On the Information Leakage in Private Information Retrieval Systems

Tao Guo, Ruida Zhou, and Chao Tian
Department of Electrical and Computer Engineering
Texas A&M University
{guotao, ruida, chao.tian}@tamu.edu

***Abstract*—We consider information leakage to the user in private information retrieval (PIR) systems. Information leakage can be measured in terms of individual message leakage or total leakage. Individual message leakage, or simply individual leakage, is defined as the amount of information that the user can obtain on any individual message that is not being requested, and the total leakage is defined as the amount of information that the user can obtain about all the other messages except the one being requested. In this work, we characterize the tradeoff between the minimum download cost and the individual leakage, and that for the total leakage, respectively. New codes are proposed to achieve these optimal tradeoffs, which are also shown to be optimal in terms of the message size. We further characterize the optimal tradeoff between the minimum amount of common randomness and the total leakage. Moreover, we show that under individual leakage, common randomness is in fact unnecessary when there are more than two messages.**

I. Introduction

The problem of *private information retrieval* (PIR) [1], [2] addresses the retrieval of one out of K messages from N replicated databases, without revealing the identity of the desired message to any individual database. The goal is to find an efficient protocol, i.e., with the minimum download cost, to privately retrieve the desired message. The capacity of a PIR system is defined as the maximum number of bits of desired message that can be retrieved per bit of downloaded information, which was shown in [3] to be $C_{\text{PIR}} = \left(1 + 1/N + 1/N^2 + \cdots + 1/N^{K-1}\right)^{-1}$.

The problem considered in this work is closely related to the *symmetric* private information retrieval (SPIR) problem [4], where "symmetric" refers to the fact that both user privacy and database privacy need to be preserved. Database privacy requires that the user obtains no information on other messages beyond the requested message; strictly speaking, this is a security requirement rather than a privacy requirement, and we shall refer to it as such in the sequel. It was shown in [4] that to ensure perfect security, the databases need to share *common randomness* (a common key) which is independent of the messages and only available to the databases. The common key is used for randomizing the answers such that no information about the non-desired messages is leaked. The capacity of SPIR was shown to be $C_{\text{SPIR}} = 1 - 1/N = \left(1 + 1/N + 1/N^2 + \cdots + 1/N^{\infty}\right)^{-1}$, as long as the amount of common randomness is at least $\frac{1}{N-1}$ bits per desired message bit.

If the user can instead obtain some information about the non-desired messages, then the system is said to have information leakage [5]–[7]. This is the type of systems that we wish to understand in this work. Allowing a small amount of information leakage lets us control the system security level in a finer grain manner. In this context, SPIR essentially requires strictly zero information leakage, while the classical PIR does not have a security constraint at all. Thus our goal is to understand the tradeoff between the download cost and the amount of information leakage in the regime other than these two extreme cases. Two different notations of information leakage can be defined: the individual leakage is defined as the amount of information that the user can obtain about any individual non-desired message, and the total leakage as the amount of information that the user can obtain about all the non-desired messages. The former is similar to the weak-security constraint cases seen in the literature [8]–[11], while the latter to the standard strong-security constraint cases.

The main result of this work is the characterizations of the optimal tradeoffs between the download cost and the amount of information leakage for both individual leakage and total leakage. By adapting the capacity-achieving PIR code (referred to as the TSC code) in [12] and the SPIR code in [4], we provide code constructions that can achieve these optimal tradeoffs; moreover, they are also shown to have the minimum message sizes. For the individual leakage case, the constructed codes do not require common randomness (unless there are only two messages). At the extreme case with perfect individual message security, the download cost is in fact the same as that with perfect total security. This is rather reassuring since it implies that the stronger security requirement does not induce any additional download cost.

The rest of the paper is organized as follows. We formally define the problem in Section II. Section III is devoted to our main results on the optimal tradeoffs and the minimum message size. The achievability proofs are given in Section IV, Section V. We conclude the paper in Section VI.

II. Preliminaries

A. Problem Statement

For positive integers K, N, let $[1:K] \triangleq \{1, 2, \cdots, K\}$ and $[1:N] \triangleq \{1, 2, \cdots, N\}$. For any $k \in [1:K]$, define the

This work was supported in part by CCF-18-16546.

complement set by $\bar{k} \triangleq \{1, 2, \cdots, K\} \backslash \{k\}$.

In a *private information retrieval* (PIR) system, there are K independent messages $W_{1:K} = (W_1, W_2, \cdots, W_K)$, each of which is comprised of L i.i.d. symbols uniformly distributed over a finite alphabet $\mathcal{X}$. In $\log_{|\mathcal{X}|}$-ary units, this implies that

$$H(W_{1:K}) = H(W_1) + H(W_2) + \cdots + H(W_K), \quad (1)$$
$$H(W_1) = H(W_2) = \cdots = H(W_K) = L. \quad (2)$$

There are a total of N databases, each of which stores all the messages $W_{1:K}$. A user aims to retrieve a message $W_k,\ k \in [1:K]$ from the N databases without revealing the identity k of the desired message to any individual database. An independent random key $\mathbf{F}$ is used to generate queries $Q_{1:N}^{[k]} = \left(Q_1^{[k]}, Q_2^{[k]}, \cdots, Q_N^{[k]}\right)$, i.e.,

$$H(Q_{1:N}^{[k]} | \mathbf{F}) = 0,\ \forall k \in [1:K], \quad (3)$$

where $Q_n^{[k]} \in \mathcal{Q}_n$ for $n \in [1:N]$. For $n \in [1:N]$, the n-th query $Q_n^{[k]}$ is sent to the n-th database. Since we wish to protect the non-desired messages from unintentional access, the databases may need to share a common random key $S \in \mathcal{S}$ that is not accessible to the user, which induces the condition

$$H(W_{1:K}, \mathbf{F}, S) = H(W_{1:K}) + H(\mathbf{F}) + H(S). \quad (4)$$

Upon receiving $Q_n^{[k]}$, the n-th database generates an answer $A_n^{[k]}$ from the query $Q_n^{[k]}$, the stored messages $W_{1:K}$, and the common key S, i.e.,

$$A_n^{[k]} = \varphi_n(Q_n^{[k]}, W_{1:K}, S),\ \forall k \in [1:K],\ \forall n \in [1:N], \quad (5)$$

which implies

$$H(A_n^{[k]} | Q_n^{[k]}, W_{1:K}, S) = 0,\ \forall k \in [1:K], \forall n \in [1:N]. \quad (6)$$

The answer symbols are from a finite alphabet $\mathcal{Y}$, i.e., $A_n^{[k]} \in \mathcal{Y}^{\ell_n}$, where ℓ_n is the length of the answer. Using all the answers $A_{1:N}^{[k]} = \left(A_1^{[k]}, A_2^{[k]}, \cdots, A_N^{[k]}\right)$ from the N databases and the values of $\mathbf{F}$ and k, the user perfectly decodes the desired message W_k, which further implies that

$$H(W_k | A_{1:N}^{[k]}, \mathbf{F}) = 0. \quad (7)$$

To satisfy the privacy requirement of keeping the desired message index private to any one of the databases, the received queries should be identically distributed, i.e.,

$$Q_n^{[k]} \sim Q_n^{[k']},\ \forall k, k' \in [1:K],\ \forall n \in [1:N]. \quad (8)$$

Since $W_{1:K}$, S, and $\mathbf{F}$ are independent, we have by (3) that

$$(Q_n^{[k]}, W_{1:K}, S) \sim (Q_n^{[k']}, W_{1:K}, S), \quad \forall k, k' \in [1:K],\ \forall n \in [1:N]. \quad (9)$$

Since $A_n^{[k]}$ is a deterministic function of $(Q_n^{[k]}, W_{1:K}, S)$, the following identical distribution constraint must also hold

$$(Q_n^{[k]}, A_n^{[k]}, W_{1:K}, S) \sim (Q_n^{[k']}, A_n^{[k']}, W_{1:K}, S), \quad \forall k, k' \in [1:K],\ \forall n \in [1:N]. \quad (10)$$

In contrast to the perfect security of the non-desired messages $W_{\bar{k}} = (W_1, \cdots, W_{k-1}, W_{k+1}, \cdots, W_K)$ in [4], we allow information leakage in this work. Define the total leakage as $I(W_{\bar{k}}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F})$, which is the amount of information that the user can obtain about all the non-desired messages. Define the individual message leakage for message $k' \in [1:K]$ $(k' \neq k)$ as $I(W_{k'}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F})$, which is the amount of information that the user can obtain about the individual non-desired message-k'. The total leakage and the individual leakage in the systems are constrained as follows

$$\tfrac{1}{L} I(W_{\bar{k}}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F}) \leq s,\ \forall k \in [1:K] \quad (11)$$
$$\tfrac{1}{L} I(W_{k'}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F}) \leq w,\ \forall k' \neq k \in [1:K], \quad (12)$$

where the parameters s and w are used to indicate the strong security requirement and the weak security requirement, respectively. For $K = 2$, since $I(W_{\bar{k}}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F}) = I(W_{k'}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F})$, the total leakage constraint is equivalent to the individual leakage constraint.

In a PIR system, the download cost is defined as

$$D \triangleq \log_{|\mathcal{X}|} |\mathcal{Y}| \sum_{n=1}^{N} \mathbb{E}(\ell_n), \quad (13)$$

where the expectation is taken over the possible query set $\mathcal{Q}_n$. Note that D is a deterministic function of queries and query distribution, but neither the particular realization of messages nor the requested message index k. The amount of common randomness is normalized by the message length L as

$$\rho \triangleq \frac{H(S)}{L}. \quad (14)$$

B. Performance of Several Existing PIR Codes

Before presenting our main result, we provide a simple analysis on the TSC code [12] and the SPIR code [4].

- The TSC code [12]:

$$L = N - 1, \quad (15)$$
$$D = \frac{N^K - 1}{N^{K-1}}, \quad (16)$$
$$\frac{1}{L} I(W_{\bar{k}}; A_{1:N}^{[k]}, \mathbf{F}) = \frac{1}{N-1}\left(1 - \frac{1}{N^{K-1}}\right), \quad (17)$$
$$\frac{1}{L} I(W_{k'}; A_{1:N}^{[k]}, \mathbf{F}) = \frac{1}{N^{K-1}},\ \forall k' \neq k \in [1:K]. \quad (18)$$

- The SPIR code [4]:

$$L = N - 1, \quad D = N, \quad \rho = 1. \quad (19)$$

III. Main Results

A. Total Leakage

The following theorem characterizes the optimal tradeoff between the minimum download cost $D_{\min}$ and total leakage constraint s. For notational convenience, define

$$D_{\min}^0 \triangleq L \cdot \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}}\right). \quad (20)$$

Theorem 1. *If the amount of common randomness satisfies* $\rho \geq \rho_{\min}^s$*, where*

$$\rho_{\min}^s \triangleq \frac{1}{N-1} - \frac{N^{K-1}}{N^{K-1} - 1} \cdot s, \quad (21)$$

then the minimum download cost $D_{\min}$ of the PIR system is given by

$$D_{\min}=\begin{cases}L\cdot\left(\frac{N}{N-1}-\frac{1}{N^{K-1}-1}\cdot s\right), & \text{if } 0\le s\le s_t\\ D^0_{\min}, & \text{otherwise,}\end{cases}\tag{22}$$

where the threshold is defined by $s_t \triangleq \frac{1}{N-1}\left(1-\frac{1}{N^{K-1}}\right)$ *and* $D^0_{\min}$ *is defined in* (20). *If* $\rho<\rho^s_{\min}$, *then* $D_{\min}=\infty$.

Proof. The proof is given in Section IV. □

Remark 1. *The case where $D_{\min}=\infty$ indicates that it is impossible to simultaneously meet all the system requirements, i.e., i) retrieval; ii) privacy; iii) total leakage constraint. In this case, the capacity $C=0$.*

For $\rho\ge\rho^s_{\min}$, the dependency of $D_{\min}$ on (ρ,s) is illustrated by the shaded area in Fig. 1. The dashed lines project the corner points to the axises that show their values. We see that for a given value of s, $D_{\min}$ is a constant and thus independent with ρ. The shaded area is projected onto the $D_{\min}$-s plane, which is drawn in Fig. 2.

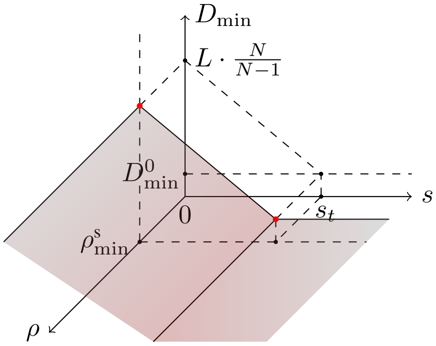

Fig. 1. The dependency of $D_{\min}$ on (ρ,s).

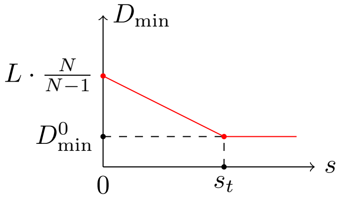

Fig. 2. Tradeoff curve between $D_{\min}$ and s.

The red line in Fig. 2 is the tradeoff curve between $D_{\min}$ and s for $\rho\ge\rho^s_{\min}$. Codes that achieve points on this optimal tradeoff curve will be referred to as *Pareto optimal* codes. We see from Fig. 2 that for $s=0$, the problem reduces to SPIR [4] for which the capacity is $C_{\text{SPIR}}=\left.\frac{L}{D_{\min}}\right|_{s=0}=1-\frac{1}{N}$. We further observe that i) the threshold s_t is equal to the normalized total leakage of the capacity-achieving TSC code in (17); ii) for $s\ge s_t$, the capacity $C=\left.\frac{L}{D_{\min}}\right|_{s=s_t}=1+\frac{1}{N}+\cdots+\frac{1}{N^{K-1}}$ is equal to C_{PIR}.

B. Individual Leakage

The following theorem characterizes the tradeoff between the minimum download cost $D_{\min}$ and the individual leakage constraint w.

Theorem 2. *If the amount of common randomness satisfies* $\rho\ge\rho^w_{\min}$, *where*

$$\rho^w_{\min}\triangleq\begin{cases}\frac{1}{N-1}-\frac{N}{N-1}\cdot w, & \text{if } K=2\\ 0, & \text{if } K\ge 3,\end{cases}\tag{23}$$

then the minimum download cost $D_{\min}$ of the PIR system is

$$D_{\min}=\begin{cases}L\cdot\left(\frac{N}{N-1}-\frac{1}{N-1}w\right), & \text{if } 0\le w\le\frac{1}{N^{K-1}}\\ D^0_{\min}, & \text{otherwise,}\end{cases}\tag{24}$$

where $D^0_{\min}$ is defined in (20). *If $\rho<\rho^w_{\min}$, then $D_{\min}=\infty$.*

Proof. The proof is given in Section V. □

Remark 2. *If $\rho<\rho^w_{\min}$, it is impossible to meet all the requirements of the PIR system simultaneously. In this case, the capacity $C=0$.*

Remark 3. *For $K=2$, the individual leakage is equal to the total leakage. We can see that Theorem 1 and Theorem 2 are equivalent.*

From (23), it is seen that $\rho^{\text{w}}_{\min}=0$ for $K\ge 3$, which is the case when one non-desired messages can be used as the encryption key to protect another individual non-desired message. For $K=2$, there is only one non-desired message, and this makes it impossible to have another non-desired messages as the encryption key. For $\rho\ge\rho^{\text{w}}_{\min}$, similar to Fig. 1 and Fig. 2, we can illustrate the dependency of $D_{\min}$ on (ρ,w) by replacing s_t with $\frac{1}{N^{K-1}}$ in the figures.

Similar to the observations of Fig. 2, for $w=0$, the problem becomes weakly secure PIR (WS-PIR) which differs from SPIR only in the type of security. The capacity of WS-PIR can be obtained by calculating $\frac{L}{D_{\min}}$ at $w=0$ from (24), which is given in the following corollary.

Corollary 1. *If the amount of common randomness satisfies*

$$\rho\ge\begin{cases}\frac{1}{N-1}, & \text{if } K=2\\ 0, & \text{if } K\ge 3,\end{cases}\tag{25}$$

the capacity of WS-PIR is given by

$$C_{\textit{WS-PIR}}=1-\frac{1}{N}.\tag{26}$$

If (25) *is not satisfied, the capacity $C_{\textit{WS-PIR}}=0$.*

From this corollary, it is seen that $C_{\text{WS-PIR}}=C_{\text{SPIR}}$. It is satisfying to see that requiring strong security does not increase download cost compared to requiring weak security, even though the required amount of common randomness is larger. It is straightforward to verify that the optimal SPIR code in [4] is also optimal for WS-PIR. In Section V-A, we propose another optimal code for WS-PIR with $K\ge 3$, where the databases do not need to share common randomness.

C. Minimum Message Size

The minimum message size (in $\log_{|\mathcal{X}|}$-ary unit) is highly dependent on the download cost. It was shown in [12] that the message size of a capacity-achieving PIR code is greater than or equal to $N-1$ if the code is uniformly decomposable (a generalization of linear code). For the definition of uniformly decomposable code, we refer the readers to [12].

We can see from the Pareto optimal code constructions in Sections IV and V that for most cases, we can simultaneously achieve a message size of $N-1$ and the minimum amount of common randomness characterized in Theorems 1 and 2. The only exception is the individual leakage case with $K \geq 3$ and $|\mathcal{X}| = N = 2$, for which the message size is twice the minimum. Except for this case, the following theorem shows that the minimum message size is the same as that of the classical PIR codes in [12].

Theorem 3. *Except for the case of $K \geq 3, |\mathcal{X}| = N = 2$, the minimum message size of any Pareto optimal uniformly decomposable PIR codes achieving the minimum amount of common random randomness with either individual leakage constraint w or total leakage constraint s is $(N-1)\log_{|\mathcal{X}|}|\mathcal{Y}|$; in particular, it equals to $N-1$ if we restrict $\mathcal{Y} = \mathcal{X}$.*

Proof. Except for the case of $K \geq 3, |\mathcal{X}| = N = 2$, we have designed Pareto optimal codes achieving simultaneously the minimum amount of common random randomness and the minimum message size. Thus, we only need to prove the lower bound $L \geq (N-1)\log_{|\mathcal{X}|}|\mathcal{Y}|$, for which the details can be found in [13]. □

IV. Proof of Theorem 1

We only present the achievability here and the converse can be found in [13]. Following the analysis of Fig. 2, we only need to prove the theorem for $s \in [0, s_t]$. Consider a message length of $L = N-1$. In addition to the query generation in (3), now the random key $\mathbf{F}$ generates one more indicator bit $F_0 \in \{0,1\}$, according to probability $P(F_0 = 0) = \frac{N^{K-1}(N-1)s}{N^{K-1}-1} \in [0,1]$ and $P(F_0 = 1) = 1 - \frac{N^{K-1}(N-1)s}{N^{K-1}-1}$. For $x, y \in [1:N]$, it is useful to define the operation $(x+y)_N$ by

$$(x+y)_N = \begin{cases} x+y, & \text{if } x+y \leq N \\ x+y-N, & \text{if } x+y > N. \end{cases} \quad (27)$$

And similarly,

$$(x-y)_N = \begin{cases} x-y, & \text{if } x-y > 0 \\ x-y+N, & \text{if } x-y \leq 0. \end{cases} \quad (28)$$

Let $(F_1, F_2, \cdots, F_{K-1})$ be chosen uniformly from $[1:N]^{K-1}$. For random key $\mathbf{F} = (F_0, F_1, F_2, \cdots, F_{K-1}) \in \{0,1\} \times [1:N]^{K-1}$, let $F^* \triangleq \left(\sum_{i=1}^{K-1} F_i\right)_N$. Then the query for the n-th database is generated as

$$Q_n^{[k]} = \left(F_0, F_1, \cdots, F_{k-1}, (n-F^*)_N, F_k, \cdots, F_{K-1}\right), \quad n \in [1:N], k \in [1:K]. \quad (29)$$

Since the query is a length-$(K+1)$ vector, we can denote it by $Q_{n,0:K}^{[k]}$. Upon receiving the queries, the databases generate the answers using the TSC code for $F_0 = 0$, and SPIR code for $F_0 = 1$. Specifically, the answer is

$$A_n^{[k]} = \begin{cases} W_{1,Q_{n,1}^{[k]}} \oplus \cdots \oplus W_{K,Q_{n,K}^{[k]}}, & \text{if } Q_{n,0}^{[k]} = 0 \\ W_{1,Q_{n,1}^{[k]}} \oplus \cdots \oplus W_{K,Q_{n,K}^{[k]}} \oplus S, & \text{if } Q_{n,0}^{[k]} = 1, \end{cases} \quad (30)$$

where S is the common randomness shared among all databases. The retrieval and privacy requirements are easily seen from the TSC and SPIR codes. The performance of information leakage, download cost, and amount of common randomness can be easily verified. This proves the achievability of Theorem 1 for $s \in [0, s_t]$.

Remark 4. *The code is obtained by combining TSC code and SPIR code probabilistically, which is done by sending one more query bit as an indicator. The method is valid only when the two codes have the same message length, and the message length of the combined code remains the same, which is $N-1$ here. This combination method outperforms time-sharing in the sense that, with real-valued combination coefficients, time-sharing may require a much larger message size.*

V. Proof of Theorem 2

We only show the achievability here and the converse can be found in [13].

A. Optimal code for WS-PIR

Even though the SPIR code in [4] achieves the capacity of WS-PIR, it does not always achieve the minimum amount of common randomness for WS-PIR. For $K \geq 3$, (25) indicates that the databases do not need to share common randomness. We consider the following three cases:

i) $|\mathcal{X}| \geq 3$ and $N \geq 2$;
ii) $|\mathcal{X}| = 2$ and $N \geq 3$;
iii) $|\mathcal{X}| = N = 2$.

Next, we propose an optimal code for WS-PIR that uses the sum of all message symbols as encryption key shared by databases and no extra randomness is needed. The code design is simply to modify TSC code by adding the shared encryption key (sum of all message symbols) to each of the answers.

Case i): The code has a message length of $L = N-1$. Specifically, by appending dummy variables $W_{k,N} = 0$, the message W_k can be written as

$$W_k = (W_{k,1}, W_{k,2} \cdots, W_{k,N-1}, W_{k,N}). \quad (31)$$

Let the random key $\mathbf{F}$ of the user be chosen uniformly from $[1:N]^{K-1}$ which gives

$$\mathbf{F} = (F_1, F_2, \cdots, F_{K-1}). \quad (32)$$

For $x, y \in [1:N]$, the operation $(x+y)_N$ and $(x-y)_N$ are defined by (27) and (28). Let $F^* \triangleq \left(\sum_{i=1}^{K-1} F_i\right)_N$. For $k \in [1:K]$ and $n \in [1:N]$, the query is a deterministic function of the random key $\mathbf{F}$, defined as

$$Q_n^{[k]} = \big(F_1, F_2, \cdots, F_{k-1}, (n-F^*)_N, F_k, \cdots, F_{K-1}\big). \quad (33)$$

The sum of all message symbols is denoted by S, which is

$$S = \sum_{k=1}^{K} \big(W_{k,1} \oplus W_{k,2} \oplus \cdots \oplus W_{k,N-1}\big). \tag{34}$$

Upon receiving the query $Q_n^{[k]}$, the n-th database generates an answer $A_n^{[k]}$ using $Q_n^{[k]}$ as linear combination indexes of all the message symbols. We further add the encryption key S to each answer and obtain that

$$\begin{aligned} A_n^{[k]} = W_{1,F_1} \oplus \cdots \oplus W_{k-1,F_{k-1}} \oplus W_{k,(n-F^*)_N} \\ \oplus W_{k+1,F_k} \oplus \cdots \oplus W_{K,F_{K-1}} \oplus S. \end{aligned} \tag{35}$$

For simplicity, we define

$$B = W_{1,F_1} \oplus \cdots \oplus W_{k-1,F_{k-1}} \oplus W_{k+1,F_k} \oplus \cdots \oplus W_{K,F_{K-1}}. \tag{36}$$

Substituting (36) into (35), we have

$$A_n^{[k]} = W_{k,(n-F^*)_N} \oplus B \oplus S. \tag{37}$$

The user receives all the answers $A_{1:N}^{[k]}$ and we see that

$$W_{k,(n-F^*)_N} = A_n^{[k]} \ominus A_{F^*}^{[k]} = A_n^{[k]} \ominus (B \oplus S), \tag{38}$$

where $\ominus$ is the subtraction operation in the Abelian group $\mathcal{X}$. The message W_k can be recovered by ranging n from 1 to N.

Since $\mathbf{F}$ is chosen uniformly from $[1:N]^{K-1}$ and all the queries are deterministic functions of $\mathbf{F}$, we see that $Q_n^{[k]}$ is chosen uniformly from the query set for any $k \in [1:K]$ and $n \in [1:N]$. Thus $Q_n^{[k]}$ provides no information about the message index k, which ensures the privacy.

The weak security (individual leakage) can be seen as follows. The coefficient of each message symbol $W_{k,i}$ ($k \in [1:K], i \in [1:N-1]$) in the expression of $B \oplus S$ can only be 1 or 2. Because of the assumption that $|\mathcal{X}| \geq 3$, none of the message symbols will vanish. Since $K \geq 3$, there is at least one $k'' \in [1:K]$ such that $k'' \neq k, k'$. Then the message symbols of $W_{k''}$ randomizes the sum of symbols from W_k and $W_{k'}$, and thus we obtain that

$$I(B \oplus S; W_{k'} W_k) = 0, \tag{39}$$

which implies

$$I(B \oplus S; W_k) = I(B \oplus S; W_{k'} | W_k) = 0. \tag{40}$$

Thus, we have

$$I(W_{k'}; B \oplus S, W_k) = I(W_{k'}; W_k) + I(W_{k'}; B \oplus S | W_k) = 0. \tag{41}$$

To see the security, we consider the following,

$$\begin{aligned} I(W_{k'}; A_{1:N}^{[k]} Q_{1:N}^{[k]} \mathbf{F}) &= I(W_{k'}; A_{1:N}^{[k]} Q_{1:N}^{[k]} W_k \mathbf{F}) && (42) \\ &= I(W_{k'}; A_{1:N}^{[k]} W_k \mathbf{F}) && (43) \\ &= I(W_{k'}; B \oplus S, W_k | \mathbf{F}) && (44) \\ &= 0, && (45) \end{aligned}$$

where (42) follows from the recovery in (38), (44) follows from the expression of answer in (37) and $I(W_{k'}; \mathbf{F}) = 0$, and the last equality follows from (41) and $B \oplus S$ is a function of messages $W_{1:K}$ and the random key $\mathbf{F}$.

Since each answer consists of one symbol, the download cost is $D = N$. Since $L = N-1$, the rate of the code is simply $R = \frac{N-1}{N} = 1 - \frac{1}{N}$ that matches the capacity in (26) of Corollary 1. Lastly, the databases do not share common randomness in addition to the messages themselves.

Case ii): It is easy to verify that the code also works for $|\mathcal{X}| = 2$ and $N \geq 3$. We omitted the details here, which can be found in [13].

Case iii): For the case that $N = 2$ and $|\mathcal{X}| = 2$, let $L = 2(N-1) = 2$. We can divide each message into two sub-messages, i.e., $W_k = (W_k^{(1)}, W_k^{(2)})$ for $k \in [1:K]$. Then we apply the code in Case i) to each sub-message, where we choose the same encryption key as $S = \sum_{k=1}^{K} W_k^{(1)} \oplus W_k^{(2)}$. The parameters $\mathbf{F}^{(1)}, \mathbf{F}^{(2)} \in \{1,2\}^{K-1}$ are chosen to be dual of each other, i.e., $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ differ in every position. The private retrieval requirement is easy. We can see the weak security from the following example of $K = 3$. Let $k = 2$ and the random key to be chosen as $\mathbf{F}^{(1)} = (2,1)$ and $\mathbf{F}^{(2)} = (1,2)$, then the queries are

$$q_1^{(1)} = (2,1,1),\ q_2^{(1)} = (2,2,1) \tag{46}$$

$$q_1^{(2)} = (1,1,2),\ q_2^{(2)} = (1,2,2). \tag{47}$$

Then $B^{(1)} \oplus S$ and $B^{(2)} \oplus S$ of the two sub-codes are thus

$$B^{(1)} \oplus S = W_{1,2}^{(1)} \oplus W_{3,1}^{(1)} \oplus S = W_1^{(1)} \oplus W_1^{(2)} \oplus W_3^{(2)}, \tag{48}$$

$$B^{(2)} \oplus S = W_{1,1}^{(2)} \oplus W_{3,2}^{(2)} \oplus S = W_1^{(1)} \oplus W_3^{(1)} \oplus W_3^{(2)}. \tag{49}$$

We can easily see that $(B^{(1)} \oplus S, B^{(2)} \oplus S)$ provides no information about either W_1 or W_3. Then using similar arguments as (39)-(45), we can obtain the weak security.

B. Achievability of Pareto Optimal Points

We only need to prove the theorem for $w \in [0, w_t]$. Since for $K = 2$, the WS-PIR code reduces to SPIR code, we consider only $K \geq 3$ here. Similar to Section IV, consider a message length of $L = N-1$. The random key $\mathbf{F}$ generates one more indicator bit $F_0 \in \{0,1\}$ according to probability $P(F_0 = 0) = wN^{K-1} \in [0,1]$ and $P(F_0 = 1) = 1 - wN^{K-1}$. The queries and answers are generated similarly as that in Section IV. The only difference is to replace S by the encryption key using (34). The amount of leakage, download cost, and common randomness can be verified accordingly. This proves the achievability of Theorem 2 for $w \in [0, w_t]$.

VI. Conclusion

In this paper, we studied the PIR problem with total and individual leakage constraints, respectively. Our main contribution was the characterization of the tradeoffs between the minimum download cost $D_{\min}$ and the total leakage constraint s and individual leakage constraint w. The minimum amount of common randomness with respect to s and w was also characterized. It was shown that the linear combination of TSC code and SPIR (WS-PIR) code is Pareto optimal (achieving the whole tradeoff curve). The proposed Pareto optimal codes for both individual and total leakage were proved to have a minimum message size of $N-1$.

References

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 41–50, Oct. 1995.

[2] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *Journal of the ACM (JACM)*, vol. 45, pp. 965–981, Nov. 1998.

[3] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 63, pp. 4075–4088, Jul. 2017.

[4] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, pp. 322–329, Jan. 2019.

[5] H. Yamamoto, "On secret sharing communication systems with two or three channels," *IEEE Trans. Inf. Theory*, vol. 32, pp. 387–393, May 1986.

[6] R. W. Yeung, *Information Theory and Network Coding*. Springer, 2008.

[7] T. Guo, X. Guang, and K. Shum, "Symmetric multilevel imperfect secret sharing," in *2018 IEEE Information Theory Workshop (ITW)*, (Guangzhou, China), Dec. 2018.

[8] K. Bhattad and K. R. Narayanan, "Weakly secure network coding," in *Proc. NetCod 2005*, (Riva del Garda, Italy), Apr. 2005.

[9] M. Yan and A. Sprintson, "Weakly secure network coding for wireless cooperative data exchange," in *IEEE Global Telecommunications Conference (GLOBECOM)*, (Kathmandu, Nepal), Dec. 2011.

[10] T. Guo, C. Tian, T. Liu, and R. W. Yeung, "Weakly secure symmetric multilevel diversity coding," in *2019 IEEE Information Theory Workshop (ITW)*, (Visby, Gotland, Sweden), Aug. 2019.

[11] Y. Chen, O. O. Koyluoglu, and A. J. H. Vinck, "On secure communication over the multiple access channel," in *International Symposium on Information Theory and Its Applications (ISITA)*, (Monterey, CA, USA), Oct. 2016.

[12] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Trans. Inf. Theory*, vol. 65, pp. 7613–7627, Nov. 2019.

[13] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *arXiv*, Sep. 2019. (full version). Available: https://arxiv.org/abs/1909.11605.