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 $\left(1+\frac{T}{N}+\cdots+\left(\frac{T}{N}\right)^{K-M-1}\right)^{-1}$. As a special case obtained by setting $T=1$, this result settles the capacity of PIR-PSI, an open problem previously noted by Kadhe et al. We also consider the problem of symmetric-TPIR with private side information (STPIR-PSI), where the answers from all N databases reveal no information about any other message besides the desired message. We show that the capacity of STPIR-PSI is $1-\frac{T}{N}$ if the databases have access to common randomness (not available to the user) that is independent of the messages, in an amount that is at least $\frac{T}{N-T}$ bits per desired message bit. Otherwise, the capacity of STPIR-PSI is zero.**

I. Introduction

The private information retrieval (PIR) problem investigates the privacy of the contents downloaded from public databases. In the classical form of PIR [1], a user wishes to, as efficiently as possible, retrieve one of K messages that are replicated across N non-colluding databases while preserving the privacy of the desired message index. Since its first formulation by Chor et al. in [1], the PIR problem has been studied extensively in computer science and cryptography under both information-theoretic and computational privacy constraints [2]–[6]. While studies of PIR typically seek to optimize both the upload and download costs, recently there has been a burst of activity aimed at *capacity* characterizations for information-theoretic PIR under the assumption of large message sizes, so that the communication cost is dominated by the download cost [7]–[12]. The capacity of PIR was defined in [9] as the maximum number of bits of the desired message that can be privately obtained per bit of total downloaded information from all the servers. In order to summarize some of the capacity results for PIR, let us define the function $\Psi(A,B)=\left(1+A+A^2+\cdots+A^{B-1}\right)^{-1}$ for positive real number A and positive integer B. Correspondingly, $\Psi(A,\infty)=1-A$ for $A<1$. The capacity of PIR was characterized in [9] as $C_{\text{PIR}}=\Psi(1/N,K)$. The capacity of T-PIR, where the privacy of the user's desired message index must be protected against collusion among any set of up to T servers, was characterized in [13] as $C_{\text{TPIR}}=\Psi(T/N,K)$. 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-SPSI, 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-SPSI.

Out of these four settings, PIR-GSI is rather trivial, and PIR-SPSI 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

The authors are with Center for Pervasive Communications and Computing (CPCC), University of California Irvine, Irvine, CA 92697, email: {zhenc4, zhiying, syed}@uci.edu.

Kadhe et al. in [20].

For PIR-SI with a single database ($N = 1$), Kadhe et al. showed in [20] that the capacity is $\lceil \frac{K}{M+1} \rceil^{-1}$. 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 $\Psi(1/N, \lceil K/(M+1) \rceil)$. 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 an 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-SI ≥ PIR-SPSI ≥ 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-PSI 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 $C_{\text{TPIR-PSI}} = \Psi(T/N, \infty)$, 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 $\{z_1, z_1+1, \cdots, z_2\}$ and $(z_1 : z_2)$ represents the vector $(z_1, z_1+1, \cdots, z_2)$. The compact notation $[z]$ represents $[1:z]$ for positive integer z. For random variables $\boldsymbol{W}_i, i = 1, 2, \ldots,$ and a set of positive integers $S = \{s_1, s_2, \cdots, s_n\}$, where $s_1 < s_2 < \cdots < s_n$, the notation $\boldsymbol{W}_S$ represents the vector $(\boldsymbol{W}_{s_1}, \boldsymbol{W}_{s_2}, \cdots, \boldsymbol{W}_{s_n})$. 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'. $\mathbb{F}_q$ 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 $\boldsymbol{W}_{[K]} = (\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K)$, each containing L independent and uniform bits, i.e., their entropy satisfies

$$H(\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K) = H(\boldsymbol{W}_1) + \cdots + H(\boldsymbol{W}_K), \tag{1}$$
$$H(\boldsymbol{W}_1) = \cdots = H(\boldsymbol{W}_K) = L. \tag{2}$$

There are N databases and each database stores all K messages $\boldsymbol{W}_1, \cdots, \boldsymbol{W}_K$. A user is equipped with a local cache and has M ($M < K$) messages as side information. Let $\boldsymbol{S} = \{\boldsymbol{i}_1, \boldsymbol{i}_2, \cdots, \boldsymbol{i}_M\}$ be M distinct indices chosen uniformly from $[K]$. These M cached messages are represented as $\boldsymbol{W}_{\boldsymbol{S}} = (\boldsymbol{W}_{\boldsymbol{i}_1}, \cdots, \boldsymbol{W}_{\boldsymbol{i}_M})$. $\boldsymbol{S}$ is not known to the databases. A user wishes to retrieve $\boldsymbol{W}_{\boldsymbol{\Theta}}$, where $\boldsymbol{\Theta}$ is a message index uniformly chosen from $[K] \setminus \boldsymbol{S}$, as efficiently as possible, while revealing no information about $(\boldsymbol{\Theta}, \boldsymbol{S})$ to any colluding subsets of up to T out of the N databases. Note the following independence,

$$H(\boldsymbol{\Theta}, \boldsymbol{S}, \boldsymbol{W}_1, \cdots, \boldsymbol{W}_K) = H(\boldsymbol{\Theta}, \boldsymbol{S}) + \sum_{i=1}^{K} H(\boldsymbol{W}_i). \tag{3}$$

In order to retrieve $\boldsymbol{W}_{\boldsymbol{\Theta}}$, the user generates N queries $\boldsymbol{Q}_1^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \cdots, \boldsymbol{Q}_N^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ with the knowledge of $(\boldsymbol{\Theta}, \boldsymbol{S}, \boldsymbol{W}_{\boldsymbol{S}})$. Since the queries are generated with no knowledge of the other $K-M$ messages, the queries must be independent of them,

$$I\left(\boldsymbol{\Theta}, \boldsymbol{S}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_1^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \cdots, \boldsymbol{Q}_N^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{W}_{[K]\setminus\boldsymbol{S}}\right) = 0. \tag{4}$$

The user sends query $\boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ to the n^{th} database and in response, the n^{th} database returns an answer $\boldsymbol{A}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ which is a deterministic function of $\boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ and $\boldsymbol{W}_{[K]}$,

$$H\left(\boldsymbol{A}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_1, \cdots, \boldsymbol{W}_K\right) = 0. \tag{5}$$

Upon collecting the answers from all N databases, the user must be able to decode the desired message $\boldsymbol{W}_{\boldsymbol{\Theta}}$ based on the queries and side information,

$$\text{[Correctness]}\ H\left(\boldsymbol{W}_{\boldsymbol{\Theta}} \mid \boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{S}, \boldsymbol{\Theta}\right) = 0. \tag{6}$$

To satisfy the user-privacy constraint that any T colluding databases learn nothing about $(\boldsymbol{\Theta}, \boldsymbol{S})$, the information available to any T databases (queries, answers and stored messages)

[1] The converse in [27] does not cover the scope of PIR-SI, because the privacy condition assumed in [27] is not a necessary condition for PIR-SI schemes.

[2] Based on progressively tighter privacy constraints, it is already immediately obvious that in terms of their capacities, the settings can be partially ordered as PIR-SI ≥ PIR-SPSI ≥ PIR-PSI, and PIR-SI≥ PIR-GSI. The main result of this work shows that PIR-PSI has the same capacity as PIR-GSI, thus allowing a complete ordering.

must be independent of $(\boldsymbol{\Theta}, \boldsymbol{S})$. [3] Let $\mathcal{T}$ be any subset of $[1:N]$, of cardinality $|\mathcal{T}| = T$. $\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ represents the vector of queries corresponding to $\boldsymbol{Q}_{n}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, n \in \mathcal{T}$. $\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ is defined as the answer vector corresponding to $\boldsymbol{A}_{n}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, n \in \mathcal{T}$. To satisfy the T-privacy requirement we must have $\forall \mathcal{T} \subset [1:N], |\mathcal{T}| = T$,

$$\text{[User privacy]}\ I\left(\boldsymbol{\Theta}, \boldsymbol{S}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}\right) = 0. \tag{7}$$

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.

$$R_{\text{TPIR-PSI}} \triangleq \lim_{L\to\infty} \frac{L}{D}, \tag{8}$$

where D is the expected (over all $\boldsymbol{\Theta}$, $\boldsymbol{S}$, $\boldsymbol{W}_{[K]}$ 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 $\boldsymbol{W}_{[K]}$ beyond the retrieved message, $\boldsymbol{W}_{\boldsymbol{\Theta}}$, and the side information, $\boldsymbol{W}_{\boldsymbol{S}}$. To facilitate database privacy, suppose the databases share a common random variable $\boldsymbol{U}$ 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

$$\begin{aligned} &H\left(\boldsymbol{\Theta}, \boldsymbol{S}, \boldsymbol{W}_1, \cdots, \boldsymbol{W}_K, \boldsymbol{U}\right) \\ &\quad = H\left(\boldsymbol{\Theta}, \boldsymbol{S}\right) + \sum_{i=1}^{K} H\left(\boldsymbol{W}_i\right) + H(\boldsymbol{U}). \end{aligned} \tag{9}$$

The answering string $\boldsymbol{A}_{n}^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ is a deterministic function of $\boldsymbol{Q}_{n}^{[\boldsymbol{\Theta},\boldsymbol{S}]}$, $\boldsymbol{W}_{[K]}$ and common randomness $\boldsymbol{U}$,

$$H\left(\boldsymbol{A}_{n}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{n}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_1, \cdots, \boldsymbol{W}_K, \boldsymbol{U}\right) = 0. \tag{10}$$

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

$$\text{[User privacy]}\ I\left(\boldsymbol{\Theta}, \boldsymbol{S}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}, \boldsymbol{U}\right) = 0. \tag{11}$$

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

$$\text{[DB privacy]}\ I\left(\boldsymbol{W}_{\overline{(\boldsymbol{\Theta},\boldsymbol{S})}}; \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}, \boldsymbol{S}, \boldsymbol{W}_{\boldsymbol{S}}\right) = 0. \tag{12}$$

An STPIR-PSI scheme is called *feasible* if it satisifes 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*

$$\begin{aligned} C_{\textit{TPIR-PSI}} &= \left(1 + \frac{T}{N} + \left(\frac{T}{N}\right)^2 + \cdots + \left(\frac{T}{N}\right)^{K-M-1}\right)^{-1} \\ &= \Psi(T/N, K-M), \end{aligned} \tag{13}$$

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

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 Π 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 Π 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

[3] Note that the joint privacy of $(\boldsymbol{\Theta}, \boldsymbol{S})$ is a stronger constraint than the marginal privacy of each of $\boldsymbol{\Theta}$ and $\boldsymbol{S}$, i.e., $I(\boldsymbol{\Theta}, \boldsymbol{S}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}) = 0$ implies both $I(\boldsymbol{\Theta}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}) = 0$ and $I(\boldsymbol{S}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}) = 0$. However, the reverse is not true, i.e., even if both $I(\boldsymbol{\Theta}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}) = 0$ and $I(\boldsymbol{S}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}) = 0$, this does not imply that $I(\boldsymbol{\Theta}, \boldsymbol{S}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}) = 0$.

setting of [19] should allow a higher capacity in general. For example, if $T = 1$ and the amount of side information is ML bits, then the capacity result of [19] corresponds to the expression $\Psi(1/N, K)/\left(1 - \frac{M}{K}\right)$. It is easy to verify that $C_{\text{TPIR-GSI}} = \Psi(1/N, K - M) < \Psi(1/N, K)/\left(1 - \frac{M}{K}\right)$ 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 $C_{\text{TPIR-GSI}}$ reflects the same insight — the role of globally known side information is to reduce the effective number of messages by M. The authors of [21] also give a similar explanation for the scheme in [19].

Remark 4. Now we can completely order the four variants of PIR with side information, in terms of their capacities as PIR-SI $\geq$ PIR-SPSI $\geq$ 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-SI is $1/2$ while the capacity of PIR-SPSI is no more than $1/3$, so that PIR-SI $>$ PIR-SPSI. 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-SPSI is $1/3$ while the capacity of PIR-PSI is only $1/5$, so that PIR-SPSI $>$ 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*

$$C_{\textit{STPIR-PSI}} = \begin{cases} 1, & \textit{if } M = K - 1, \\ 1 - \frac{T}{N}, & \textit{if } M < K - 1 \textit{ and } \rho \geq \frac{T}{N-T}, \\ 0, & \textit{otherwise}, \end{cases} \tag{14}$$

where $\rho = \frac{H(\boldsymbol{U})}{L}$ 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 TPIR 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 TPIR-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., $C_{\text{TPIR-PSI}} > C_{\text{STPIR-PSI}}$ for any finite K and $C_{\text{TPIR-PSI}} \to C_{\text{STPIR-PSI}}$ when $K \to \infty$. 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.[4] 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 TPIR-PSI with parameters (K, M, N, T) is the achievable scheme of TPIR [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 TPIR [13], the messages $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3 \in \mathbb{F}_q^{27}$ are 27×1 column vectors and let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3 \in \mathbb{F}_q^{27 \times 27}$ represent random matrices chosen privately by the user, independently and uniformly from all 27×27 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 27×1 column vectors $\boldsymbol{a}_{(1:27)}, \boldsymbol{b}_{(1:27)}, \boldsymbol{c}_{(1:27)} \in \mathbb{F}_q^{27}$ as follows.

$$\boldsymbol{a}_{(1:27)} = \boldsymbol{Y}_1 \boldsymbol{W}_1, \tag{15}$$

$$\boldsymbol{b}_{(1:18)} = G_{18 \times 12} \boldsymbol{Y}_2[(1:12), :] \boldsymbol{W}_2, \tag{16}$$

$$\boldsymbol{c}_{(1:18)} = G_{18 \times 12} \boldsymbol{Y}_3[(1:12), :] \boldsymbol{W}_3, \tag{17}$$

$$\boldsymbol{b}_{(19:27)} = G_{9 \times 6} \boldsymbol{Y}_2[(13:18), :] \boldsymbol{W}_2, \tag{18}$$

$$\boldsymbol{c}_{(19:27)} = G_{9 \times 6} \boldsymbol{Y}_3[(13:18), :] \boldsymbol{W}_3, \tag{19}$$

where $\boldsymbol{Y}_2[(1:18), :]$ and $\boldsymbol{Y}_3[(1:18), :]$ are 18×27 matrices comprised of the first 18 rows of $\boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$, respectively.

[4] The explanation is similar to that for TPIR with globally known side information as in Remark 2.

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 $R_{\text{TPIR}} = 9/19$, namely, in this scheme the user recovers 9 desired symbols from a total of 19 downloads symbols from each database.

TABLE I
ACHIEVABLE SCHEME OF TPIR [13]

DB_1	DB_2	DB_3
$\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4$	$\boldsymbol{a}_5, \boldsymbol{a}_6, \boldsymbol{a}_7, \boldsymbol{a}_8$	$\boldsymbol{a}_9, \boldsymbol{a}_{10}, \boldsymbol{a}_{11}, \boldsymbol{a}_{12}$
$\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3, \boldsymbol{b}_4$	$\boldsymbol{b}_5, \boldsymbol{b}_6, \boldsymbol{b}_7, \boldsymbol{b}_8$	$\boldsymbol{b}_9, \boldsymbol{b}_{10}, \boldsymbol{b}_{11}, \boldsymbol{b}_{12}$
$\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \boldsymbol{c}_4$	$\boldsymbol{c}_5, \boldsymbol{c}_6, \boldsymbol{c}_7, \boldsymbol{c}_8$	$\boldsymbol{c}_9, \boldsymbol{c}_{10}, \boldsymbol{c}_{11}, \boldsymbol{c}_{12}$
$\boldsymbol{a}_{13} + \boldsymbol{b}_{13}$	$\boldsymbol{a}_{15} + \boldsymbol{b}_{15}$	$\boldsymbol{a}_{21} + \boldsymbol{b}_{17}$
$\boldsymbol{a}_{14} + \boldsymbol{b}_{14}$	$\boldsymbol{a}_{16} + \boldsymbol{b}_{16}$	$\boldsymbol{a}_{22} + \boldsymbol{b}_{18}$
$\boldsymbol{a}_{17} + \boldsymbol{c}_{13}$	$\boldsymbol{a}_{19} + \boldsymbol{c}_{15}$	$\boldsymbol{a}_{23} + \boldsymbol{c}_{17}$
$\boldsymbol{a}_{18} + \boldsymbol{c}_{14}$	$\boldsymbol{a}_{20} + \boldsymbol{c}_{16}$	$\boldsymbol{a}_{24} + \boldsymbol{c}_{18}$
$\boldsymbol{b}_{19} + \boldsymbol{c}_{19}$	$\boldsymbol{b}_{21} + \boldsymbol{c}_{21}$	$\boldsymbol{b}_{23} + \boldsymbol{c}_{23}$
$\boldsymbol{b}_{20} + \boldsymbol{c}_{20}$	$\boldsymbol{b}_{22} + \boldsymbol{c}_{22}$	$\boldsymbol{b}_{24} + \boldsymbol{c}_{24}$
$\boldsymbol{a}_{25} + \boldsymbol{b}_{25} + \boldsymbol{c}_{25}$	$\boldsymbol{a}_{26} + \boldsymbol{b}_{26} + \boldsymbol{c}_{26}$	$\boldsymbol{a}_{27} + \boldsymbol{b}_{27} + \boldsymbol{c}_{27}$

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 $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ as side information and wishes to retrieve $\boldsymbol{W}_1$. He does not need to download individual symbols of $\boldsymbol{W}_2, \boldsymbol{W}_3$, or the linear combinations comprised of only $\boldsymbol{W}_2, \boldsymbol{W}_3$ symbols, i.e., $\boldsymbol{b}_i, \boldsymbol{c}_i, 1 \leq i \leq 12$ and $\boldsymbol{b}_j + \boldsymbol{c}_j, 19 \leq j \leq 24$ 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 $G^s_{e\times f}$ 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 $\boldsymbol{X}_i \in \mathbb{F}_q^{19}$ 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 $\boldsymbol{X}_i$ with a systematic $(28, 19)$ MDS code generator matrix $G^s_{28\times 19} = [V_{19\times 9} \mid I_{19\times 19}]'$, where $I_{19\times 19}$ is the identity matrix, and downloads only the 9 linear combinations corresponding to the parity part, $V'_{19\times 9}\boldsymbol{X}_i$.

The correctness constraint is satisfied because of the property of MDS code and the correctness of the original TPIR scheme. Given $(\boldsymbol{b}_i)_{i\in[12]}$, $(\boldsymbol{c}_i)_{i\in[12]}$, $(\boldsymbol{b}_i + \boldsymbol{c}_i)_{i\in[19:24]}$, $V'_{19\times 9}\boldsymbol{X}_1$, $V'_{19\times 9}\boldsymbol{X}_2$ and $V'_{19\times 9}\boldsymbol{X}_3$, the user is able to decode $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$, 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 $(\boldsymbol{\Theta}, \boldsymbol{S})$. Thus, the rate achieved with private side information is $R_{\text{TPIR-PSI}} = 27/27 = 1$ 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 $1 + \frac{T}{N} + \cdots + (\frac{T}{N})^{K-1}$. 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[5] 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 Π. 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 Π. Out of these p_1 symbols, let p_2 $(p_2 < p_1)$ denote the number of user known symbols. Denote by vector $\boldsymbol{X}_i \in \mathbb{F}_q^{p_1}$ the symbols downloaded from DB_i in Π. For each database, use a systematic $(2p_1 - p_2, p_1)$ MDS code with generator matrix $G^s_{(2p_1-p_2)\times p_1} = \left[V_{p_1\times(p_1-p_2)} \mid I_{p_1\times p_1}\right]'$ 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, $V'_{p_1\times(p_1-p_2)}\boldsymbol{X}_i$.

Note that the user does not need to know the realization of side information $\boldsymbol{S}$ or $\boldsymbol{W_S}$ in order to construct the queries. This is because the systematic MDS code in the query redundancy removal does not depend on $\boldsymbol{S}$ or $\boldsymbol{W_S}$. During the decoding, $\boldsymbol{S}$ and $\boldsymbol{W_S}$ 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., $V'_{p_1\times(2p_1-p_2)}\boldsymbol{X}_i$, 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 $\boldsymbol{X}_i$. Then

[5] The term $(N-T)^{k-1}T^{K-k}$ comes from the undesired message exploitation step (Step 4) of achievability in [13] and can be verified recursively. A detailed analysis of a similar flavor can be found in [9].

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 Π, 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 $\boldsymbol{W}_{\boldsymbol{i}_1}, \cdots, \boldsymbol{W}_{\boldsymbol{i}_M}$, $\{\boldsymbol{i}_1, \cdots, \boldsymbol{i}_M\} \in [K]$ 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 ($k \le M$), the user knows all the instances of k-sum that contain only symbols $\boldsymbol{W}_{\boldsymbol{j}_1}, \boldsymbol{W}_{\boldsymbol{j}_2}, \cdots, \boldsymbol{W}_{\boldsymbol{j}_k}$, where $\{\boldsymbol{j}_1, \boldsymbol{j}_2, \cdots, \boldsymbol{j}_k\} \subset \{\boldsymbol{i}_1, \cdots, \boldsymbol{i}_M\}$. 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,

$$p_1 = \sum_{k=1}^{K}(N-T)^{k-1}T^{K-k}\binom{K}{k} \tag{20}$$

$$= \frac{\sum_{k=0}^{K}(N-T)^{k}T^{K-k}\binom{K}{k} - T^K}{N-T} \tag{21}$$

$$= \frac{N^K - T^K}{N-T}. \tag{22}$$

And p_2 can be simplified as,

$$p_2 = \sum_{k=1}^{M}(N-T)^{k-1}T^{K-k}\binom{M}{k} \tag{23}$$

$$= T^{K-M}\sum_{k=1}^{M}(N-T)^{k-1}T^{M-k}\binom{M}{k} \tag{24}$$

$$= \frac{T^{K-M}(N^M - T^M)}{N-T}. \tag{25}$$

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

$$m = \sum_{k=1}^{K}(N-T)^{k-1}T^{K-k}\binom{K-1}{k-1} = N^{K-1}. \tag{26}$$

Therefore, the rate achieved is

$$R_{\text{TPIR-PSI}} = \frac{Nm}{N(p_1 - p_2)} \tag{27}$$

$$= \frac{N^{K-1}(N-T)}{(N^K - T^K) - T^{K-M}(N^M - T^M)} \tag{28}$$

$$= \frac{1 - \frac{T}{N}}{1 - (\frac{T}{N})^{K-M}} \tag{29}$$

$$= \left(1 + \frac{T}{N} + \cdots + \left(\frac{T}{N}\right)^{K-M-1}\right)^{-1}. \tag{30}$$

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

B. Converse

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

Lemma 1. *For all $S_1 \in \mathcal{S}$, $\theta \in [K] \setminus S_1$, $S_2 \subseteq [K] \setminus S_1$, and $\mathcal{T} \subset [N], |\mathcal{T}| = T$, given $\boldsymbol{S} = S_1, \boldsymbol{\Theta} = \theta$, $\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \leftrightarrow \left(\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{S_1 \cup S_2}\right) \leftrightarrow \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ is a Markov chain.*

Proof. In this proof, to be convenient, denote $\mathcal{E}_1 = S_1 \cup S_2$ and $\mathcal{E}_2 = [K] \setminus (S_1 \cup S_2)$. It is equivalent to prove

$$I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) = 0.$$

By the chain rule of mutual information,

$$\begin{aligned}
&I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_2}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&= I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&\quad + I\left(\boldsymbol{W}_{\mathcal{E}_2}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&= I\left(\boldsymbol{W}_{\mathcal{E}_2}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&\quad + I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&= I\left(\boldsymbol{W}_{\mathcal{E}_2}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&+ I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&- I\left(\boldsymbol{W}_{\mathcal{E}_2}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right).
\end{aligned} \tag{31}$$

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

$$\begin{aligned}
&I\left(\boldsymbol{W}_{\mathcal{E}_2}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&= I\left(\boldsymbol{W}_{\mathcal{E}_2}; \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{S_1 \cup S_2}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \\
&\quad - I\left(\boldsymbol{W}_{[K]\setminus(S_1 \cup S_2)}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) \quad (32) \\
&= 0, \quad (33)
\end{aligned}$$

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

$$I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) = 0$$

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

$$I\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}; \boldsymbol{Q}_{[N]\setminus\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\mathcal{E}_1}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S_1\right) = 0.$$

■

Lemma 2. *For all $S \in \mathcal{S}$, $\theta, \theta' \in [K] \backslash S$, and $\mathcal{T} \subset [N], |\mathcal{T}| = T$,*

$$\begin{aligned}
&H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{\Theta}}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S\right) \\
&= H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{\Theta}}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{\Theta} = \theta', \boldsymbol{S} = S\right), \quad (34) \\
&H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S\right) \\
&= H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{\Theta} = \theta', \boldsymbol{S} = S\right). \quad (35)
\end{aligned}$$

Proof. It follows from the user-privacy constraint (11) and the non-negativity of mutual information, that for all $S \in \mathcal{S}$, $\mathcal{T} \subset [N], |\mathcal{T}| = T$

$$I\left(\boldsymbol{\Theta}; \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]} \mid \boldsymbol{S} = S\right) = 0, \tag{36}$$

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

$$\begin{aligned} &H\left(\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}} \mid \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S\right) \\ &= H\left(\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}} \mid \boldsymbol{\Theta} = \theta', \boldsymbol{S} = S\right), \quad (37) \\ &H\left(\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}} \mid \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S\right) \\ &= H\left(\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}} \mid \boldsymbol{\Theta} = \theta', \boldsymbol{S} = S\right). \quad (38) \end{aligned}$$

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,

$$\begin{aligned} D &\geq H(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\mathbf{S}]} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{\Theta}, \boldsymbol{S}) && (39) \\ &\geq \min_{\substack{S \in \mathcal{S} \\ \theta \in [K] \setminus S}} H(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\mathbf{S}]} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{\Theta} = \theta, \boldsymbol{S} = S). && (40) \end{aligned}$$

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

For $(K, M, N, T) = (3, 1, 3, 2)$, without loss of generality suppose message $\boldsymbol{W}_1$ is known as side information and $\boldsymbol{W}_2$ is desired. Let $S = \{1\}$. We bound the total download as,

$$\begin{aligned} D &\geq H\left(\boldsymbol{A}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_1, \boldsymbol{\Theta} = 2, \boldsymbol{S} = S\right) && (41) \\ &\overset{(6)}{=} H\left(\boldsymbol{A}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_2 \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_1, \boldsymbol{\Theta} = 2, \boldsymbol{S} = S\right) && (42) \\ &= H\left(\boldsymbol{W}_2 \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_1, \boldsymbol{\Theta} = 2, \boldsymbol{S} = S\right) \\ &\quad + H\left(\boldsymbol{A}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 2, \boldsymbol{S} = S\right) && (43) \\ &\geq L + H\left(\boldsymbol{A}_{[2]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 2, \boldsymbol{S} = S\right) && (44) \\ &= L + H\left(\boldsymbol{A}_{[2]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[2]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 2, \boldsymbol{S} = S\right) && (45) \\ &= L + H\left(\boldsymbol{A}_{[2]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[2]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right) && (46) \\ &\geq L + H\left(\boldsymbol{A}_{[2]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right) && (47) \end{aligned}$$

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,

$$\begin{aligned} D &\geq L + H\left(\boldsymbol{A}_{\{2,3\}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right), && (48) \\ D &\geq L + H\left(\boldsymbol{A}_{\{1,3\}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right). && (49) \end{aligned}$$

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

$$\begin{aligned} D &\geq L + \frac{1}{3} H\left(\boldsymbol{A}_{\{1,2\}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right) \\ &\quad + \frac{1}{3} H\left(\boldsymbol{A}_{\{2,3\}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right) \\ &\quad + \frac{1}{3} H\left(\boldsymbol{A}_{\{1,3\}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right) && (50) \\ &\geq L + \frac{2}{3} H\left(\boldsymbol{A}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[2]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right) && (51) \\ &= L + \frac{2}{3} L && (52) \\ &= \frac{5}{3} L. && (53) \end{aligned}$$

Here (51) follows from Han's inequality, and (52) holds because from $\left(\boldsymbol{W}_{[2]}, \boldsymbol{A}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{Q}_{[3]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta} = 3, \boldsymbol{S} = S\right)$ one can recover $\boldsymbol{W}_3$ with vanishing probability of error. Since the same argument holds for all realizations $(\boldsymbol{\Theta}, \boldsymbol{S}) = (\theta, S)$, this gives us the upper bound on the capacity of TPIR-PSI with $(K, M, N, T) = (3, 1, 3, 2)$ as $C_{\text{TPIR-PSI}} \leq \frac{3}{5}$.

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

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

Here $\boldsymbol{W}_{[V]} = (\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_V)$ represents the messages that appear in the conditioning. Also, define an arbitrary $\mathcal{T} \subset [N]$ with cardinality $|\mathcal{T}| = T$ which represents the set of indices of colluding databases.

Without loss of generality, suppose messages $\boldsymbol{W}_1, \cdots, \boldsymbol{W}_M$ are known as side information and $\boldsymbol{W}_{M+1}$ is desired. Then, we have

$$\begin{aligned} &D(K, [M], M+1, M) \\ &= H(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} | \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]) \\ &\overset{(6)}{=} H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{\boldsymbol{\Theta}} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]\right) \\ &= H\left(\boldsymbol{W}_{\boldsymbol{\Theta}} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]\right) \\ &\quad + H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M+1]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]\right). \end{aligned}$$

Note that

$$H\left(\boldsymbol{W}_{\boldsymbol{\Theta}} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]\right) = L$$

since messages are independent, and queries are independent of the messages. And

$$\begin{aligned} &H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M+1]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]\right) \\ &\geq H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M+1]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]\right) && (54) \\ &= H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M+1]}, \boldsymbol{\Theta} = M+1, \boldsymbol{S} = [M]\right) && (55) \\ &= H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M+1]}, \boldsymbol{\Theta} = M+2, \boldsymbol{S} = [M]\right) && (56) \\ &\geq H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M+1]}, \boldsymbol{\Theta} = M+2, \boldsymbol{S} = [M]\right), && (57) \end{aligned}$$

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

$$\begin{aligned} &D(K, [M], M+1, M) \\ &\geq \frac{T}{N} H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[M+1]}, \boldsymbol{\Theta} = M+2, \boldsymbol{S} = [M]\right) \end{aligned}$$

$$
\begin{aligned}
&+L && (58)\\
=&L+\frac{T}{N}D(K,[M],M+2,M+1), && (59)
\end{aligned}
$$

where (58) follows from Han's inequality. Proceeding along these lines, we have

$$
\begin{aligned}
&D(K,[M],M+1,M)\\
&\geq L+\frac{T}{N}D(K,[M],M+2,M+1) && (60)\\
&\geq L+\frac{T}{N}\left(L+\frac{T}{N}D(K,[M],M+3,M+2)\right) && (61)\\
&\geq\cdots && (62)\\
&\geq L+\frac{T}{N}\left(L+\cdots+\frac{T}{N}\left(L+\frac{T}{N}D(K,[M],K,K-1)\right)\right) && (63)
\end{aligned}
$$

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

$$
\begin{aligned}
&D(K,[M],M+1,M)\\
&\geq L+\frac{T}{N}L+\cdots+\left(\frac{T}{N}\right)^{K-M-1}L && (64)\\
&=L\left(1+\frac{T}{N}+\cdots+\left(\frac{T}{N}\right)^{K-M-1}\right). && (65)
\end{aligned}
$$

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

$$
\begin{aligned}
R&=\lim_{L\to\infty}\frac{L}{D}\\
&\leq\left(1+\frac{T}{N}+\left(\frac{T}{N}\right)^2+\cdots+\left(\frac{T}{N}\right)^{K-M-1}\right)^{-1}.
\end{aligned}
$$

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 $\boldsymbol{S}$ 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 $I\left(\boldsymbol{\Theta},\boldsymbol{S};\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{W}_{[K]}\right)=0$, which in turn implies that $I\left(\boldsymbol{\Theta};\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{W}_{[K]}\mid\boldsymbol{S}\right)=0$. Therefore,

$$
\begin{aligned}
C_{\text{TPIR-PSI}}&\leq C_{\text{TPIR-GSI}}\\
&=\left(1+\frac{T}{N}+\left(\frac{T}{N}\right)^2+\cdots+\left(\frac{T}{N}\right)^{K-M-1}\right)^{-1}.
\end{aligned}
$$

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 $R\leq 1-T/N$: Let us start with an intuitive understanding of the upper bound, $R\leq 1-T/N$. 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 $S\in\mathcal{S}$, 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),

$$
0=I\left(\boldsymbol{W}_{\overline{(\boldsymbol{\Theta},\boldsymbol{S})}};\boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{S},\boldsymbol{\Theta}\right) \tag{66}
$$

Therefore, $\forall\mathcal{T}\subset[N],|\mathcal{T}|=T,\forall S\in\mathcal{S}$, and for all distinct $\theta,\theta'\in[K]\setminus S$,

$$
\begin{aligned}
0&=I\left(\boldsymbol{W}_{\theta'};\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{\Theta}=\theta,\boldsymbol{S}=S\right) && (67)\\
&=I\left(\boldsymbol{W}_{\theta'};\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{\Theta}=\theta,\boldsymbol{S}=S\right)\\
&\quad+I\left(\boldsymbol{W}_{\theta'};\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{\Theta}=\theta,\boldsymbol{S}=S\right) && (68)\\
&=H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{\Theta}=\theta,\boldsymbol{S}=S\right)\\
&\quad-H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{W}_{\theta'},\boldsymbol{\Theta}=\theta,\boldsymbol{S}=S\right) && (69)\\
&\overset{(34)}{=}H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{\Theta}=\theta,\boldsymbol{S}=S\right)\\
&\quad-H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{W}_{\theta'},\boldsymbol{\Theta}=\theta',\boldsymbol{S}=S\right) && (70)
\end{aligned}
$$

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

$$
\begin{aligned}
L&=H\left(\boldsymbol{W}_{\theta'}\right)\\
&\overset{(6)}{=}I\left(\boldsymbol{W}_{\theta'};\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{\Theta}=\theta',\boldsymbol{S}=S\right) && (71)\\
&=H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{\Theta}=\theta',\boldsymbol{S}=S\right)\\
&\quad-H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\theta'},\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{\Theta}=\theta',\boldsymbol{S}=S\right) && (72)\\
&\leq H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{\Theta}=\theta',\boldsymbol{S}=S\right)\\
&\quad-H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\theta'},\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{\Theta}=\theta',\boldsymbol{S}=S\right) && (73)\\
&=H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\mid\boldsymbol{W}_{\boldsymbol{S}},\boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]},\boldsymbol{\Theta}=\theta',\boldsymbol{S}=S\right)
\end{aligned}
$$

$$
\begin{aligned}
&\quad - H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\theta'}, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right) \qquad (74)\\
&\overset{(70)}{=} H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right)\\
&\quad - H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right) \qquad (75)\\
&\overset{(35)}{=} H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right)\\
&\quad - H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right) \qquad (76)\\
&\leq H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right)\\
&\quad - H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right), \qquad (77)
\end{aligned}
$$

where (74) follows due to Lemma 1. Writing (77) for all $\mathcal{T} \subset [1:N], |\mathcal{T}|=T$, adding those inequalities and divided by $\binom{N}{T}$ we obtain,

$$
\begin{aligned}
L &\leq H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right)\\
&\quad - \frac{1}{\binom{N}{T}} \sum_{\mathcal{T}} H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right) \qquad (78)\\
&\leq H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right)\\
&\quad - \frac{T}{N} H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right) \qquad (79)\\
&= \left(1-\frac{T}{N}\right) H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right) \qquad (80)
\end{aligned}
$$

where (79) is due to Han's inequality. Since this inequality is true for all $S \in \mathcal{S}, \theta' \in [K] \setminus S$, it is also true when averaged across them, so,

$$
\begin{aligned}
L &\leq \left(1-\frac{T}{N}\right) H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}, \boldsymbol{S}\right) \qquad (81)\\
&\leq \left(1-\frac{T}{N}\right) H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}\right) \qquad (82)\\
&\leq \left(1-\frac{T}{N}\right) D, \qquad (83)
\end{aligned}
$$

where (82) holds because dropping conditioning does not reduce entropy. Therefore, $R = \lim_{L\to\infty} \frac{L}{D} \leq 1 - \frac{T}{N}$, and we have shown that the rate of any feasible STPIR-SI 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 $\boldsymbol{U}$ 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 $\boldsymbol{S} = S \in \mathcal{S}$ and for $\Theta = \theta \in [K] \setminus S$, consider the answering strings $\boldsymbol{A}_1^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \cdots, \boldsymbol{A}_N^{[\boldsymbol{\Theta},\boldsymbol{S}]}$ and the side information $\boldsymbol{W}_{\boldsymbol{S}}$, from which the user can retrieve $\boldsymbol{W}_\theta$. According to the database-privacy constraint, we have

$$
\begin{aligned}
0 &= I\left(\boldsymbol{W}_{\overline{(\theta,S)}} ; \boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\overset{(6)}{=} I\left(\boldsymbol{W}_{\overline{(\theta,S)}} ; \boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_\theta \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\overset{(9)}{=} I\left(\boldsymbol{W}_{\overline{(\theta,S)}} ; \boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\geq I\left(\boldsymbol{W}_{\overline{(\theta,S)}} ; \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&= H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\quad - H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{[K]}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\overset{(10)}{=} H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\quad - H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{[K]}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\quad + H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{[K]}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{U}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&= H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\quad - I\left(\boldsymbol{U}; \boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{[K]}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right)\\
&\geq H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_\theta, \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right) - H(\boldsymbol{U})\\
&\overset{(70)}{=} H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta', \boldsymbol{S}=S\right) - H(\boldsymbol{U})\\
&\overset{(35)}{=} H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right) - H(\boldsymbol{U}).
\end{aligned}
$$

Therefore,

$$
H(\boldsymbol{U}) \geq H\left(\boldsymbol{A}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right). \qquad (84)
$$

Adding (84) for all $\mathcal{T} \subset [N], |\mathcal{T}|=T$ and divided by $\binom{N}{T}$, we obtain,

$$
\begin{aligned}
H(\boldsymbol{U}) &\geq \frac{T}{N} H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{\mathcal{T}}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right) \qquad (85)\\
&\geq \frac{T}{N} H\left(\boldsymbol{A}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]} \mid \boldsymbol{W}_{\boldsymbol{S}}, \boldsymbol{Q}_{[N]}^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{\Theta}=\theta, \boldsymbol{S}=S\right) \qquad (86)\\
&\overset{(80)}{\geq} \frac{T}{N}\left(\frac{N}{N-T}\right) L = \left(\frac{T}{N-T}\right) L. \qquad (87)
\end{aligned}
$$

$$
\Rightarrow \quad \rho = \frac{H(\boldsymbol{U})}{L} \geq \frac{T}{N-T} \quad \text{(letting } L \to \infty\text{)}. \qquad (88)
$$

Note that (85) is due to Han's inequality. Thus the amount of common randomness normalized by the message size for any feasible STPIR-PSI scheme cannot be less than $T/(N-T)$.

VI. Conclusion

In this paper, the capacity of TPIR-PSI and the capacity of STPIR-PSI are characterized. As a special case of TPIR-PSI obtained by setting $T=1$, the result settles the capacity of PIR-PSI, an open problem highlighted by Kadhe et al. in [20]. Notably, the results of our work (initially limited to capacity of PIR-PSI for $T=1$ as reported in our original ArXiv posting in 2017 [32]) have subsequently been generalized to multi-message PIR-PSI in [33]. Other generalizations, e.g., PIR-PSI with multi-round communication, secure and/or coded storage, remain promising directions for future work, as are the capacity characterizations for PIR-SI (multiple databases) and PIR-SPSI which remain open.

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., $I\left(\boldsymbol{\Theta}; \boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}\right) = 0, \forall n \in [N]$.
- **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., $I\left(\boldsymbol{\Theta}; \boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}\right) = I\left(\boldsymbol{S}; \boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}\right) = 0, \forall n \in [N]$.
- **PIR-PSI**, 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., $I\left(\boldsymbol{\Theta}, \boldsymbol{S}; \boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]}\right) = 0, \forall n \in [N]$.
- **PIR-GSI**, or PIR with global side information. The side information is globally known, i.e., the databases are also fully knowledgeable about the side information. In this case, the privacy of the desired message index must be preserved in spite of the globally known side information, $I\left(\boldsymbol{\Theta}; \boldsymbol{Q}_n^{[\boldsymbol{\Theta},\boldsymbol{S}]}, \boldsymbol{W}_{[K]} \mid \boldsymbol{S}\right) = 0, \forall n \in [N]$.

From the result of Theorem 1 we know the capacity of PIR-PSI 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 $\lceil \frac{K}{M+1} \rceil^{-1}$ 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 $\left(\frac{K}{2}\right)^{-1} = \lceil \frac{K}{2} \rceil^{-1}$, 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 θ 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: $P_1 = W_{i_1} + W_{i_2}, P_2 = W_{i_3} + W_{i_4}, P_3 = W_{i_5} + W_{i_6}$. The indices $(i_1, i_2, \cdots, i_6)$ are obtained by first randomly permuting $(1, 2, \cdots, 6)$ 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 θ, i.e., one of the partitions must contain $W_\theta + W_s$. The scheme is correct because the user can recover W_θ from the sum $W_\theta + W_s$ (because W_s is already available to the user as side information). It is easily verified that θ and s are each uniformly distributed over $(i_1, i_2, \cdots, i_6)$, so the scheme preserves their separate privacy. However, since θ, s must appear in the same partition, it is also clear that their *joint* privacy is not preserved. For example, (θ, s) cannot be equal to (i_1, i_3). The general scheme in [20] works for any even K by partitioning the messages into sets of size 2, one of which contains both θ and s. Each of θ and s is uniformly distributed over the indices but they are not jointly uniform.

B. PIR-SPSI: $N=1$, $M=1$, K odd

For this setting also the capacity of PIR-SPSI is $\left(\frac{K+1}{2}\right)^{-1} = \lceil \frac{K}{2} \rceil^{-1}$, the same as the capacity of PIR-SI. Once again, the converse is trivially inherited 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 $\mathbb{F}_5$. 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 $P_4 = W_{i_5} + 2W_{i_6} + 3W_{i_7}$. In fact, P_3, P_4 can be the non-systematic symbols of any $(5,3)$ systematic MDS code applied to $W_{i_5}, W_{i_6}, W_{i_7}$. Once again, the indices $(i_1, i_2, \cdots, i_7)$ are obtained by first randomly permuting $(1, 2, \cdots, 7)$ 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 θ. If W_θ and W_s appear in P_1 or P_2 then W_θ is decoded by subtracting the side-information, while if W_θ and W_s appear in partitions P_3, P_4 with interfering message W_i, then after eliminating the known side information W_s, the two equations can be solved for the remaining two variables W_θ, W_i (equivalently, the MDS property guarantees decodability). Once again, it is easily verified that θ and s are each uniformly distributed over $(i_1, i_2, \cdots, i_7)$, so the scheme preserves their separate privacy. However, since θ, 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 $W_{i_1} + W_{i_2}$, $W_{i_3} + W_{i_4}$, $\cdots$, $W_{i_{K-4}} + W_{i_{K-3}}$, $W_{i_{K-2}} + W_{i_{K-1}} + W_{i_K}$ and $W_{i_{K-2}} + 2W_{i_{K-1}} + 3W_{i_K}$, and generating the indices $(i_1, i_2, \cdots, i_K)$ by first randomly permuting $(1, 2, \cdots, K)$ 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 θ. This ensures that θ and s are each uniformly distributed over $(i_1, i_2, \cdots, i_K)$, so the scheme preserves their separate privacy. However, since θ, s must appear in the same partition, it is also clear that their *joint* privacy is not preserved.

C. PIR-SPSI: $N=1$, $M=2$, $K=6$

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-SPSI 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 θ. 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-SPSI 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 bits. We will show that conditioned on each realization of the query, the download from the database must be at least $3L$ bits, which also proves that the average download must be at least $3L$ bits. To set up a proof by contradiction, let us assume that conditioned on the query realization $\mathbf{Q} = q$, the download $\mathbf{A}$ from the database is less than $3L$ bits. This assumption implies that,

$$H(\mathbf{A} \mid \mathbf{Q} = q) < 3L. \tag{89}$$

The conditioning on $\mathbf{Q} = q$ will be assumed throughout the remainder of this proof.

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

$$H_{\mathbf{A}}(W_P) \triangleq H\left(\mathbf{A} \mid \mathbf{Q} = q, W_{[K] \setminus P}\right). \tag{90}$$

Intuitively, $H_{\mathbf{A}}(W_P)$ represents the entropy that remains in the answer $\mathbf{A}$ due to messages W_P (after all other messages are known), i.e., the 'space' occupied by the messages W_P in $\mathbf{A}$. We need the following facts.

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

1) *If P is a singleton set, e.g., $P = \{k\}$, then we must have*

$$H_{\mathbf{A}}(W_k) \geq L, \ \forall k \in [K]. \tag{91}$$

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

$$H_{\mathbf{A}}(P_1) \leq H_{\mathbf{A}}(P_2). \tag{92}$$

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

$$H_{\mathbf{A}}(W_l, W_m, W_n) < 2L, \tag{93}$$

$$H_{\mathbf{A}}(W_\theta, W_i) \geq 2L, \ \forall i \in \{l, m, n\}. \tag{94}$$

Proof. The first fact, (91) holds because given the answer $\mathbf{A}$ and all messages except W_k (which must include the side information), the user must be able to decode W_k, therefore $L = I(W_k; \mathbf{A} \mid \mathbf{Q} = q, W_{[K] \setminus \{k\}}) \leq H_{\mathbf{A}}(W_k)$. 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.

$$\begin{aligned} L &= I(W_\theta; \mathbf{A} \mid \mathbf{Q} = q, W_{s_1}, W_{s_2}) && (95) \\ &= H(\mathbf{A} \mid \mathbf{Q} = q, W_{s_1}, W_{s_2}) \\ &\quad - H(\mathbf{A} \mid \mathbf{Q} = q, W_{s_1}, W_{s_2}, W_\theta) && (96) \\ &\leq H(\mathbf{A} \mid \mathbf{Q} = q) - H_{\mathbf{A}}(W_l, W_m, W_n) && (97) \\ &< 3L - H_{\mathbf{A}}(W_l, W_m, W_n) && (98) \end{aligned}$$

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 $\mathbf{A}$ and his side information,

$$\begin{aligned} L &= I(W_\theta; \mathbf{A} \mid \mathbf{Q} = q, W_{[K] \setminus \{\theta, i\}}) && (99) \\ &= H_{\mathbf{A}}(W_\theta, W_i) - H_{\mathbf{A}}(W_i) && (100) \\ &\leq H_{\mathbf{A}}(W_\theta, W_i) - L && (101) \end{aligned}$$

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 $\mathbf{Q} = q$, 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,

$$H_{\mathbf{A}}(W_d, W_e, W_f) < 2L, \tag{102}$$

$$H_{\mathbf{A}}(W_c, W_i) \geq 2L, \ \forall i \in \{d, e, f\}. \tag{103}$$

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 $H_{\mathbf{A}}(W_d, W_e) \geq 2L$, which contradicts the fact that $H_{\mathbf{A}}(W_d, W_e) \leq H_{\mathbf{A}}(W_d, W_e, W_f) < 2L$ 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

$$H_{\mathbf{A}}(W_a, W_i) \geq 2L, \ \forall i \in \{d, e, f\}. \tag{104}$$

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 $H_{\mathbf{A}}(W_b, W_c, W_d) < 2L$. Along with (92) this implies that $H_{\mathbf{A}}(W_c, W_d) < 2L$ 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

$$H_{\mathbf{A}}(W_a, W_b, W_c) < 2L. \tag{105}$$

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 $H_{\mathbf{A}}(W_b, W_c) \geq 2L$, 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

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

[2] S. Yekhanin, "Private Information Retrieval," *Communications of the ACM*, vol. 53, no. 4, pp. 68–73, 2010.

[3] A. Beimel, Y. Ishai, and E. Kushilevitz, "General constructions for information-theoretic private information retrieval," *Journal of Computer and System Sciences*, vol. 71, no. 2, pp. 213–247, 2005.

[4] W. Gasarch, "A Survey on Private Information Retrieval," in *Bulletin of the EATCS*, 2004.

[5] R. Ostrovsky and W. E. Skeith III, "A Survey of Single-database Private Information Retrieval: Techniques and Applications," in *Public Key Cryptography–PKC 2007*. Springer, 2007, pp. 393–411.

[6] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998, pp. 151–160.

[7] N. Shah, K. Rashmi, and K. Ramchandran, "One Extra Bit of Download Ensures Perfectly Private Information Retrieval," in *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, 2014, pp. 856–860.

[8] H. Sun and S. A. Jafar, "Blind interference alignment for private information retrieval," *IEEE International Symposium on Information Theory (ISIT)*, pp. 560–564, 2016.

[9] ——, "The Capacity of Private Information Retrieval," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4075–4088, July 2017.

[10] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private Information Retrieval from MDS Coded Data in Distributed Storage Systems," *IEEE Transactions on Information Theory*, 2018.

[11] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private Information Retrieval for Coded Storage," *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, pp. 2842–2846, 2015.

[12] K. Banawan and S. Ulukus, "The Capacity of Private Information Retrieval from Coded Databases," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1945–1956, 2018.

[13] H. Sun and S. A. Jafar, "The Capacity of Robust Private Information Retrieval with Colluding Databases," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2361–2370, 2018.

[14] ——, "The capacity of symmetric private information retrieval," *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 322–329, 2019.

[15] Q. Wang and M. Skoglund, "Symmetric Private Information Retrieval For MDS Coded Distributed Storage," *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2017.

[16] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *arXiv preprint arXiv:1702.01739*, 2017.

[17] H. Sun and S. A. Jafar, "Multiround Private Information Retrieval: Capacity and Storage Overhead," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5743–5754, August 2018.

[18] Z. Jia, H. Sun, and S. A. Jafar, "Cross Subspace Alignment and the Asymptotic Capacity of X-Secure T-Private Information Retrieval," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5783–5798, 2019.

[19] R. Tandon, "The capacity of cache aided private information retrieval," *55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2017.

[20] S. Kadhe, B. Garcia, A. Heidarzadeh, S. E. Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Transactions on Information Theory*, to appear, 2019.

[21] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3215–3232, 2019.

[22] ——, "Cache-aided private information retrieval with partially known uncoded prefetching: fundamental limits," *IEEE Jour. on Selected Areas in Communications*, vol. 36, no. 6, pp. 1126–1139, 2018.

[23] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "Capacity of single-server single-message private information retrieval with coded side information," *arXiv preprint arXiv:1806.00661*, 2018.

[24] 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," *arXiv preprint arXiv:1807.09908*, 2018.

[25] A. Heidarzadeh, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "Single-server multi-message individually-private information retrieval with side information," *arXiv preprint arXiv:1901.07509*, 2019.

[26] F. Kazemi, E. Karimi, A. Heidarzadeh, and A. Sprintson, "Single-server single-message online private information retrieval with side information," *arXiv preprint arXiv:1901.07748*, 2019.

[27] S. Li and M. Gastpar, "Converse for multi-server single-message pir with side information," *arXiv preprint arXiv:1809.09861*, 2018.

[28] ——, "Single-server multi-user private information retrieval with side information," *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, 2018.

[29] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "The role of coded side information in single-server private information retrieval," *arXiv preprint arXiv:1910.07612*, 2019.

[30] Q. Wang and M. Skoglund, "Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers," *arXiv preprint arXiv:1708.05673*, 2017.

[31] S. Lin and D. J. Costello, *Error control coding*. Prentice hall, 2001, vol. 2.

[32] Z. Chen, Z. Wang, and S. Jafar, "The capacity of private information retrieval with private side information," *arXiv preprint arXiv:1709.03022v1*, Sep. 2017.

[33] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," *arXiv preprint arXiv:1805.11892*, 2018.