Weakly Private Information Retrieval Under the Maximal Leakage Metric

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

***Abstract*—In the canonical *private information retrieval* (PIR) problem, a user retrieves a message from a set of databases without allowing any individual database to obtain any knowledge regarding the identity of the requested message. This perfect privacy requirement may be too stringent in many cases, and the user may only wish to control the amount of the privacy leakage to below a given level, and in return, can retrieve the message at a lower communication cost. In this work, we study the tradeoff between the download cost and the amount of privacy leakage under the maximal leakage metric. A new scheme is proposed by allowing a more flexible query structure and probability distributions in a code previously proposed by Tian et al., which utilized a fixed query set and a uniform distribution. It is shown that the optimal probability distribution in the proposed scheme has a particularly simple structure, which leads to a closed form achievability bound for the optimal tradeoff between the download cost and the privacy leakage. The proposed scheme includes several known schemes, such as those proposed by Lin et al., by Samy et al., and by Jia, as special cases.**

I. Introduction

In the canonical *private information retrieval* (PIR) problem, a user wishes to retrieve one out of K messages from N replicated databases, without revealing any information about the identity of the requested message to any individual database. Efficient protocols need to be designed to minimize the communication cost when completing this task. The PIR capacity is defined as the maximum number of bits of desired message that can be retrieved per bit of downloaded information, which was characterized recently by Sun and Jafar [1]. A class of capacity-achieving codes with a small query set was proposed more recently by Tian, Sun, and Chen [2], which is instrumental in our work and will be referred to as the TSC scheme in the sequel. Many variations and extensions to the canonical PIR problem have also been considered [3]–[10], for example, using more general storage codes, allowing servers to collude, or imposing other security constraints.

The perfect privacy required in the canonical PIR setting can be overly stringent, and a small amount of privacy leakage may be acceptable in practice; we refer to this setting as weakly private information retrieval (WPIR). This relaxation was considered in a few recent works for the case of two databases. In [11], mutual information was used to measure the privacy leakage, differential privacy was adopted in [12], and a less explicit metric involving conditional entropy was used in [13]. Several codes were proposed in these works to minimize the download cost, however, the case for more general number of databases was not considered.

This work was supported in part by the National Science Foundation under Grant CCF-18-16546.

In this paper, we study the WPIR problem for systems with a general number of databases, and adopt the maximal leakage metric [14] to measure the amount of privacy leakage. The advantages of the maximal leakage metric, in contrast to others, such as the mutual information [11], are as follows. Firstly, as argued in [14], maximal leakage has a clear operational meaning, which is also applicable in the PIR setting; secondly, the amount of leakage under this metric depends only on the retrieval strategy, but not the probability distribution of the requested message, i.e., it is not necessary to assume a priori a (uniform) probability distribution of the request.

We propose a new code construction by allowing a more flexible query structure and a nonuniform probability distribution in the TSC scheme. In contrast to the symmetry-guided code construction given in [1], the small query set in the TSC scheme allows us to control more conveniently the query probability distribution. The probability distribution in the proposed scheme needs to be optimized, and we show that the optimal distribution in fact has a particular simple structure. To identify this optimal solution, we first provide a reduced scheme that is shown to yield a lower bound to the tradeoff achieved by the proposed scheme, then show that this optimal solution to the reduced scheme is also a solution in the proposed scheme. The optimal solution in the reduced system can be obtained by analyzing the Karush-Kuhn-Tucker (KKT) conditions, which yields the given optimal solution in the proposed scheme. It should be noted that the proposed scheme in fact includes the recently proposed schemes in [11]–[13] as special cases.

The rest of the paper is organized as follows. In Section II, we first fix the notations, and then review the maximal leakage metric and the TSC scheme. The proposed scheme and its performance are given in Section III. Section IV is devoted to establishing the performance of the proposed system by a combination of bounding and analysis of the optimization problem. We conclude the paper in Section V.

II. Preliminaries

A. System Setup

There are a total of K mutually independent messages $W_{0:K-1} = (W_0, W_1, \cdots, W_{K-1})$, each of which is uniformly

distributed over the same finite alphabet. Each message is stored in N non-colluding databases. The user wishes to retrieve a single message by querying the databases. The index of the requested message can be viewed as a random variable M, which follows an unknown probability distribution. From here on, for positive integers n_1, n_2 where $n_1 \leq n_2$, we denote $[n_1 : n_2] \triangleq \{n_1, n_1+1, \cdots, n_2\}$.

A random key $\mathbf{F}$, independent of everything else, is used to generate the queries $Q_{0:N-1}^{[k]} = \left(Q_0^{[k]}, Q_1^{[k]}, \cdots, Q_{N-1}^{[k]}\right)$ when $M = k$, i.e.,

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

where $Q_n^{[k]} \in \mathcal{Q}_n$ is the query sent to the n-th database when message W_k is requested. Upon receiving $Q_n^{[k]}$, the n-th database generates an answer $A_n^{[k]}$ as a deterministic function of the query $Q_n^{[k]}$ and the stored messages $W_{0:K-1}$, i.e.,

$$H(A_n^{[k]}|Q_n^{[k]}, W_{0:K-1}) = 0, \quad \forall k \in [0:K-1], n \in [0:N-1]. \tag{2}$$

Using all the answers $A_{0:N-1}^{[k]} = \left(A_0^{[k]}, A_1^{[k]}, \cdots, A_{N-1}^{[k]}\right)$ from the N databases, as well as $\mathbf{F}$ and k, the user must be able to decode the desired message W_k, i.e.,

$$H(W_k|A_{0:N-1}^{[k]}, \mathbf{F}) = 0. \tag{3}$$

The *informational* normalized download cost of this information retrieval can be defined as

$$D \triangleq \max_{k\in[0:K-1]} \frac{\sum_{n=0}^{N-1} H(A_n^{[k]}|\mathbf{F})}{H(W_k)}. \tag{4}$$

We can also define the *operational* download cost as the downloaded number of bits. The details of these two definitions and their differences can be found in [15].

B. The Maximal Leakage Metric

We consider the problem of *weakly-private information retrieval* (WPIR) where the identity of the requested message may not be kept perfectly private, i.e., the user wishes to control the amount of knowledge that a database can deduce from the queries. The privacy leaked on M needs to be measured by a meaningful metric, and we adopt the maximal leakage metric in this work.

The maximal leakage from a random variable X to another random variable Y has an operational meaning as follows. When guessing a function of X upon observing Y, the leakage is the logarithm of the ratio of the probability of a correct guess when Y is observed, to the probability of a correct guess when Y is not observed. The maximal leakage metric $\mathcal{L}(X \to Y)$ is then defined as the maximum leakage over all such functions. It was shown in [14] that

$$\mathcal{L}(X \to Y) = \log \sum_{y\in\mathcal{Y}} \max_{\substack{x\in\mathcal{X}:\\ P_X(x)>0}} P_{Y|X}(y|x). \tag{5}$$

For the WPIR system, we shall measure the leakage from M to $Q_n^{[M]}$ using this metric, which then by (5) reduces to

$$\begin{aligned}\mathcal{L}(M \to Q_n^{[M]}) &= \log \sum_{q\in\mathcal{Q}_n} \max_{\substack{k\in[0:K-1]:\\ P_M(k)>0}} P_{Q_n^{[M]}|M}(q|M=k) \\ &= \log \sum_{q\in\mathcal{Q}_n} \max_{k\in[0:K-1]} P_{Q_n^{[k]}}(q), \end{aligned} \tag{6}$$

assuming all messages will be requested with nonzero probabilities. The worst-case maximal leakage among the databases is then

$$\rho \triangleq \max_{n\in[0:N-1]} \mathcal{L}(M \to Q_n^{[M]}). \tag{7}$$

It is clear that there is a tradeoff between ρ and D, which is the focus in this work.

We note that the perfect privacy requirement in the canonical PIR problem stipulates the queries to be identically distributed

$$Q_n^{[k]} \sim Q_n^{[k']}, \ \forall k, k' \in [0:K-1], \ \forall n \in [0:N-1], \tag{8}$$

which implies the (independence) relation

$$\mathcal{L}(M \to Q_n^{[M]}) = 0, \ \forall n \in [0:N-1]. \tag{9}$$

C. The TSC Scheme

The TSC scheme was proposed in [2] for the canonical PIR system to achieve perfect privacy, which we briefly review here. Each message W_k consists of $L = N-1$ binary symbols, and thus the message W_k can be denoted as $W_k = (W_{k,0}, W_{k,1}, \cdots, W_{k,N-1})$, where $W_{k,0} = 0$ is a prepending dummy element. Let the random key $\mathbf{F} = (F_0, F_1, \cdots, F_{K-2})$ be uniformly distributed in the set $\mathcal{F} \triangleq [0:N-1]^{K-1}$, i.e.,

$$P(\mathbf{F} = \boldsymbol{f}) = N^{-(K-1)}, \ \forall \boldsymbol{f} \in \mathcal{F}. \tag{10}$$

The query $\tilde{Q}_n^{[k]}$ is then generated as

$$\tilde{Q}_n^{[k]} = \left(F_0, \cdots, F_{k-1}, \left(n - \sum_{i=0}^{K-2} F_i\right)_N, F_k, \cdots, F_{K-2}\right), \tag{11}$$

where $(\cdot)_N$ denotes the module N operation. Since each query is a length-K vector, we can denote it by $\tilde{Q}_{n,0:K-1}^{[k]}$. Upon receiving the query, the answer $\tilde{A}_n^{[k]}$ can be generated as

$$\tilde{A}_n^{[k]} = W_{0,\tilde{Q}_{n,0}^{[k]}} \oplus W_{1,\tilde{Q}_{n,1}^{[k]}} \oplus \cdots W_{K-1,\tilde{Q}_{n,K-1}^{[k]}} \tag{12}$$

$$= W_{k,\left(n-\sum_{i=1}^{K-1} F_i\right)_N} \oplus \mathscr{I}, \tag{13}$$

where $\oplus$ is addition in the binary field and

$$\begin{aligned}\mathscr{I} \triangleq\ & W_{0,F_0} \oplus \cdots \oplus W_{k-1,F_{k-1}} \\ & \oplus W_{k+1,F_k} \oplus \cdots W_{K-1,F_{K-2}}\end{aligned} \tag{14}$$

is the interference signal. Since the answers $\tilde{A}_{1:N}^{[k]}$ only differ at the element of message W_k and $W_{k,0} = 0$, the user can retrieve W_k by recovering each element $W_{k,i}$ for all $i \in [1:N-1]$. The scheme is private because the uniform distribution of $\mathbf{F}$ on $\mathcal{F}$ induces a uniform query distribution regardless of the requested message.

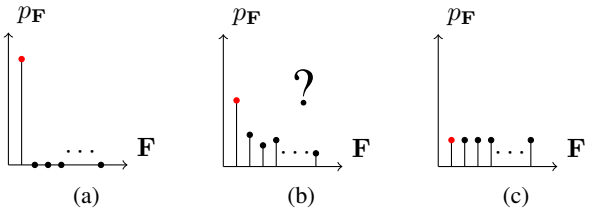

Fig. 1: Between the extreme case of (a) where download cost is minimized and the other extreme case of (c) where the privacy is prefect, we can adjust the probability distribution to achieve the tradeoff between ρ and D.

III. The Proposed Weakly-Private Information Retrieval Scheme

A. Motivation

The TSC scheme uses a uniformly distributed random key $\mathbf{F}$, which leads to $\rho = 0$. To reduce the download cost and allow a certain amount of privacy leakage, a natural generalization is to let the random key $\mathbf{F}$ take other distributions. Particularly, in the extreme case of perfect privacy, a uniform distribution is required, and for the other extreme of minimum download, we can directly download the requested message, i.e., $P(\mathbf{F} = \mathbf{0}) = 1$. Our question is thus how to adjust the probability distribution in the intermediate regime to achieve a meaningful tradeoff; see Fig. 1 for an illustration.

B. The Weakly-Private TSC (WP-TSC) Scheme

Intuitively speaking, in addition to allowing more general probability distributions on the queries, we also wish to allow a permutation of the query strategy across the databases, such that the unbalance among the databases in the TSC scheme does not adversely affect the performance. For this purpose, more notations need to be introduced first. Let $\mathcal{P}$ be the set of all permutations of $[0 : N-1]$. For each $k \in [0 : K-1]$, let Π_k be a random variable in $\mathcal{P}$, with a distribution $P(\Pi_k = \pi) = w_\pi^k$. Additionally, for any $\pi \in \mathcal{P}$ and $k \in [0 : K-1]$, let $F^{k,\pi} \triangleq (F_0^{k,\pi}, F_1^{k,\pi}, \cdots, F_{K-2}^{k,\pi})$ be a random vector distributed as follows

$$P(F^{k,\pi} = \boldsymbol{f}) = p_{\boldsymbol{f}}^{k,\pi}, \ \forall \ \boldsymbol{f} \in \mathcal{F}. \tag{15}$$

The random key $\mathbf{F}$ in the WP-TSC scheme is

$$\mathbf{F} \triangleq \Big\{ \{F^{k,\pi}, \pi \in \mathcal{P}, k \in [0 : K-1]\}, \{\Pi_k, k \in [0 : K-1]\} \Big\}, \tag{16}$$

i.e., $\mathbf{F}$ has a total of $N!K + K$ components. It is clear that $\mathbf{F}$ is indeed independent of M and the messages.

Similar to (11), the query $Q_n^{[k]}$ can be now generated using F^{k,Π_k} directly,

$$Q_n^{[k]} = \left(F_0^{k,\Pi_k}, \cdots, F_{k-1}^{k,\Pi_k}, \left(\Pi_k(n) - \sum_{i=1}^{K-1} F_i^{k,\Pi_k} \right)_N, F_k^{k,\Pi_k}, \cdots, F_{K-2}^{k,\Pi_k} \right). \tag{17}$$

The answers are also generated in the same way as in (13), with $Q_{n,i}^{[k]}$ replacing $\tilde{Q}_{n,i}^{[k]}$, i.e.,

$$A_n^{[k]} = W_{0,Q_{n,0}^{[k]}} \oplus W_{1,Q_{n,1}^{[k]}} \oplus \cdots W_{K-1,Q_{n,K-1}^{[k]}}. \tag{18}$$

The random variable Π_k essentially selects a permutation to apply on the original TSC scheme, when message W_k is requested; then for this permuted TSC scheme, the set of queries is exactly the same as that in the original scheme, however, with a non-uniform distribution specified by $p_{\boldsymbol{f}}^{k,\pi}$. Following this view, the correctness of the scheme is immediate. Note that the query $Q_n^{[k]}$ can take any possible values in $[0, N-1]^K$, i.e., $\mathcal{Q}_n = [0, N-1]^K$ for any $n \in [0, N-1]$, therefore, we shall simply write it as $\mathcal{Q}$ in the sequel.

Let us illustrate the idea using an example where $K = 2$ and $N = 3$, and the two messages are $a = (a_1, a_2)$ and $b = (b_1, b_2)$, respectively. For $\pi = (0, 1, 2)$ and $\pi = (1, 0, 2)$, the retrieval strategy and the corresponding probabilities are given in Tables I and II. Observe that in Table I, DB0 plays a special role, and thus there is a natural unbalance among the databases in the TSC scheme. The permutations Π_k helps to remove such unbalance.

Denote $\|\cdot\|$ as the zero norm of a vector and let $\mathcal{F}_j = \{\boldsymbol{f} \in \mathcal{F} : \|\boldsymbol{f}\| = j\}$ for $j \in [0 : K-1]$. Denote the set of cyclic (round-robin) permutations by $\mathcal{P}_0 = \{\pi \in \mathcal{P} : \exists \ m \text{ s.t. } \pi\big([0 : N-1]\big) = (m : N-1, 0 : m-1)\}$. Denote the optimal download cost of the canonical PIR problem by D_{PIR}, which is $D_{\text{PIR}} = 1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}$.

The following theorem characterizes the optimal tradeoff between the maximal leakage ρ and the normalized download cost D of the WP-TSC scheme, as well as the optimal probability distribution that achieves this tradeoff. The proof can be found in Section IV.

Theorem 1. *The optimal tradeoff between ρ and D of the WP-TSC scheme is*

$$\rho = \log \left[1 + \frac{(N-1)(K-1)(N^K-1)}{N^K - N} - \frac{N^{K-1}(N-1)^2(K-1)}{N^K - N} D \right], \ D \in [1, D_{PIR}] \tag{19}$$

which is achieved by the following distribution

$$w_\pi^k = \begin{cases} \frac{1}{N}, & \forall \ k \in [0 : K-1], \pi \in \mathcal{P}_0 \\ 0, & \textit{otherwise} \end{cases} \tag{20}$$

and

$$p_{\boldsymbol{f}}^{k,\pi} = \begin{cases} N - (N-1)D, & \forall \ k \in [0 : K-1], \pi \in \mathcal{P}, \boldsymbol{f} \in \mathcal{F}_0 \\ \frac{1-[N-(N-1)D]}{N^{K-1}-1}, & \forall \ k \in [0 : K-1], \pi \in \mathcal{P}, \boldsymbol{f} \notin \mathcal{F}_0. \end{cases} \tag{21}$$

The optimal probability distribution can be understood as follows: 1) we only need to use the cyclic permutations of the TSC scheme uniformly at random, instead of all possible permutations among the databases, and 2) the probability of retrieving the message directly is given a higher value, while

TABLE I: Retrieval strategy when $\pi = (0, 1, 2)$

Requesting message a				Requesting message b			
prob.	DB0	DB1	DB2	prob.	DB0	DB1	DB2
$p_{(0)}^{a,012}$		a_1	a_2	$p_{(0)}^{b,012}$		b_1	b_2
$p_{(1)}^{a,012}$	$a_2 \oplus b_1$	b_1	$a_1 \oplus b_1$	$p_{(1)}^{b,012}$	$a_1 \oplus b_2$	a_1	$a_1 \oplus b_1$
$p_{(2)}^{a,012}$	$a_1 \oplus b_2$	$a_2 \oplus b_2$	b_2	$p_{(2)}^{b,012}$	$a_2 \oplus b_1$	$a_2 \oplus b_2$	a_2

TABLE II: Retrieval strategy when $\pi = (1, 0, 2)$

Requesting message a				Requesting message b			
prob.	DB0	DB1	DB2	prob.	DB0	DB1	DB2
$p_{(0)}^{a,102}$	a_1		a_2	$p_{(0)}^{b,102}$	b_1		b_2
$p_{(1)}^{a,102}$	b_1	$a_2 \oplus b_1$	$a_1 \oplus b_1$	$p_{(1)}^{b,102}$	a_1	$a_1 \oplus b_2$	$a_1 \oplus b_1$
$p_{(2)}^{a,102}$	$a_2 \oplus b_2$	$a_1 \oplus b_2$	b_2	$p_{(2)}^{b,102}$	$a_2 \oplus b_2$	$a_2 \oplus b_1$	a_2

all the other possible query combinations in the TSC scheme are given the same probability.

C. Discussions

The WPIR problem was also studied in [11]–[13], where only the case $N = 2$ was considered. Though they used different leakage metrics, the schemes proposed in [11]–[13] are very similar, which can be viewed as related to the TSC scheme in different ways. The scheme in [11] is related by reassigning non-uniform probabilities to all the possible queries in the TSC scheme for each database when $N = 2$. The proposed schemes in both [12] and [13] can be viewed as a round-robin of the TSC scheme when $N = 2$. The WP-TSC scheme proposed here allows arbitrary probability distributions on the queries and also a permutation of the query strategy across the databases. Thus, the WP-TSC scheme includes all the existing schemes in [11]–[13] as special cases.

The random key $\mathbf{F}$ given in (16) is rather complicated, and it is trivial to see that it is independent of everything else. However, it is in fact not necessary to require this amount of randomness in the protocol. Essentially, the randomness in $\mathbf{F}$ needs to be sufficiently sophisticated to produce the random variable Π_k, and for each (k, π) also produce the required distribution specified by $p_{\boldsymbol{f}}^{k,\pi}$.

IV. Optimizing the WP-TSC Scheme

The normalized download cost of the WP-TSC scheme can be computed as

$$D = \max_{k\in[0:K-1]} \sum_{\pi\in\Pi} w_\pi^k \frac{(N-1)p_{\mathbf{0}}^{k,\pi} + N(1-p_{\mathbf{0}}^{k,\pi})}{N-1} = \frac{N - \min_{k\in[0:K-1]} \sum_{\pi\in\mathcal{P}} w_\pi^k p_{\mathbf{0}}^{k,\pi}}{N-1}. \tag{22}$$

For $q \in \mathcal{Q}$, denote $q|k \triangleq (q_0, \cdots, q_{k-1}, q_{k+1}, \cdots, q_{N-1})$. We can analyze the maximal leakage of the WP-TSC scheme, which is

$$\rho\left(\{w_\pi^k\}, \{p_{\boldsymbol{f}}^{k,\pi}\}\right) \triangleq \max_{n\in[0:N-1]} \{\mathcal{L}(M \to Q_n^{[M]})\} \tag{23}$$

$$= \log \max_{n\in[0:N-1]} \left[\sum_{q\in\mathcal{Q}} \max_{k\in[0:K-1]} P_{Q_n^{[k]}}(q)\right] \tag{24}$$

$$= \log \max_{n\in[0:N-1]} \left[\sum_{q\in\mathcal{Q}} \max_{k\in[0:K-1]} \sum_{\pi:\pi(n)=\left(\sum_i q_i\right)_N} w_\pi^k \cdot p_{q|k}^{k,\pi}\right]. \tag{25}$$

For any $q \in \mathcal{Q}$ and $k \in [0 : K-1]$, we have $q|k \in \mathcal{F}$, and thus $\rho\left(\{w_\pi^k\}, \{p_{\boldsymbol{f}}^{k,\pi}\}\right)$ is indeed a function of $\{w_\pi^k\}$ and $\{p_{\boldsymbol{f}}^{k,\pi}\}$.

Under the constraint of download cost $D \le D^*$ ($1 \le D^* \le D_{\text{PIR}}$), the problem of minimizing the leakage using the proposed scheme can then be written as the following optimization problem $P1$:

$$\text{minimize: } \rho\left(\{w_\pi^k\}, \{p_{\boldsymbol{f}}^{k,\pi}\}\right) \tag{26}$$

$$\text{subject to: } w_\pi^k \ge 0,\ \forall k \in [0:K-1], \pi \in \mathcal{P} \tag{27}$$

$$p_{\boldsymbol{f}}^{k,\pi} \ge 0,\ \forall k \in [0:K-1], \pi \in \mathcal{P}, \boldsymbol{f} \in \mathcal{F} \tag{28}$$

$$\sum_{\pi\in\Pi} w_\pi^k = 1,\ \forall k \in [0:K-1] \tag{29}$$

$$\sum_{\boldsymbol{f}\in\mathcal{F}} p_{\boldsymbol{f}}^{k,\pi} = 1,\ \forall k \in [0:K-1], \pi \in \mathcal{P} \tag{30}$$

$$\frac{N - \min_{k\in[0:K-1]} \sum_{\pi\in\mathcal{P}} w_\pi^k p_{\mathbf{0}}^{k,\pi}}{N-1} \le D^*. \tag{31}$$

Problem $P1$ in this form is not linear or convex, and it is not easy to solve. In order to solve $P1$, we consider a reduced scheme in Section IV-A, which belongs to the WP-TSC scheme but has a simple representation. The performance of the reduced scheme can be written as an optimization problem $P2$, and we show in Section IV-B that the optimal value of $P1$ is essentially equal to the optimal value of $P2$ which is solved in Section IV-C.

A. The Reduced WP-TSC Scheme

The random strategy $\mathbf{F}$ is chosen from $\mathcal{F}$ with the following distribution

$$P(\mathbf{F}=\boldsymbol{f})=p_j,\ \forall \boldsymbol{f}\in\mathcal{F}_j, j\in[0:K-1]. \tag{32}$$

The queries are generated by applying the original TSC strategy in (11) in a cyclic manner uniformly at random, i.e.,

$$P\left(Q^{[k]}_{0:N-1}=\tilde{Q}^{[k]}_{m:N-1,0:m-1}\right)=\frac{1}{N},\ \forall m\in[0,N-1]. \tag{33}$$

The answers are generated in the same way as in (13) with $Q^{[k]}_{n,i}$ replacing $\tilde{Q}^{[k]}_{n,i}$, i.e.,

$$A^{[k]}_n=W_{0,Q^{[k]}_{n,0}}\oplus W_{1,Q^{[k]}_{n,1}}\oplus\cdots W_{K-1,Q^{[k]}_{n,K-1}}. \tag{34}$$

The scheme described above can be viewed as a reduced version of the WP-TSC scheme by setting the permutation weights following (20) and the random key distribution obeying (32). Note that the query $Q^{[k]}_n$ can take any possible values in $\mathcal{Q}$. Denote $t_j\triangleq|\{q\in\mathcal{Q}:\|q\|=j\}|$, which is obtained as

$$t_j=\begin{pmatrix}K\\ j\end{pmatrix}(N-1)^j,\forall\ j\in[0:K]. \tag{35}$$

For notational simplicity, let $p_{-1}=p_K=0$. The download cost and maximal leakage of the reduced WP-TSC scheme is given in the following lemma, whose proof can be found in [16].

Lemma 1. *The reduced WP-TSC scheme achieves the download cost and maximal leakage pair (D,ρ) such that*

$$D=\frac{N-p_0}{N-1}, \tag{36}$$

$$\rho=\log\frac{1}{N}\left(\sum_{j=0}^{K}t_j\max\{p_{j-1},p_j\}\right), \tag{37}$$

for $p_0\in\left[\frac{1}{N^{K-1}},1\right]$.

B. The Reduced Optimization Problem

We denote $s_j\triangleq|\mathcal{F}_j|$, which is

$$s_j=\begin{pmatrix}K-1\\ j\end{pmatrix}(N-1)^j,\forall\ j\in[0:K-1]. \tag{38}$$

For a given download cost constraint $D\le D^*$ $(1\le D^*\le D_{\text{PIR}})$, by Lemma 1, the problem of minimizing the leakage in the reduced scheme can be written as the following optimization problem $P2$:

$$\text{minimize:}\quad \log\frac{1}{N}\left(\sum_{j=0}^{K}t_j\max\{p_{j-1},p_j\}\right) \tag{39}$$

$$\text{subject to:}\quad p_j\ge 0,\ \forall\ j\in[0:K-1] \tag{40}$$

$$\sum_{j=0}^{K-1}s_jp_j=1 \tag{41}$$

$$\frac{N-p_0}{N-1}\le D^*. \tag{42}$$

Denote the optimal value of $P1$ as $(P1)$, and similarly for $P2$. Then we have the following lemma.

Lemma 2. *Under the same constraint D^* where $1\le D^*\le D_{PIR}$, we have $(P1)=(P2)$.*

Proof. (Outline) On the one hand, for any valid parameters $\{w^k_\pi,p^{k,\pi}_{\boldsymbol{f}}\}$ for problem $P1$, we assign p_j in problem $P2$ as

$$p_j=\frac{1}{Ks_j}\sum_{k=0}^{K-1}\sum_{\pi\in\mathcal{P}}\sum_{\boldsymbol{f}\in\mathcal{F}_j}w^k_\pi\cdot p^{k,\pi}_{\boldsymbol{f}}. \tag{43}$$

With relation (43), it can be shown that $\{p_j\}$ satisfies the constraints (40)-(42), and the objective function in problem $P1$ is lower bounded by that in $P2$, for which the detailed proof is given in [16]. Thus, $(P1)\ge(P2)$.

On the other hand, since the reduced scheme is a special case of the WP-TSC scheme, we have $(P1)\le(P2)$. Combining the other inequality, we arrive at $(P1)=(P2)$. □

C. The Optimal Solution and Optimal Value

The problem (P2) is not yet linear, however since $\log(\cdot)$ function is monotonically increasing, we can instead minimize the function inside the logarithm, which becomes a linear program. The following lemma gives an optimal solution to $P2$, whose proof can be found in [16].

Lemma 3. *An optimal solution to the optimization problem $P2$ defined in* (39)-(42) *is given as*

$$p_0=N-(N-1)D^*, \tag{44}$$

$$p_j=\frac{1-p_0}{N^{K-1}-1},\ \textit{for}\ j\in[1:K-1]. \tag{45}$$

By Lemma 3, the optimal maximal leakage of the reduced scheme can then be calculated as

$$\rho=\log\frac{1}{N}\left[\frac{N^K-K(N-1)-1}{N^{K-1}-1}+\frac{N^{K-1}(N-1)(K-1)}{N^{K-1}-1}p_0\right]. \tag{46}$$

With the constraint $D\le D^*$, we have

$$p_0\ge N-(N-1)D^*. \tag{47}$$

Then Theorem 1 is proved by substituting (47) into (46).

V. Conclusion

In this paper, we studied the WPIR problem under the maximal leakage metric. We propose a new coding scheme where the probability distribution can be optimized to control the privacy leakage. The optimal probability distribution turns out to have a very simple structure, where only the retrieval combination of directly obtaining the message itself is given a higher probability, and all other combinations are given the same probability. We prove the optimality of this solution through a combination of lower bounding and analysis of the KKT conditions.

REFERENCES

[1] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.

[2] 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.

[3] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[4] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.

[5] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

[6] Q. Wang, H. Sun, and M. Skoglund, "The capacity of private information retrieval with eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3198–3214, May 2019.

[7] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 2852–2856.

[8] H. Yang, W. Shin, and J. Lee, "Private information retrieval for secure distributed storage systems," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 12, pp. 2953–2964, Dec. 2018.

[9] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," *arXiv*, Apr. 2019, (full version). Available: http://arxiv.org/abs/1903.08229. [Online]. Available: http://arxiv.org/abs/1903.08229

[10] 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. [Online]. Available: https://arxiv.org/abs/1909.11605

[11] H. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "Weakly-private information retrieval," in *2019 IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 1257–1261.

[12] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *2019 IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 1262–1266.

[13] Z. Jia, "On the capacity of weakly-private information retrieval," Master's thesis, Department of Electrical Engineering, University of California, Irvine, CA, 2019.

[14] I. Issa, S. Kamath, and A. B. Wagner, "An operational measure of information leakage," in *2016 Annual Conference on Information Science and Systems (CISS)*, Princeton, NJ, Mar. 2016, pp. 234–239.

[15] C. Tian, "On the storage cost of private information retrieval," 2019. [Online]. Available: https://arxiv.org/pdf/1910.11973.pdf

[16] R. Zhou, T. Guo, and C. Tian, "Weakly private information retrieval under the maximal leakage metric." [Online]. Available: https://tiangroup.engr.tamu.edu/wp-content/uploads/sites/150/2020/01/weakprivacylong.pdf