W_{(\theta,S)} : A_{[N]}^{[S]} | W_{S}, Q_{T}^{[S]}, \Theta = \theta, S = S)$$
VII. ACKNOWLEDGMENT
This work was supported in part by NSF grants CNS-1731384, CCF-1907053 and by ONR grant N00014-18-1-2057.

APPENDIX A
SOME INSIGHTS ON THE CAPACITY OF PIR-SPSI
The four variants of PIR with side information are defined as follows.

- **PIR-SI**, or PIR with (non-private) side information. Only the privacy of the desired message is preserved, i.e.,
  \[ I\left(\Theta; Q_n^{[0,s]}\left|W_n\right.\right) = 0, \forall n \in [N]. \]

- **PIR-SPSI**, or PIR with separately private side information. The privacy of the desired message and the privacy of the side information are preserved individually, i.e.,
  \[ I\left(\Theta; Q_n^{[0,s]}\left|W_n\right.\right) = I\left(S; Q_n^{[0,s]}\left|W_n\right.\right) = 0, \forall n \in [N]. \]

- **PIR-GSI**, or PIR with jointly private side information. The privacy of the desired message and the privacy of the side information are preserved jointly, i.e.,
  \[ I\left(\Theta, S; Q_n^{[0,s]}\left|W_n\right.\right) = 0, \forall n \in [N]. \]

- **PIR-PSI**, or PIR with jointly private side information. The privacy of the desired message index and the privacy of the side information index are preserved individually, i.e.,
  \[ I\left(\Theta, S; Q_n^{[0,s]}\left|W_n\right.\right) = I\left(s; Q_n^{[0,s]}\left|W_n\right.\right) = 0, \forall n \in [N]. \]

From the result of Theorem 1 we know the capacity of PIR-SI is \(\Psi(1/N, K - M)\), and from Remark 2 that follows Theorem 1 we also know the capacity of PIR-GSI is \(\Psi(1/N, K - M)\). The capacity of PIR-SI is known to be \(\left[\frac{K}{2}\right]^{-1}\) for \(N = 1\) database from [20]. In spite of various attempts the capacity of PIR-SI remains in general an open problem for multiple databases. The remaining setting of PIR-SPSI has not been studied, perhaps due to lack of practical motivation for this setting. Nevertheless, out of technical curiosity, let us present some insights into the capacity of PIR-SPSI. We will focus only on the single database setting, i.e., \(N = 1\) in this section.

A. **PIR-SPSI: \(N = 1, M = 1, K\) even**

For this setting the capacity of PIR-SPSI is \(\left[\frac{K}{2}\right]^{-1}\), i.e., the same as the capacity of PIR-SI. Since PIR-SPSI is a more constrained version of PIR-SI, its capacity cannot be higher than that of PIR-SI. Thus, the converse is trivial. It turns out that the achievability is also straightforward because the Partition and Code scheme in [20] already preserves the separate privacy of side information. Let us present just an example to illustrate this. Suppose \(N = 1, M = 1, K = 6\), and suppose each message is comprised of one bit. Let \(s\) denote the desired message index and \(s\) denote the index of the message available as side information to the user. The user asks the database for three bits, corresponding to the three partitions: \(P_1 = W_1 + W_{i_2}, P_2 = W_3 + W_{i_1}, P_3 = W_{i_5} + W_{i_6}\). The indices \((i_1, i_2, \ldots, i_6)\) are obtained by first randomly permuting \((1, 2, \ldots, 6)\) and then switching the position of the side information index \(s\) with another index (if needed) so that it appears within the same partition as \(\theta\), i.e., one of the partitions must contain \(W_\theta + W_{s'}\). The scheme is correct because the user can recover \(W_\theta\) from the sum \(W_\theta + W_{s'}\) (because \(W_{s'}\) is already available to the user as side information). It is easily verified that \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_6)\), so the scheme preserves their separate privacy. However, since \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_6)\), so the scheme preserves their separate privacy. However, since \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_6)\), so the scheme preserves their separate privacy. Hence, since \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_6)\), so the scheme preserves their separate privacy. However, since \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_6)\), so the scheme preserves their separate privacy. However, since \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_6)\), so the scheme preserves their separate privacy.

B. **PIR-PSI: \(N = 1, M = 1, K\) odd**

For this setting also the capacity of PIR-PSI is \(\left[\frac{K+1}{2}\right]^{-1}\), the same as the capacity of PIR-SI. Once again, the converse is trivially inferred from PIR-SI. Achievability requires a small modification to the Partition and Code scheme of [20], as explained next. Let us also illustrate this through an example. Suppose \(N = 1, M = 1, K = 7\) and each message is comprised of one symbol from, say \(\mathbb{F}_2\). The user asks the database for 4 symbols, corresponding to \(P_1 = W_{i_1} + W_{i_2}, P_2 = W_{i_3} + W_{i_4}, P_3 = W_{i_5} + W_{i_6} + W_{i_7}\), and \(P_4 = W_{i_8} + 2W_{i_9} + 3W_{i_7}\). In fact, \(P_3, P_4\) can be the non-systematic symbols of any \((5, 3)\) systematic MDS code applied to \(W_{i_8}, W_{i_9}, W_{i_7}\). Once again, the indices \((i_1, i_2, \ldots, i_7)\) are obtained by first randomly permuting \((1, 2, 3, 4, 5, 6, 7)\) and then switching the position of the side information index \(s\) with another index (if needed) so that it appears within the same partition as \(\theta\). If \(W_\theta\) and \(W_s\) appear in \(P_1\) or \(P_2\) then \(W_\theta\) is decoded by subtracting the side-information, while if \(W_\theta\) and \(W_s\) appear in partitions \(P_3, P_4\) with interfering message \(W_{i_7}\), then after eliminating the known side information \(W_{i_7}\), the two equations can be solved for the remaining two variables \(W_\theta, W_s\) (equivalently, the MDS property guarantees decodability). Once again, it is easily verified that \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_7)\), so the scheme preserves their separate privacy. However, since \(\theta\) and \(s\) must appear in the same partition, it is also clear that their joint privacy is not preserved. The example generalizes to any odd value of \(K\), by constructing \((K+1)/2\) partitions of the form \(W_{i_1} + W_{i_2}, W_{i_3} + W_{i_4}, \ldots, W_{i_K} + W_{i_{K+1}}\), \(W_{i_k} - W_{i_{k+1}} + W_{i_k}\) and \(W_{i_k} - W_{i_{k+1}} + 2W_{i_{k+1}} + 3W_{i_k}\), and generating the indices \((i_1, i_2, \ldots, i_K)\) by first randomly permuting \((1, 2, \ldots, K)\) and then switching the position of the side information index \(s\) with another index (if needed) so that it appears within the same partition as \(\theta\). This ensures that \(\theta\) and \(s\) are each uniformly distributed over \((i_1, i_2, \ldots, i_K)\), so the scheme preserves their separate privacy. However, since \(\theta\) and \(s\) must appear in the same partition, it is also clear that their joint privacy is not preserved.

C. **PIR-PSI: \(N = 1, M = 2, K = 6\)**

The preceding discussion shows that PIR-SI and PIR-SPSI have the same capacity for \(N = 1, M = 1\). Let us now present
an example to show that the capacity of PIR-PSPI can be strictly less than the capacity of PIR-SI in general. For this example, let us consider $K = 6$ messages stored at $N = 1$ database, out of which $M = 2$ messages are available to the user as side information. From [20] we know that the capacity of PIR-SI for this example is $1/2$. Incidentally, this is achieved by downloading two partitions, namely $W_{i_1} + W_{i_2} + W_{i_3}$ and $W_{i_4} + W_{i_5} + W_{i_6}$, where the indices $(i_1, i_2, \cdots, i_6)$ are generated by first randomly permuting $(1, 2, \cdots, 6)$ and then switching indices if necessary to place the two side information indices into the same partition as $\theta$. Note that this scheme does not preserve the privacy of side information indices, e.g., $(i_1, i_4)$ cannot be both side information indices (because side information indices must be within the same partition). We will show that for this example the capacity of PIR-PSPI is no more than $1/3$, i.e., strictly smaller than the capacity of PIR-SI.

Let us denote the entropy of each message as $L$. We need some preliminary work before we start the core of the converse proof. To have compact notation, for any subset $P \subset [K]$, let us define

$$H(A | Q = q) = H(\{W_i | \theta = q, W_{[K] \setminus i}\}).$$

(90)

Intuitively, $H(A | Q = q)$ represents the entropy that remains in the answer $A$ due to messages $W_P$ (after all other messages are known), i.e., the ‘space’ occupied by the messages $W_P$ in $A$. We need the following facts.

**Lemma 3.** The following facts hold for PIR-PSPI with $N = 1, M = 2, K = 6$.

1) If $P$ is a singleton set, e.g., $P = \{k\}$, then we must have $H(A | Q = q) < 3L$. (89)

2) If $P_1 \subset P_2 \subset [K]$, then

$$H(A | Q = q) \leq 3L, \forall k \in [K].$$

(91)

3) If $P = \theta$ is the desired message index, $S = (s_1, s_2)$ are the $M = 2$ side information indices, and $l, m, n$ are the $3$ remaining indices representing interfering messages, then we must have,

$$H(A | Q = q) < 2L, \forall i \in \{l, m, n\}.$$

(93)

**Proof.** The first fact, (91) holds because given the answer $A$ and all messages except $W_k$ (which must include the side information), the user must be able to decode $W_k$, therefore $L = I(W_k; A | Q = q, W_{[K] \setminus i}) \leq H(A | W_k)$. The next fact, (92) is simply the statement that conditioning reduces entropy. The third fact, (93) is quite intuitive, as it says that the space occupied by interference must be less than $2L$ bits because the overall download is less than $3L$ bits, out of which $L$ bits are needed for the desired message. Formally, this can be seen as follows.

$$L = I(W_\theta; A | Q = q, W_{s_1}, W_{s_2})$$

$$= H(A | Q = q, W_{s_1}, W_{s_2})$$

$$- H(A | Q = q, W_{s_1}, W_{s_2}, W_\theta)$$

(95)

$$\leq H(A | Q = q) - H(A | W_i, W_m, W_n)$$

(96)

$$< 2L - H(A | W_i, W_m, W_n)$$

(97)

which implies (93). Finally, the last fact, (94) is also quite intuitive. It says that the desired information must not align with interference so that the user is able to resolve the two. Formally, for any $i \in \{l, m, n\}$, because the user must be able to decode his desired message from $A$ and his side information,

$$L = I(W_\theta; A | Q = q, W_{[K] \setminus \{\theta, i\}})$$

$$= H(A | W_\theta, W_i) - H(A | W_\theta, W_i)$$

(99)

$$\leq H(A | W_\theta, W_i) - L$$

(100)

which implies (94). Note that we used (91) to obtain (101).

With these preliminary facts established, let us now proceed with the core of the converse argument. Since the query preserves the privacy of the side information, all choices of $(s_1, s_2)$ must be equally likely. In particular they must all be feasible (have non-zero probability) from the database’s perspective. Note that because the database knows $Q = q$, it can evaluate $H(W_P)$ for all $P \subset [K]$. Let $(a, b, c, d, e, f)$ represent some permutation of $(1, 2, \cdots, 6)$. The main reasoning now proceeds through the following steps.

1) Consider $(s_1, s_2) = (a, b)$. Since this must be feasible, there must exist another index in $[K]$ that could be a desired message, i.e., that satisfies facts (93), (94). Without loss of generality, let $c$ be this index, so that,

$$H(A | W_d, W_e, W_f) < 2L,$$

(102)

$$H(A | W_c, W_d) \geq 2L, \forall i \in \{d, e, f\}.$$  

(103)

2) Now consider $(s_1, s_2) = (b, c)$. This must also be feasible, so there must exist an index in $[K]$ which can be a desired message. Based on (102), and the fact (94) the database can conclude that this index must be $a$. This is because all other indices lead to contradictions. For example, if the desired message is $W_d$, then from (94) we must have $H(A | W_d, W_c) \geq 2L$, which contradicts the fact that $H(A | W_d, W_c) \leq H(A | W_d, W_c, W_f) < 2L$ according to (92) and (102). Similarly, the desired message index cannot be $e$ or $f$ either, leaving $a$ as the only possibility. Now (94) implies that we must have

$$H(A | W_d, W_e) \geq 2L, \forall i \in \{d, e, f\}.$$  

(104)

3) Next, consider $(s_1, s_2) = (e, f)$. This must also be feasible, so there must exist an index in $[K]$ which can
be a desired message. Based on (103), (104) and the fact (93) the database can conclude that this index must be \( d \). This is because all other indices lead to contradictions. For example, if the desired message is \( a \), then from (93) we must have \( H(A(W_b, W_c, W_d) < 2L) \). Along with (92) this implies that \( H(A(W_c, W_d) < 2L \) which contradicts (103). Similarly, the desired message index cannot be \( b \) or \( c \) either, leaving \( d \) as the only possibility. Now (93) implies that we must have

\[
H(A(W_a, W_b, W_c)) < 2L. \tag{105}
\]

4) Finally, consider \( (s_1, s_2) = (a, d) \). This must also be feasible, so there must exist an index in \( [K] \) which can be a desired message. However, it turns out that every choice of this desired message index leads to a contradiction. For example, suppose the desired message index is \( b \). Then according to (94) we must have \( H(A(W_b, W_c) \geq 2L) \), which contradicts with the combination of (105) and (92). All other indices are similarly ruled out, leaving us with an unavoidable contradiction.

The contradiction proves that the download must be at least 3L bits, which in turn implies that the average download must be at least 3L bits, and therefore the capacity cannot be more than \( 1/3 \). The exact capacity even for this simple setting remains an intriguing open problem. Remarkably, if the capacity is less than \( 1/3 \) then that would imply that having more side-information is counterproductive for PIR-SPSI (because if \( M \) is reduced from 2 to 1 then we do know from the preceding discussion in this section that the capacity of PIR-SPSI is \( 1/3 \)).

REFERENCES