Weakly-Private Information Retrieval from MDS-Coded Distributed Storage

Weakly-Private Information Retrieval From MDS-Coded Distributed Storage

Asbjørn O. Orvedal, Hsuan-Yin Lin, and Eirik Rosnes
Simula UiB, N–5006 Bergen, Norway
Emails: asbjorn.orvedal@gmail.com, {lin, eirikrosnes}@simula.no

***Abstract*—We consider the problem of weakly-private information retrieval (WPIR) when data is encoded by a maximum distance separable code and stored across multiple servers. In WPIR, a user wishes to retrieve a piece of data from a set of servers without leaking too much information about which piece of data she is interested in. We study and provide the first WPIR protocols for this scenario and present results on their optimal trade-off between download rate and information leakage using the maximal leakage privacy metric.**

I. Introduction

Private information retrieval (PIR), introduced in a seminal paper by Chor *et al.* [1], [2], has been extensively studied for more than two decades in both the computer science and information theory communities, see, e.g., [3]–[8] and references therein. In PIR, the objective is to download a piece of data stored on a set of servers without leaking any information about which piece of data is being requested to the servers storing the data, while minimizing the overall communication cost. As the upload cost is typically much lower than the download cost, the download rate, defined as the ratio between the amount of requested information and the amount of downloaded information, is used as a measure to compare different PIR protocols. When data is replicated across several servers, the maximum achievable download rate, referred to as the PIR capacity, was derived in [9], while the capacity for the case where the data is encoded by a maximum distance separable (MDS) code and stored across a set of servers was settled in [10]. Arbitrary linear storage codes were considered in [11], [12].

Weakly-private information retrieval (WPIR), introduced independently by Lin *et al.* [13] and Samy *et al.* [14], is a relaxed version of PIR that allows for reducing the download cost at the expense of some information leakage on the identity of the requested piece of data to the servers storing it. So far, only the case of replicated data (across servers) and the single server case have been considered in the literature [15]–[21], while in this work we consider for the first time the case where the data is encoded by an MDS code and stored across multiple servers. WPIR protocols allow for a trade-off between download rate and privacy leakage, and the optimal trade-off curve for the case of multiple servers is still an open problem. As in previous works, we consider the maximal leakage (MaxL) privacy metric [22]–[24]. Our main contributions are as follows.

- We adapt the PIR protocols in [25], [26] for MDS-coded databases to allow for information leakage. The adapted protocols from [25], [26], referred to as the ZYQT and ZTSL MDS-WPIR schemes, respectively, yield a trade-off between download rate and information leakage, and we show that for the MaxL privacy metric the optimal trade-off is the solution of a convex optimization problem (see Theorem 1). The optimized ZYQT MDS-WPIR scheme yields the best trade-off but also has the largest query space.
- We propose a *new* WPIR protocol, referred to as the OLR MDS-WPIR scheme, with a much smaller query space than the ZYQT scheme while providing an equally good or better trade-off between download rate and information leakage. As for the ZYQT and ZTSL MDS-WPIR schemes, the optimal trade-off is the solution of a convex optimization problem (see Theorem 1).

II. Preliminaries and System Model

A. Notation

We denote by $\mathbb{N}$ the set of all positive integers, and $[a:b] \triangleq \{a, a+1, \ldots, b\}$ for $a, b \in \{0\} \cup \mathbb{N}$, $a \leq b$. Vectors (normally row-wise) are denoted by bold letters, random variables (RVs) (either scalar or vector) by uppercase letters, and sets by calligraphic uppercase letters, e.g., $\boldsymbol{x}$, X, and $\mathcal{X}$, respectively. Matrices are denoted by sans serif letters, while random matrices are represented by bold sans serif capital letters, e.g., $\mathbf{X}$, and x represents its realization. The all-one (all-zero) row vector is denoted by $\mathbf{1}$ ($\mathbf{0}$), and its length will be clear from the context. When a set of indices $\mathcal{S}$ is given, $\boldsymbol{X}_{\mathcal{S}}$ denotes $\{\boldsymbol{X}_s \colon s \in \mathcal{S}\}$. $\mathsf{E}_X[\cdot]$ denotes expectation with respect to the RV X. $X \sim P_X$ denotes an RV distributed according to a probability mass function (PMF) $P_X(x)$, $x \in \mathcal{X}$, and $X \sim \mathrm{U}(\mathcal{S})$ a uniformly-distributed RV over a set $\mathcal{S}$. $\mathsf{H}(\cdot)$ denotes the entropy function, $(\cdot)^\mathsf{T}$ the transpose of a matrix, and $\gcd(a, b)$ the greatest common divisor of two positive integers a and b.

B. System Model

We consider an MDS-coded distributed storage system (DSS) with N noncolluding servers that store M independent files $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(\mathsf{M})}$, where each file is represented as a random matrix $\mathbf{W}^{(m)} = \left(W_{i,j}^{(m)}\right)$ of size $\lambda \times \mathsf{K}$, $\lambda, \mathsf{K} \in \mathbb{N}$. Each file $\mathbf{W}^{(m)}$ is encoded row-wise using an $[\mathsf{N}, \mathsf{K}]$ MDS code $\mathcal{C}$ over some finite field $\mathbb{F}_q$ of size $q \geq \mathsf{N}$ resulting in

the codewords $\left(X_{i,1}^{(m)},\dots,X_{i,\mathsf{N}}^{(m)}\right) = (W_{i,1}^{(m)},\dots,W_{i,\mathsf{K}}^{(m)})\mathsf{G}^{\mathcal{C}}$, $i \in [0:\lambda-1]$, where $\mathsf{G}^{\mathcal{C}}$ denotes a generator matrix for $\mathcal{C}$. Denote by $\boldsymbol{X}_j^{(m)} \triangleq \left(X_{0,j}^{(m)},\dots,X_{\lambda-1,j}^{(m)}\right)^{\intercal}$ a vector consisting of λ code symbols generated by the code $\mathcal{C}$. Then, the j-th server stores $\boldsymbol{X}_j \triangleq \left((\boldsymbol{X}_j^{(1)})^{\intercal}|\cdots|(\boldsymbol{X}_j^{(\mathsf{M})})^{\intercal}\right)^{\intercal}$, $j \in [1:\mathsf{N}]$.

To retrieve a file $\mathbf{W}^{(M)}$, from the MDS-coded DSS, the user sends a query $\mathbf{Q}_j$ to the j-th server for all $j \in [1:\mathsf{N}]$. Here, $M \sim \mathrm{U}([1:\mathsf{M}])$ is an RV representing the desired file index. In response to the received query, server j returns the answer $\mathbf{A}_j$, which is a function of $\mathbf{Q}_j$ and the code symbols $\boldsymbol{X}_j$ stored in the server, back to the user. We formally describe an MDS-coded $(\mathsf{M},\mathsf{N},\mathsf{K})$ WPIR scheme as follows.

Definition 1 (MDS-WPIR Scheme)**.** *An $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR scheme for an $[\mathsf{N},\mathsf{K}]$ MDS-coded DSS with N non-colluding servers consists of:*

- *M independent files $\mathbf{W}^{(m)}$ of size $\lambda\times\mathsf{K}$, for some $\lambda\in\mathbb{N}$, $m\in[1:\mathsf{M}]$.*
- *A global random strategy $\mathbf{S}$, whose alphabet is $\mathcal{S}$. In general, the realization of $\mathbf{S}$ is a matrix.*
- *An (N,K) MDS storage code $\mathcal{C}$ that encodes the file $\mathbf{W}^{(m)}$ into the matrix $\mathbf{X}^{(m)} = \left(\boldsymbol{X}_1^{(m)}|\cdots|\boldsymbol{X}_{\mathsf{N}}^{(m)}\right)$ as described above, $m\in[1:\mathsf{M}]$.*
- *N queries $\mathbf{Q}_j = \phi_j(M,\mathbf{S})$ with alphabet $\mathcal{Q}_j$, $j\in[1:\mathsf{N}]$, that are generated by the query-encoding functions ϕ_j. Query $\mathbf{Q}_j$ is sent to the j-th server.*
- *N answers $\mathbf{A}_j = \psi_j(\mathbf{Q}_j,\boldsymbol{X}_j)$ with alphabet $\mathcal{A}=\mathbb{F}_q$, $j\in[1:\mathsf{N}]$, that are constructed by the answer functions ψ_j. All answers $\mathbf{A}_j$ are sent back to the user.*
- *N answer lengths $\ell_j(\mathbf{Q}_j)\in\{0\}\cup\mathbb{N}$, $j\in[1:\mathsf{N}]$, each being a function of the corresponding query $\mathbf{Q}_j$.*

In addition, the scheme should satisfy the following condition of perfect retrievability:

$$\mathsf{H}\big(\mathbf{W}^{(M)}\,\big|\,\mathbf{A}_{[1:\mathsf{N}]},\mathbf{Q}_{[1:\mathsf{N}]},M\big)=0.$$

C. Maximal Leakage Metric

From Definition 1, one can notice that at the j-th server, the requested file index M can be inferred by observing the query distribution $P_{\mathbf{Q}_j}$, which results in an information leakage on M to the servers. In this work, we adopt a meaningful information-theoretic privacy metric from the computer science literature, the MaxL metric, to measure information leakage. Formally, given the query distributions $P_{M,\mathbf{Q}_j}$, $j\in[1:\mathsf{N}]$, of a given $(\mathsf{M},\mathsf{N},\mathsf{K})$ WPIR scheme $\mathscr{C}$, the overall MaxL about M of $\mathscr{C}$ is defined as

$$\rho^{(\mathrm{MaxL})}(\mathscr{C}) \triangleq \max_{j\in[1:\mathsf{N}]} \mathrm{MaxL}(M;\mathbf{Q}_j),$$

where

$$\mathrm{MaxL}(M;\mathbf{Q}) \triangleq \log_2\left(\sum_{\mathsf{q}\in\mathcal{Q}}\max_{m\in[\mathsf{M}]} P_{\mathbf{Q}|M}(\mathsf{q}|m)\right).$$

Note that an $[\mathsf{N},\mathsf{K}]$ MDS-coded PIR scheme is an $(\mathsf{M},\mathsf{N},\mathsf{K})$ WPIR scheme $\mathscr{C}$ that satisfies $\rho^{(\mathrm{MaxL})}(\mathscr{C})=0$, such a condition is refereed to as the *perfect privacy* constraint.

D. WPIR Download Cost and Rate

The overall download cost (in number of symbols over $\mathbb{F}_q$) and rate of a WPIR scheme $\mathscr{C}$, denoted by $\mathrm{D}(\mathscr{C})$ and $\mathsf{R}(\mathscr{C})$, respectively, are given by

$$\mathrm{D}(\mathscr{C}) = \sum_{j=1}^{\mathsf{N}} \mathsf{E}_{\mathbf{Q}_j}[\ell_j(\mathbf{Q}_j)] \text{ and } \mathsf{R}(\mathscr{C}) \triangleq \frac{\lambda\mathsf{K}}{\mathrm{D}(\mathscr{C})}.$$

III. General MDS-WPIR Schemes

In this section, we give a general description of the $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR schemes we consider in this work. We start by reviewing two MDS-PIR capacity-achieving schemes for small file sizes, namely the ZYQT scheme [25] and the ZTSL scheme [26].[1]

A. The ZYQT Scheme and the ZTSL Scheme

1) Storage Data Structure: The following *effective code parameters* are universally defined for an MDS-coded DSS:

$$n \triangleq \frac{\mathsf{N}}{\gcd(\mathsf{N},\mathsf{K})},\quad k \triangleq \frac{\mathsf{K}}{\gcd(\mathsf{N},\mathsf{K})},\quad r \triangleq n-k.$$

Moreover, the subpacketization size for each file is given by $\lambda = n-k$. For ease of exposition, we further append k dummy variables $X_{i,j}^{(m)} \equiv 0$ for $i\in[n-k:n-1]$, $j\in[1:\mathsf{N}]$, such that for all $m\in[1:\mathsf{M}]$,

$$\mathbf{X}^{(m)} = \left.\begin{pmatrix} X_{0,1}^{(m)} & X_{0,2}^{(m)} & \cdots & X_{0,\mathsf{N}}^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n-k-1,1}^{(m)} & X_{n-k-1,2}^{(m)} & \cdots & X_{n-k-1,\mathsf{N}}^{(m)} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}\right\} k \text{ rows} \tag{1}$$

2) Query Generation: The query generation is the main difference among the $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR schemes. In our context, we will make use of the set

$$\mathcal{P}_k^n \triangleq \big\{\boldsymbol{s}^{\intercal}=(s_1,\dots,s_k)^{\intercal}\colon s_i,s_{i'}\in[0:n-1],\ s_i\neq s_{i'},\ \forall i,i'\in[1:k], i\neq i'\big\}$$

of column vectors. The global random strategy alphabet for the ZYQT and ZTSL schemes are, respectively, given by

$$\mathcal{S}_{\mathrm{ZYQT}} \triangleq \{\mathsf{s}=(\boldsymbol{s}_1^{\intercal},\dots,\boldsymbol{s}_{\mathsf{M}}^{\intercal})\colon \boldsymbol{s}_{m'}^{\intercal}\in\mathcal{P}_k^n,\ m'\in[1:\mathsf{M}]\},$$

$$\mathcal{S}_{\mathrm{ZTSL}} \triangleq \left\{\boldsymbol{s}\in[0:n-1]^{\mathsf{M}}\colon \left(\sum_{m'=1}^{\mathsf{M}} s_{m'}\right) \bmod n = 0\right\}.$$

Note that $|\mathcal{S}_{\mathrm{ZYQT}}| = \left(\binom{n}{k}k!\right)^{\mathsf{M}}$ and $|\mathcal{S}_{\mathrm{ZTSL}}| = n^{\mathsf{M}-1}$. Since the cost of uploading the queries for an MDS-PIR scheme depends on the cardinality of the global random strategy alphabet, it is apparent that the ZTSL scheme has a lower upload cost than the ZYQT scheme. It is also worth mentioning that MDS-PIR schemes are generally constructed using an $\mathbf{S}$ that is uniformly distributed over the set $\mathcal{S}$.

[1] Precisely, the ZTSL scheme we consider here is the so-called Construction-A MDS-PIR code that is referred in [26, Sec. III].

We next present the original query generation for the ZYQT and ZTSL MDS-PIR schemes for retrieving the m-th file $\mathbf{X}^{(m)}$, $m \in [1:\mathsf{M}]$. Notice that we do not adopt the uniformly-distributed $\mathbf{S}$ here. Thus, the leakage $\rho^{(\text{MaxL})}$ is not necessarily equal to 0. We refer to the corresponding proposed schemes as the ZYQT MDS-WPIR and ZTSL MDS-WPIR schemes and denote them by $\mathscr{C}_{\text{ZYQT}}$ and $\mathscr{C}_{\text{ZTSL}}$, respectively.

$\mathscr{C}_{\text{ZYQT}}$: The query $\mathsf{q}_j \in \mathcal{Q}_j$, $j \in [1:\mathsf{N}]$, generated from the query-encoding function ϕ_j is defined as

$$\mathsf{q}_j = (\boldsymbol{s}_1^{\mathsf{T}}, \ldots, \boldsymbol{s}_{m-1}^{\mathsf{T}}, \big(\boldsymbol{s}_m^{\mathsf{T}} + (j-1)\mathbf{1}^{\mathsf{T}}\big) \bmod n, \\ \boldsymbol{s}_{m+1}^{\mathsf{T}}, \ldots, \boldsymbol{s}_{\mathsf{M}}^{\mathsf{T}}), \quad \boldsymbol{s}_m^{\mathsf{T}} \in \mathcal{P}_k^n,\ m \in [1:\mathsf{M}].$$

$\mathscr{C}_{\text{ZTSL}}$: The query $\mathsf{q}_j \in \mathcal{Q}_j$, $j \in [1:\mathsf{N}]$, is generated by

$$\mathsf{q}_j = \left[\underbrace{\begin{pmatrix} s_1 \cdots s_{m-1} \, (s_m + (j-1)) \, s_{m+1} \cdots s_{\mathsf{M}} \\ \vdots \quad \ddots \quad \vdots \quad\quad \vdots \quad \ddots \quad \vdots \\ s_1 \cdots s_{m-1} \, (s_m + (j-1)) \, s_{m+1} \cdots s_{\mathsf{M}} \end{pmatrix}}_{k \text{ rows}} + \underbrace{\begin{pmatrix} 0 & 0 \cdots \cdots 0 \\ 1 & 1 \cdots \cdots 1 \\ \vdots & \vdots \quad\quad \ddots \quad\quad \vdots \\ k-1 & k-1 \cdots k-1 \end{pmatrix}}_{\mathsf{M} \text{ columns}} \right] \bmod n,$$

where $\boldsymbol{s} \in \mathcal{S}_{\text{ZTSL}}$.

3) Answer Construction: Upon receiving a query (matrix)

$$\mathsf{q}_j = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,\mathsf{M}} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k,1} & q_{k,2} & \cdots & q_{k,\mathsf{M}} \end{pmatrix},$$

the j-th server uses the answer function ψ_j to construct the answer

$$\mathbf{A}_j = \psi_j(\mathsf{q}_j, \boldsymbol{X}_j) = \left(\sum_{m'=1}^{\mathsf{M}} X^{(m')}_{q_{1,m'},j}, \cdots, \sum_{m'=1}^{\mathsf{M}} X^{(m')}_{q_{k,m'},j} \right)^{\mathsf{T}}$$

consisting of k sub-responses. With the storage data defined in (1), the length of the answer is given by the number of nonzero components in $\mathbf{A}_j$, which is equal to

$$\ell_j(\mathsf{q}_j) = \sum_{i=1}^{k} \mathbb{1}\Big\{ \min_{m' \in [1:\mathsf{M}]} q_{i,m'} \le n-k-1 \Big\},$$

where $\mathbb{1}\{\text{statement}\}$ is the indicator function whose value is 1 if the statement is true and 0 otherwise.

Finally, we remark that according to the query constructions for both the ZYQT and ZTSL MDS-WPIR schemes, the file $\mathbf{W}^{(m)}$ can always be reconstructed by the MDS property of the storage code $\mathcal{C}$ (the so-called K-out-of-N property).

B. Time-Sharing MDS-WPIR Scheme

Clearly, selecting a different global random strategy $\mathbf{S}$ leads to a different WPIR rate and privacy leakage of an MDS-WPIR scheme. This work aims to achieve the best trade-off between download rate and privacy leakage by using the best $\mathbf{S}$ for an MDS-WPIR scheme. However, the minimization problem of the information leakage for a given WPIR rate over the global random strategy for an MDS-WPIR scheme is generally not convex. Hence, in order to easily tackle the optimization problem, we make use of a time-sharing principle to *convexify* the optimization problem for determining the best rate-leakage trade-off [16, Sec. VII].

Definition 2 (Time-Sharing MDS-WPIR Scheme)**.** *Consider an MDS-WPIR scheme $\mathring{\mathscr{C}}$ with query-encoding functions $\mathring{\phi}_j$, answer functions $\mathring{\psi}_j$, and a global random strategy $\mathring{\mathbf{S}}$. The time-sharing MDS-WPIR scheme of $\mathring{\mathscr{C}}$ is made by the query-encoding functions $\phi_j = \mathring{\phi}_{\sigma^{T-1}(j)}(M, \mathbf{S})$ and the answer functions $\psi_j = \mathring{\psi}_{\sigma^{T-1}(j)}\big(\mathring{\phi}_{\sigma^{T-1}(j)}(M, \mathbf{S}), \boldsymbol{X}_j\big)$, $j \in [1:\mathsf{N}]$, for a given requested file index M, where $T \sim \mathrm{U}([1:\mathsf{N}])$, and $\sigma(\cdot)$ denotes a left circular shift, while l left circular shifts are obtained through function composition and denoted by $\sigma^l(\cdot)$. Such an MDS-WPIR scheme $\mathscr{C}$ is called the time-sharing scheme of $\mathring{\mathscr{C}}$.*

Remark 1.

- *A time-sharing MDS-WPIR scheme always has equal information leakage at each server [16, Th. 1].*
- *In the following, unless specified otherwise, all the MDS-WPIR schemes we discuss are assumed to be already post-processed by applying the time-sharing principle, and the minimization of MaxL is also done for the time-sharing scheme of an MDS-WPIR scheme.*

C. Minimization of MaxL for MDS-WPIR Schemes

Denote by $z_{\mathsf{s}} \triangleq P_{\mathbf{S}}(\mathsf{s})$ the PMF of the random strategy $\mathbf{S}$. It can be shown that both the MaxL $\rho^{(\text{MaxL})}(\mathscr{C})$ and the WPIR download cost $\mathsf{D}(\mathscr{C})$ of a given MDS-WPIR scheme $\mathscr{C}$ can be expressed in terms of z_{s}, $\mathsf{s} \in \mathcal{S}$. Thus, the minimization of $\rho^{(\text{MaxL})}(\mathscr{C})$ under a download cost constraint $\mathsf{D}(\mathscr{C}) \le \mathsf{D}$ can be re-written in terms of the variables $\{z_{\mathsf{s}}\}_{\mathsf{s}\in\mathcal{S}}$ as the optimization problem

$$\begin{aligned} \text{minimize} \quad & \rho^{(\text{MaxL})}(\{z_{\mathsf{s}}\}_{\mathsf{s}\in\mathcal{S}}) && \text{(2a)} \\ \text{subject to} \quad & \mathsf{D}(\{z_{\mathsf{s}}\}_{\mathsf{s}\in\mathcal{S}}) \le \mathsf{D}, && \text{(2b)} \\ & \sum_{\mathsf{s}\in\mathcal{S}} z_{\mathsf{s}} = 1. && \text{(2c)} \end{aligned}$$

The following theorem can be proved using a similar argument as in [16, Sec. VII].

Theorem 1. *The optimization problem* (2) *is convex.*

All the rate-leakage trade-off curves of the MDS-WPIR schemes we study in this work are based on solving the convex optimization problem above.

IV. New Proposed MDS-WPIR Scheme

This section presents a new MDS-WPIR scheme, referred to as the OLR MDS-WPIR scheme. We first present an example illustrating the motivation for studying the new MDS-WPIR scheme in Section IV-A. In particular, we will show that the ZTSL MDS-WPIR scheme is naturally not a good scheme as it is not functional in the high-rate region when there is leakage.

A. Motivating Example: $(\mathsf{M},\mathsf{N},\mathsf{K})=(2,3,2)$

For $(\mathsf{N},\mathsf{K})=(3,2)$, we have the effective code parameters

$$n=\frac{\mathsf{N}}{\gcd(\mathsf{N},\mathsf{K})}=3,\ k=\frac{\mathsf{K}}{\gcd(\mathsf{N},\mathsf{K})}=2,\ r=n-k=1,$$

and the subpacketization size for each file is $\lambda=n-k=1$.

For the $(2,3,2)$ ZTSL MDS-WPIR scheme, we have $\mathcal{S}_{\text{ZTSL}}=\{(0,0),(1,2),(2,1)\}$, and the corresponding conditional query PMF $P_{\mathsf{Q}_j|M}(\mathsf{q}_j\mid m)$ and answer lengths are as follows:

$$P_{\mathsf{Q}_j|M}(\mathsf{q}_j\mid m)\quad \tbinom{0\,0}{1\,1}\ \tbinom{1\,2}{2\,0}\ \tbinom{2\,1}{0\,2}\ \tbinom{1\,0}{2\,1}\ \tbinom{2\,2}{0\,0}\ \tbinom{0\,1}{1\,2}\ \tbinom{2\,0}{0\,1}\ \tbinom{0\,2}{1\,0}\ \tbinom{1\,1}{2\,2}$$
$$m\begin{cases}1\\2\end{cases}\left(\begin{array}{ccccccccc}\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}\\ \frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_3}{3}&\frac{z_1}{3}&\frac{z_2}{3}\\ \hline \frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_1+z_2}{6}&\frac{z_2+z_3}{6}&\frac{z_1+z_3}{6}&\frac{z_1+z_3}{6}&\frac{z_1+z_2}{6}&\frac{z_2+z_3}{6}\\ \hline 1&1&1&1&1&1&2&2&0\end{array}\right),\tag{3}$$
$$P_{\mathsf{Q}_j}(\mathsf{q}_j)$$
$$\ell_j(\mathsf{q}_j)$$

where $z_j\triangleq\Pr(\boldsymbol{s}_j)$ for $\boldsymbol{s}_j=(j-1,(n-j+1)\bmod n)\in\mathcal{S}_{\text{ZTSL}}$, $j\in[1:n]$. A simple calculation gives

$$\mathsf{D}(\mathscr{C}_{\text{ZTSL}})=3+z_1,\quad 0\le z_1\le 1,$$

which indicates that $\mathsf{D}(\mathscr{C}_{\text{ZTSL}})$ can only range between 3 and 4, and never reaches $\mathsf{R}=\lambda\mathsf{K}/\mathsf{D}=2/\mathsf{D}=1$. Thus, the ZTSL MDS-WPIR scheme can not operate in the high-rate region.

B. New $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR Scheme

We now describe the new proposed $(\mathsf{M},\mathsf{N},\mathsf{K})$ MDS-WPIR scheme, referred to as the OLR MDS-WPIR scheme and denoted by $\mathscr{C}_{\text{OLR}}$. Here, only the query generation is presented, as its answer construction is the same as Section III-A3.

1) Query Generation: The strategy set for our new MDS-WPIR scheme is defined as

$$\mathcal{S}_{\text{OLR}}\triangleq\Bigg\{\mathsf{s}=(\boldsymbol{s}_1^{\mathsf{T}},\dots,\boldsymbol{s}_{\mathsf{M}-1}^{\mathsf{T}})\colon \boldsymbol{s}_{m'}^{\mathsf{T}}\in\mathcal{P}_k^n,\ m'\in[1:\mathsf{M}],\ \left(\sum_{m'=1}^{\mathsf{M}}\boldsymbol{s}_{m'}^{\mathsf{T}}\right)\bmod n=\mathbf{0}^{\mathsf{T}}\Bigg\}.$$

By definition, $|\mathcal{S}_{\text{OLR}}|\le\left(\binom{n}{k}k!\right)^{\mathsf{M}-1}<|\mathcal{S}_{\text{ZYQT}}|=\left(\binom{n}{k}k!\right)^{\mathsf{M}}$, as we do not include all the possible vectors $\boldsymbol{s}_{m'}^{\mathsf{T}}\in\mathcal{P}_k^n$.

The query $\mathsf{q}_j\in\mathcal{Q}_j$, $j\in[1:\mathsf{N}]$, for retrieving the m-th file, $m\in[1:\mathsf{M}]$, is defined as

$$\mathsf{q}_j=(\boldsymbol{s}_1^{\mathsf{T}},\dots,\boldsymbol{s}_{m-1}^{\mathsf{T}},\boldsymbol{q}_m^{\mathsf{T}},\boldsymbol{s}_m^{\mathsf{T}},\dots,\boldsymbol{s}_{\mathsf{M}-1}^{\mathsf{T}}),\tag{4}$$

where $(\boldsymbol{s}_1^{\mathsf{T}},\dots,\boldsymbol{s}_{\mathsf{M}-1}^{\mathsf{T}})=\mathsf{s}\in\mathcal{S}_{\text{OLR}}$ and

$$\boldsymbol{q}_m^{\mathsf{T}}\triangleq\left((j-1)\mathbf{1}^{\mathsf{T}}-\sum_{m'\in[1:\mathsf{M}-1]}\boldsymbol{s}_{m'}^{\mathsf{T}}\right)\bmod n.$$

Example 1. *Consider the same code parameters* $(\mathsf{M},\mathsf{N},\mathsf{K})=(2,3,2)$ *as in Section IV-A. We consider the strategy set*

$$\mathcal{S}_{\text{OLR}}=\{\underbrace{\tbinom{0}{1}}_{z_1},\underbrace{\tbinom{0}{2}}_{z_2},\underbrace{\tbinom{1}{0}}_{z_3},\underbrace{\tbinom{1}{2}}_{z_4},\underbrace{\tbinom{2}{0}}_{z_5},\underbrace{\tbinom{2}{1}}_{z_6}\}.$$

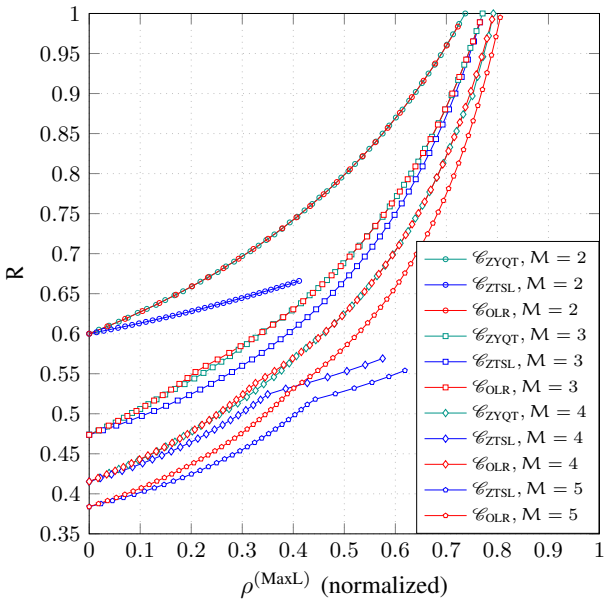

Fig. 1. Rate-leakage trade-off curve for the proposed MDS-WPIR protocols from $(3,2)$ MDS-coded storage with $\mathsf{M}=2$ (circle markers), $\mathsf{M}=3$ (square markers), $\mathsf{M}=4$ (diamond markers), and $\mathsf{M}=5$ (pentagon markers).

Similar to (3)*, we illustrate* 9 *out of the* 18 *query matrices based on* (4) *and the corresponding query distributions and answer lengths of the OLR MDS-WPIR scheme below:*

$$P_{\mathsf{Q}_j|M}(\mathsf{q}_j\mid m)\quad \tbinom{0\,0}{2\,1}\ \tbinom{0\,0}{1\,2}\ \tbinom{2\,1}{0\,0}\ \tbinom{2\,1}{1\,2}\ \tbinom{1\,2}{0\,0}\ \tbinom{1\,2}{2\,1}\ \tbinom{1\,0}{0\,1}\ \tbinom{1\,0}{2\,2}\ \tbinom{0\,1}{1\,0}$$
$$m\begin{cases}1\\2\end{cases}\left(\begin{array}{ccccccccc}\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}&\frac{z_4}{3}&\frac{z_5}{3}&\frac{z_6}{3}&\frac{z_1}{3}&\frac{z_2}{3}&\frac{z_3}{3}\\ \frac{z_2}{3}&\frac{z_1}{3}&\frac{z_5}{3}&\frac{z_6}{3}&\frac{z_3}{3}&\frac{z_4}{3}&\frac{z_3}{3}&\frac{z_4}{3}&\frac{z_1}{3}\\ \hline \frac{z_1+z_2}{3}&\frac{z_1+z_2}{6}&\frac{z_3+z_5}{3}&\frac{z_4+z_6}{6}&\frac{z_3+z_5}{6}&\frac{z_4+z_6}{6}&\frac{z_1+z_3}{6}&\frac{z_2+z_4}{6}&\frac{z_1+z_3}{6}\\ \hline 1&1&1&0&1&0&2&1&2\end{array}\right).$$
$$P_{\mathsf{Q}_j}(\mathsf{q}_j)$$
$$\ell_j(\mathsf{q}_j)$$

As a result, one can compute the download cost $\mathsf{D}(\mathscr{C}_{\text{OLR}})$ *and obtain*

$$\mathsf{D}(\mathscr{C}_{\text{OLR}})=2+2(z_1+z_2+z_3+z_5)\ge 2,$$

which shows that $\mathsf{R}(\mathscr{C}_{\text{OLR}})$ *can reach* $(n-k)\mathsf{K}/2=1$*, demonstrating a complete rate-leakage trade-off for the new MDS-WPIR scheme.*

V. Numerical Results

Here, we compare the optimal rate-leakage trade-off curves for our three proposed MDS-WPIR schemes $\mathscr{C}_{\text{ZYQT}}$, $\mathscr{C}_{\text{ZTSL}}$, and $\mathscr{C}_{\text{OLR}}$. The optimal trade-off curve is obtained by solving the corresponding convex optimization problems as outlined in (2). For the sake of presentation, the leakage is normalized by $\log_2\mathsf{M}$ bits so that its range is from 0 to 1.

In Fig. 1, we consider the case of $\mathsf{N}=3$ servers and $\mathsf{K}=2$, and with different number of files M. As can be seen from the figure by comparing the green and the blue curves, $\mathscr{C}_{\text{ZYQT}}$ gives a better rate-leakage trade-off curve than $\mathscr{C}_{\text{ZTSL}}$ for all considered values of M. Moreover, the ZTSL scheme cannot be extended to a high information leakage. On the other, the OLR scheme performs equally well as the ZYQT scheme

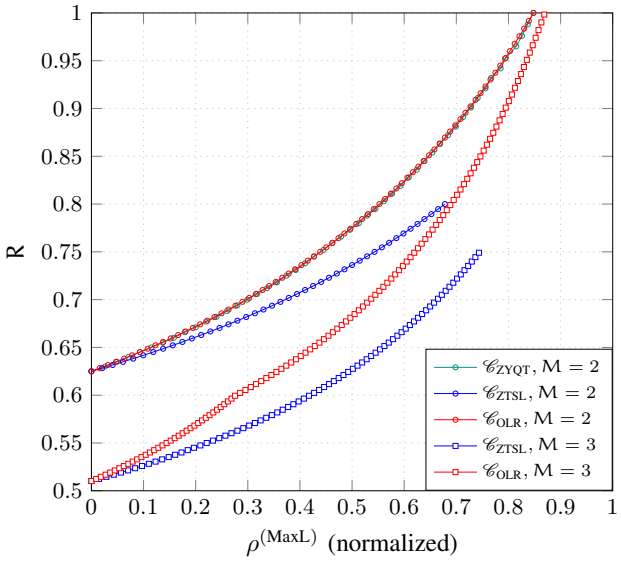

Fig. 2. Rate-leakage trade-off curve for the proposed MDS-WPIR protocols from $(5, 3)$ MDS-coded storage with $\mathrm{M} = 2$ (circle markers) and $\mathrm{M} = 3$ (square markers).

for $\mathrm{M} = 2$ files and slightly better for a certain range of information leakage for $\mathrm{M} = 3$ and $\mathrm{M} = 4$ files, while at the same time allowing for a much smaller query space.

The corresponding rate-leakage trade-off curves for $\mathrm{N} = 5$ servers with $\mathrm{K} = 3$ are provided in Fig. 2. The same observations as in Fig. 1 can be made, i.e., the ZYQT scheme outperforms the ZTSL scheme, while the proposed OLR scheme yields an equal trade-off curve as the ZYQT scheme for $\mathrm{M} = 2$ files. As the query space is significant for the ZYQT scheme for $\mathrm{M} = 3$ files, we were not able to solve the corresponding convex optimization problem as outlined in (2) and therefore no curve for $\mathrm{M} > 2$ is presented. However, as mentioned previously, a nice feature of the OLR scheme is its smaller query space, and hence the corresponding optimization problem in (2) can be readily solved even for $\mathrm{M} = 3$. In particular, we have $|\mathcal{S}_{\mathrm{ZYQT}}| = 216000 > |\mathcal{S}_{\mathrm{OLR}}| = 1500$ for $\mathrm{M} = 3$.

VI. Conclusion

This work is the first to consider WPIR for coded storage. In particular, we proposed and compared three WPIR protocols for the case where the data is encoded by an MDS code and stored across multiple servers. Allowing for some leakage on the identity of the requested file index allows for a higher download rate, and we showed that the optimal trade-off of download rate and information leakage using the MaxL privacy metric is the solution to a convex optimization problem for all three proposed protocols.

References

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. 36th Annu. IEEE Symp. Found. Comp. Sci. (FOCS)*, Milwaukee, WI, USA, Oct. 23–25, 1995, pp. 41–50.

[2] ——, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–982, Nov. 1998.

[3] A. Beimel, Y. Ishai, E. Kushilevitz, and J.-F. Raymond, "Breaking the $O(n^{1/(2k-1)})$ barrier for information-theoretic private information retrieval," in *Proc. 43rd Annu. IEEE Symp. Found. Comp. Sci. (FOCS)*, Vancouver, BC, Canada, Nov. 16–19, 2002, pp. 261–270.

[4] S. Yekhanin, "Private information retrieval," *Commun. ACM*, vol. 53, no. 4, pp. 68–73, Apr. 2010.

[5] H. Corrigan-Gibbs and D. Kogan, "Private information retrieval with sublinear online time," in *Proc. 39th Annu. Int. Conf. Theory Appl. Crypto. Techn. (EUROCRYPT)*, Zagreb, Croatia, May 10–14, 2020, pp. 44–75.

[6] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 14–19, 2015, pp. 2842–2846.

[7] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM J. Appl. Algebra Geom.*, vol. 1, no. 1, pp. 647–664, Nov. 2017.

[8] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[9] 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.

[10] 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.

[11] S. Kumar, H.-Y. Lin, E. Rosnes, and A. Graell i Amat, "Achieving maximum distance separable private information retrieval capacity with linear codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4243–4273, Jul. 2019.

[12] H.-Y. Lin, S. Kumar, E. Rosnes, and A. Graell i Amat, "Asymmetry helps: Improved private information retrieval protocols for distributed storage," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guangzhou, China, Nov. 25–29, 2018.

[13] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "Weakly-private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1257–1261.

[14] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019, pp. 1262–1266.

[15] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "The capacity of single-server weakly-private information retrieval," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 415–427, Mar. 2021.

[16] ——, "Multi-server weakly-private information retrieval," *IEEE Trans. Inf. Theory*, vol. 68, no. 2, pp. 1197–1219, Feb. 2022.

[17] C. Qian, R. Zhou, C. Tian, and T. Liu, "Improved weakly private information retrieval codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Espoo, Finland, Jun. 26–Jul. 1, 2022, pp. 2840–2845.

[18] I. Samy, M. Attia, R. Tandon, and L. Lazos, "Asymmetric leaky private information retrieval," *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 5352–5369, Aug. 2021.

[19] R. Zhou, T. Guo, and C. Tian, "Weakly private information retrieval under the maximal leakage metric," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Los Angeles, CA, USA, Jun. 21–26, 2020, pp. 1089–1094.

[20] Y. Yakimenka, H.-Y. Lin, E. Rosnes, and J. Kliewer, "Optimal rate-distortion-leakage tradeoff for single-server information retrieval," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 832–846, Mar. 2022.

[21] C.-W. Weng, Y. Yakimenka, H.-Y. Lin, E. Rosnes, and J. Kliewer, "Generative adversarial user privacy in lossy single-server information retrieval," *IEEE Trans. Inf. Forens. Secur.*, vol. 17, pp. 3495–3510, 2022.

[22] G. Smith, "On the foundations of quantitative information flow," in *Proc. 12th Int. Conf. Found. Softw. Sci. Comput. Struct. (FoSSaCS)*, York, U.K., Mar. 22–29, 2009, pp. 288–302.

[23] G. Barthe and B. Köpf, "Information-theoretic bounds for differentially private mechanisms," in *Proc. 24th IEEE Comput. Secur. Found. Symp. (CSF)*, Cernay-la-Ville, France, Jun. 27–29, 2011, pp. 191–204.

[24] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[25] J. Zhu, Q. Yan, C. Qi, and X. Tang, "A new capacity-achieving private information retrieval scheme with (almost) optimal file length for coded servers," *IEEE Trans. Inf. Forens. Secur.*, vol. 15, pp. 1248–1260, 2020.

[26] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4904–4916, Aug. 2020.