Weakly-Private Information Retrieval

Citation for the original published paper (version of record):

Lin, H., Kumar, S., Rosnes, E. et al (2019)
Weakly-Private Information Retrieval
IEEE International Symposium on Information Theory - Proceedings, 2019: 1257-1261
http://dx.doi.org/10.1109/ISIT.2019.8849444

(article starts on next page)

Weakly-Private Information Retrieval

Hsuan-Yin Lin[*], **Siddhartha Kumar**[*], **Eirik Rosnes**[*], **Alexandre Graell i Amat**[†*], and **Eitan Yaakobi**[‡]
[*]Simula UiB, N–5008 Bergen, Norway
[†]Department of Electrical Engineering, Chalmers University of Technology, SE–41296 Gothenburg, Sweden
[‡]Department of Computer Science, Technion — Israel Institute of Technology, Haifa, 3200009 Israel

***Abstract*—Private information retrieval (*PIR*) protocols make it possible to retrieve a file from a database without disclosing any information about the identity of the file being retrieved. These protocols have been rigorously explored from an information-theoretic perspective in recent years. While existing protocols strictly impose that no information is leaked on the file's identity, this work initiates the study of the tradeoffs that can be achieved by relaxing the requirement of perfect privacy. In case the user is willing to leak some information on the identity of the retrieved file, we study how the PIR rate, as well as the upload cost and access complexity, can be improved. For the particular case of replicated servers, we propose two *weakly-private* information retrieval schemes based on two recent PIR protocols and a family of schemes based on partitioning. Lastly, we compare the performance of the proposed schemes.**

I. Introduction

In 1995 Chor *et al.* introduced the notion of private information retrieval (PIR) [1]. A PIR scheme allows a user to privately retrieve an arbitrary file from a database that is stored in multiple noncolluding servers without revealing any information about the requested file index to any server. The efficiency of a PIR scheme is usually measured in terms of the communication load, which is the sum of the number of uploaded and downloaded bits for retrieval of a single file. It has been extensively studied how it is possible to reduce the communication load using several copies of the database [2]–[4]. To achieve a more efficient PIR scheme, PIR protocols have also been considered jointly with *coded distributed storage systems* (DSSs), where the data is encoded by a linear code to store the files on n servers in a distributed manner [5], [6].

Recently, there has been a renewed interest to study the PIR problem from an information-theoretic formulation [5], [7], [8]. Under this setting, the file size is assumed to be arbitrarily large, and hence the upload cost can be ignored compared to the download cost. This then defines the *PIR rate* for a PIR scheme, which is equal to the amount of information retrieved per downloaded symbol. Recently, Sun and Jafar derived the optimal achievable PIR rate, the so-called *PIR capacity*, for the classical PIR model of replicated servers [7], [9]. Since then, several works have extended the results on PIR to different setups, e.g., coded DSSs [10]–[12], colluding servers [9], [12], and different figures of merit such as the access complexity [13].

The PIR model has also been extended in different interesting directions, for example PIR with side information [14] and more. All of the aforementioned models impose the strict requirement of perfect privacy, i.e., no information leakage. However, this assumption is quite restrictive and may be relaxed for practical applications. How to quantify the amount of sensitive information leaked from different privacy-enhancing technologies has been studied massively in the computer science society [15]. Therein, many information-theoretic privacy leakage metrics have been proposed, e.g., mutual information (MI) and worst-case entropy measures [16].

This paper takes a first step towards another parameter of the PIR framework, namely the information leakage. In particular, the goal of this paper is to study the tradeoffs of the different parameters of PIR protocols, such as the rate, upload cost, and access complexity, while the user is willing to leak some information on the identity of the retrieved file. We refer to such a scenario as *weakly-private* information retrieval (WPIR). Although related, our model is different from the one considered by Toledo *et al.* [17] in the computer science literature, where a modified metric based on *differential privacy* that relies on a particular scenario between an adversary and a number of users, is considered. In several scenarios, leaking part of the information of the retrieved file's identity is legitimate as long as there is still enough ambiguity on the file to meet the privacy requirement specified by the user. For example, the user may be willing to share with the servers that the file is a movie (and not a book or other forms of files), or only the movie's genre, however the identity of the movie should be kept private.

The rest of the paper is organized as follows. Section II presents the notation, definitions, and system model used throughout the paper. In Section III, a basic solution for WPIR is presented in which the database is partitioned into several partitions and the user is willing to expose only the partition that the requested file belongs to. In Section IV, we propose a WPIR scheme for replicated databases building upon a PIR protocol recently introduced in [18] and study its tradeoffs between different parameters while relaxing the privacy constraint. A second WPIR scheme is presented in Section V, based on the PIR scheme from [12]. Lastly, Section VI presents numerical results and compares the schemes studied in the paper with respect to rate, upload cost, and access complexity.

II. Preliminaries

A. Notation

We denote by $\mathbb{N}$ the set of all positive integers, $[a] \triangleq \{1,2,\ldots,a\}$, and $[a:b] \triangleq \{a,a+1,\ldots,b\}$ for $a,b \in \{0\}\cup\mathbb{N}$, $a \leq b$. Vectors 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. For a given index set $\mathcal{S}$, we write $X^{\mathcal{S}}$ and $Y_{\mathcal{S}}$ to represent $\{X^{(m)}\colon m \in \mathcal{S}\}$ and $\{Y_l\colon l \in \mathcal{S}\}$, respectively. $X \perp\!\!\!\perp Y$ means that the two RVs X and Y are independent. $\mathbb{E}_X[\cdot]$ denotes expectation over the RV X. $X \sim \text{Bernoulli}(p)$ denotes a Bernoulli-distributed RV with $\Pr[X=1] = p = 1-\Pr[X=0]$ and $X \sim \text{Uniform}(\mathcal{S})$ a uniformly-distributed RV over the set $\mathcal{S}$. $(\cdot)^{\mathsf{T}}$ denotes the transpose of its argument. The Hamming weight of a binary vector $\boldsymbol{x}$ is denoted by $w_{\mathsf{H}}(\boldsymbol{x})$ and the inner product of $\boldsymbol{x}$ and

This work was partially funded by the Research Council of Norway (grant 240985/F20), the Swedish Research Council (grant #2016-04253), and the Israel Science Foundation (grant #1817/18).

$\boldsymbol{y}$ is denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. $\mathsf{H}(X)$ represents the entropy X and $\mathsf{I}(X;Y)$ the MI between X and Y. The Galois field with q elements is denoted by GF(q).

B. System Model

We consider a DSS with n noncolluding replicated servers, each storing M independent files $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(\mathsf{M})}$, where each file $\boldsymbol{X}^{(m)} = \bigl(X_1^{(m)}, \ldots, X_\beta^{(m)}\bigr)^{\mathsf{T}}$, $m \in [\mathsf{M}]$, can be seen as a $\beta \times 1$ vector over GF(q). Assume that each element of $\boldsymbol{X}^{(m)}$ is chosen independently and uniformly at random from GF(q). Thus, in q-ary units, we have $\mathsf{H}\bigl(\boldsymbol{X}^{(m)}\bigr) = \beta$, $\forall\, m \in [\mathsf{M}]$.

In information retrieval (IR), a user wishes to efficiently retrieve one of the M files stored in the replicated DSS. Similar to the detailed mathematical description in [18], we assume that the requested file index M is a RV and $M \sim \text{Uniform}([\mathsf{M}])$. We give the following definition of IR schemes.

Definition 1. *An (M, n) IR scheme $\mathscr{C}$ for a DSS with n servers storing M files consists of*

- *a global random strategy $\boldsymbol{S}$, whose alphabet is $\mathcal{S}$,*
- *n query-encoding functions ϕ_l, $l \in [n]$, that generate n queries $\boldsymbol{Q}_l = \phi_l(M, \boldsymbol{S})$ with alphabet $\mathcal{Q}_l$, where query $\boldsymbol{Q}_l$ is sent to server l,*
- *n answer functions φ_l that return the answers $\boldsymbol{A}_l = \varphi_l(\boldsymbol{Q}_l, \boldsymbol{X}^{[\mathsf{M}]})$, with alphabet $\mathcal{A}$ for all $l \in [n]$,*
- *n answer-length functions $L_l(\boldsymbol{A}_l)$, with range $\{0\} \cup \mathbb{N}$, that define the length of the answers,*
- *n access-number functions $\delta_l(\boldsymbol{Q}_l)$, with range $\{0\} \cup \mathbb{N}$, that define the number of symbols accessed by $\boldsymbol{Q}_l$.*

This scheme should satisfy the condition of perfect retrievability,

$$\mathsf{H}\bigl(\boldsymbol{X}^{(M)} \bigm| \boldsymbol{A}_{[n]}, \boldsymbol{Q}_{[n]}, M\bigr) = 0. \tag{1}$$

Note that a PIR scheme is an (M, n) IR scheme that satisfies full privacy for all servers, i.e., for every $m, m' \in [\mathsf{M}]$ with $m \neq m'$, the condition

$$\Pr[\boldsymbol{Q}_l = \boldsymbol{q}_l \,|\, M = m] = \Pr[\boldsymbol{Q}_l = \boldsymbol{q}_l \,|\, M = m'] \tag{2}$$

holds for all $\boldsymbol{q}_l \in \mathcal{Q}_l$, $l \in [n]$. The privacy constraint (2) is equivalent to the statement that $M \perp\!\!\!\perp \boldsymbol{Q}_l$. We denote by $\boldsymbol{Q}_l^{(m)}$ the query sent to server l if file $\boldsymbol{X}^{(m)}$ is requested, which is a RV with probability mass function (PMF) $P_{\boldsymbol{Q}_l^{(m)}}(\boldsymbol{q}_l) \triangleq \Pr[\boldsymbol{Q}_l = \boldsymbol{q}_l \,|\, M = m]$.

We refer to an (M, n) IR scheme that does not satisfy (2) as a *WPIR scheme*, as opposed to a PIR scheme that leaks no information.

C. Metrics of Information Leakage

We consider both the MI and the worst-case information leakage (WIL) [16] between M and $\boldsymbol{Q}_l$ to define suitable measures of information leakage for an IR scheme. For the former, we use the following proposition to motivate the definition of information leakage for an (M, n) IR scheme.

Proposition 1 (Time-Sharing Principle for the MI Metric)**.** *Consider an (M, n) IR scheme $\mathscr{C}$, where the leakage of the l-th server is defined as $\mathsf{I}(M;\boldsymbol{Q}_l)$, $l \in [n]$. Then, there exists an (M, n) IR scheme $\bar{\mathscr{C}}$ with leakage $\bar{\rho} \triangleq \frac{1}{n}\sum_{l\in[n]} \mathsf{I}(M;\boldsymbol{Q}_l)$ for every server.*

Proposition 1 indicates that we can obtain an (M, n) IR scheme with equal MI leakage at each server by cyclically shifting the servers' queries of an existing (M, n) IR scheme n times.

Hence, to characterize the overall leakage of a given (M, n) IR scheme $\mathscr{C}$ in terms of MI, we consider the information leakage metric $\rho^{(\mathsf{MI})}(\mathscr{C}) \triangleq \frac{1}{n}\sum_{l\in[n]} \mathsf{I}(M;\boldsymbol{Q}_l)$.

Definition 2. *The WIL of the l-th server is defined as $\mathsf{I}^{\mathsf{worst}}(M;\boldsymbol{Q}_l) = \mathsf{H}(M) - \min_{\boldsymbol{q}_l\in\mathcal{Q}_l} \mathsf{H}(M \,|\, \boldsymbol{Q}_l = \boldsymbol{q}_l)$. The overall WIL of a given (M, n) IR scheme $\mathscr{C}$ is then given as $\rho^{(\mathsf{WIL})}(\mathscr{C}) \triangleq \max_{l\in[n]} \mathsf{I}^{\mathsf{worst}}(M;\boldsymbol{Q}_l)$.*

D. Download Cost, Rate, Upload Cost, and Access Complexity of an (M, n) IR Scheme

The download cost of an IR scheme $\mathscr{C}$, denoted by $\mathsf{D}(\mathscr{C})$, is defined as the expected number of downloaded symbols among all servers for the retrieval of a single file,

$$\mathsf{D}(\mathscr{C}) \triangleq \sum_{l=1}^{n} \mathbb{E}_{\boldsymbol{Q}_l}[L_l(\boldsymbol{A}_l)] = \frac{1}{\mathsf{M}} \sum_{m=1}^{\mathsf{M}} \sum_{l=1}^{n} \mathbb{E}_{\boldsymbol{Q}_l^{(m)}}[L_l(\boldsymbol{A}_l)].$$

Accordingly, the IR rate of an IR scheme $\mathscr{C}$ is defined as $\mathsf{R}(\mathscr{C}) \triangleq \frac{\beta}{\mathsf{D}(\mathscr{C})}$. The upload cost $\mathsf{U}(\mathscr{C})$ of an IR scheme $\mathscr{C}$ is defined as the sum of the entropies of the queries $\boldsymbol{Q}_{[n]}$,

$$\mathsf{U}(\mathscr{C}) \triangleq \sum_{l=1}^{n} \mathsf{H}(\boldsymbol{Q}_l).$$

Moreover, the access complexity $\Delta(\mathscr{C})$ of an IR scheme $\mathscr{C}$ is defined as the expected number of accessed symbols among all servers for the retrieval of a single file,

$$\Delta(\mathscr{C}) \triangleq \sum_{l=1}^{n} \mathbb{E}_{\boldsymbol{Q}_l}[\delta_l(\boldsymbol{Q}_l)] = \frac{1}{\mathsf{M}} \sum_{m=1}^{\mathsf{M}} \sum_{l=1}^{n} \mathbb{E}_{\boldsymbol{Q}_l^{(m)}}[\delta_l(\boldsymbol{Q}_l)].$$

An achievable 4-tuple of an IR scheme is defined as follows.

Definition 3. *Consider a DSS with n noncolluding servers storing M files. A 4-tuple $(\mathsf{R}, \mathsf{U}, \Delta, \rho)$ is said to be* achievable *with information leakage metric $\rho^{(\cdot)}$ if there exists an (M, n) IR scheme $\mathscr{C}$ such that $\mathsf{R}(\mathscr{C}) = \mathsf{R}$, $\mathsf{U}(\mathscr{C}) = \mathsf{U}$, $\Delta(\mathscr{C}) = \Delta$, and $\rho^{(\cdot)}(\mathscr{C}) = \rho$.*

We remark that a PIR scheme is equivalent to an (M, n) IR scheme with $\rho^{(\cdot)} = 0$. It was shown in [7] that for n noncolluding replicated servers and for a given number of files M, the PIR capacity, denoted by $\mathsf{C}_{\mathsf{M},n}$, is $\mathsf{C}_{\mathsf{M},n} = \bigl(1 + 1/n + \cdots + 1/n^{\mathsf{M}-1}\bigr)^{-1}$.

III. Partition WPIR Scheme

A simple approach for the construction of WPIR schemes is to first partition the database into η equally-sized partitions, each consisting of M/η files[1], and then use a given $(\mathsf{M}/\eta, n)$ IR scheme to retrieve a file from the corresponding partition. Obviously, the resulting scheme is not a PIR scheme, since the servers gain the knowledge of which partition the requested file belongs to. In this section, we pursue this approach to construct an (M, n) IR scheme building on a given $(\mathsf{M}/\eta, n)$ IR scheme as a subscheme.

The partition (M, n) IR scheme is formally described as follows. Assume the requested file $\boldsymbol{X}^{(m)}$ belongs to the j-th partition, where $j \in [\eta]$. Then, the query $\boldsymbol{Q}_l$ is constructed as

$$\boldsymbol{Q}_l = \bigl(\tilde{\boldsymbol{Q}}_l, j\bigr) \in \tilde{\mathcal{Q}}_l \times [\eta], \quad l \in [n], \tag{3}$$

where $\tilde{\boldsymbol{Q}}_l$ is the query of an existing $(\mathsf{M}/\eta, n)$ IR scheme $\tilde{\mathscr{C}}$.

The following theorem states the achievable 4-tuple of the partition scheme.

[1] While it is not necessary that each partition has an equal number of files, for simplicity in this paper we make this assumption.

Theorem 1. *Consider a DSS with n noncolluding servers storing M files, and let $\tilde{\mathscr{C}}$ be an $(\mathsf{M}/\eta, n)$ IR scheme with achievable 4-tuple $\left(\tilde{\mathsf{R}}, \tilde{\mathsf{U}}, \tilde{\Delta}, \tilde{\rho}^{(\cdot)}\right)$. Then, the 4-tuple*

$$\begin{aligned}&\left(\mathsf{R}(\mathscr{C}), \mathsf{U}(\mathscr{C}), \Delta(\mathscr{C}), \rho^{(\cdot)}(\mathscr{C})\right)\\ &\quad = \left(\tilde{\mathsf{R}}, \tilde{\mathsf{U}} + n\log_2\eta, \tilde{\Delta}, \tilde{\rho}^{(\cdot)} + \log_2\eta\right) \qquad (4)\end{aligned}$$

is achievable by the (M, n) partition scheme $\mathscr{C}$ constructed from $\tilde{\mathscr{C}}$ as described in (3).

Since a PIR scheme is also an IR scheme, this simple approach for the construction of WPIR schemes can also be adapted to use one of the existing $(\mathsf{M}/\eta, n)$ PIR schemes in the literature as a subscheme. We refer to the partition scheme that uses a PIR scheme as the underlying subscheme and the query generation in (3) as a *basic scheme* and denote it by $\mathscr{C}^{\mathsf{basic}}$ (it gives the 4-tuple as in (4) with $\tilde{\rho}^{(\cdot)} = 0$). In Section IV-B, we will present another partition WPIR scheme $\mathscr{C}'$ based on our proposed IR scheme.

IV. (M, n) Scheme 1

In [18, Sec. 3.2], a PIR scheme that achieves both the minimum upload and download costs was proposed. The queries $\boldsymbol{Q}_{[n]}$ of the scheme in [18, Sec. 3.2] are randomly generated according to a random strategy $\boldsymbol{S} = (S_1, \ldots, S_{\mathsf{M}-1})$ with independent and identically distributed (i.i.d.) entries according to Uniform($[0 : n-1]$). In this section, we introduce an (M, n) WPIR scheme, referred to as *Scheme 1* and denoted by $\mathscr{C}_1$, based on the PIR scheme in [18]. Scheme 1 can be seen as a generalization of the PIR scheme in [18] where we lift the perfect privacy condition (2).

For the proposed scheme, we assume the file size to be $\beta = n-1$, and we represent a query by a length-M vector $\boldsymbol{q}_l = (q_{l,1}, \ldots, q_{l,\mathsf{M}}) \in \mathcal{Q}_l \subseteq [0 : n-1]^{\mathsf{M}}$. Also, the realization of $\boldsymbol{S}$ is denoted as a length-$(\mathsf{M}-1)$ vector $\boldsymbol{s} = (s_1, \ldots, s_{\mathsf{M}-1})$, $s_j \in [0 : n-1]$, $j \in [\mathsf{M}-1]$.

Before describing Scheme 1 in detail for the general case, for simplicity we first present Scheme 1 for the case of $\mathsf{M} = 2$ files and $n = 2$ servers (i.e., both servers 1 and 2 store $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$) in the following example.

Example 1. *We illustrate the $(2, 2)$ Scheme 1 obtained by adopting a nonuniformly-distributed random strategy $\boldsymbol{S}$ giving a joint PMF $P_{\boldsymbol{Q}_1,\boldsymbol{Q}_2}(\boldsymbol{q}_1, \boldsymbol{q}_2)$ as*

$$P_{\boldsymbol{Q}_1^{(1)},\boldsymbol{Q}_2^{(1)}}(\boldsymbol{q}_1, \boldsymbol{q}_2) = \begin{cases} 1-p & \text{if } \boldsymbol{q}_1 = (0,0), \boldsymbol{q}_2 = (1,0), \\ p & \text{if } \boldsymbol{q}_1 = (1,1), \boldsymbol{q}_2 = (0,1), \\ 0 & \text{otherwise,} \end{cases}$$

$$P_{\boldsymbol{Q}_1^{(2)},\boldsymbol{Q}_2^{(2)}}(\boldsymbol{q}_1, \boldsymbol{q}_2) = \begin{cases} 1-p & \text{if } \boldsymbol{q}_1 = (0,0), \boldsymbol{q}_2 = (0,1), \\ p & \text{if } \boldsymbol{q}_1 = (1,1), \boldsymbol{q}_2 = (1,0), \\ 0 & \text{otherwise.} \end{cases}$$

Files $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are composed of one stripe each ($\beta = n-1 = 1$). The answers $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are given by $(\boldsymbol{A}_1(\boldsymbol{q}_1), \boldsymbol{A}_2(\boldsymbol{q}_2)) = \left(X^{(1)}_{q_{1,1}} + X^{(2)}_{q_{1,2}}, X^{(1)}_{q_{2,1}} + X^{(2)}_{q_{2,2}}\right)$, where $X^{(m)}_0 = 0$ for all $m \in [2]$.

One can easily verify that perfect retrievability is satisfied for the above $(2, 2)$ IR scheme. Its IR rate is a function of p and is given by $\mathsf{R}(p) = (p+(1-p)+p)^{-1} = (1+p)^{-1}$. Observe that $M \perp\!\!\!\perp \boldsymbol{Q}_1$, which implies that $\mathsf{I}(M; \boldsymbol{Q}_1) = \mathsf{I}^{\mathsf{worst}}(M; \boldsymbol{Q}_1) = 0$.

The information leakage is $\rho^{(\mathsf{MI})} = \frac{1-\mathsf{H}_\mathsf{b}(p)}{2}$ and $\rho^{(\mathsf{WIL})} = 1 - \mathsf{H}_\mathsf{b}(p)$ for $0 \leq p \leq \frac{1}{2}$, where $\mathsf{H}_\mathsf{b}(p) \triangleq -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function. From this derivation, it follows that the $(2, 2)$ Scheme 1 achieves perfect privacy for $p = \frac{1}{2}$. The IR rate of the $(2, 2)$ Scheme 1, $\mathsf{R}(\mathscr{C}_1)$, is depicted in Fig. 1 as a function of the information leakage $\rho^{(\cdot)}$. Interestingly, by sacrificing perfect privacy, it is possible to achieve an IR rate larger than the 2-server PIR capacity for 2 files. As expected, the IR rate increases with increasing information leakage.

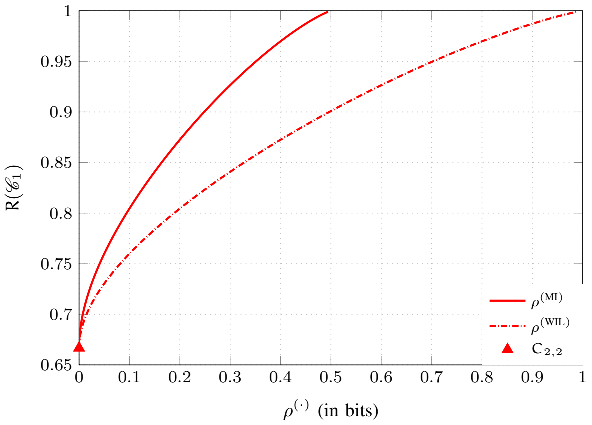

Fig. 1. The IR rate $\mathsf{R}(\mathscr{C}_1) \in \left[\frac{2}{3}, 1\right]$ of the proposed $(2, 2)$ Scheme 1, as a function of $\rho^{(\cdot)}$. The triangle marks the 2-server PIR capacity for $\mathsf{M} = 2$.

Now, we describe Scheme 1 for the general case of M files and n servers. We assume that the user wants to download file $\boldsymbol{X}^{(m)}$ and has a random strategy $\boldsymbol{S}$ that takes on values $\boldsymbol{s} \in [0 : n-1]^{\mathsf{M}-1}$ with PMF $P_{\boldsymbol{S}}(\boldsymbol{s})$.

1) Query Generation: The query $\boldsymbol{q}_l \in \mathcal{Q}_l$ sent to the l-th server, resulting from the query-encoding function ϕ_l, is defined as $\boldsymbol{q}_l = (s_1, \ldots, s_{m-1}, q_{l,m}, s_m, \ldots, s_{\mathsf{M}-1})$, where $q_{l,m} = \left(l - 1 - \sum_{j\in[\mathsf{M}-1]} s_j\right) \bmod n$. It follows that $\left(\sum_{m'\in[\mathsf{M}]} q_{l,m'}\right) \bmod n = l-1$. Note that the PMF of $\boldsymbol{Q}_l$ conditioned on the file index M satisfies $P_{\boldsymbol{Q}_l^{(m)}}(\boldsymbol{q}_l) = P_{\boldsymbol{S}}(\boldsymbol{s})$.

2) Answer Construction: The answer function φ_l maps the query $\boldsymbol{q}_l$ into $\boldsymbol{A}_l = \varphi_l(\boldsymbol{q}_l, \boldsymbol{X}^{[\mathsf{M}]}) = X^{(1)}_{q_{l,1}} + \cdots + X^{(\mathsf{M})}_{q_{l,\mathsf{M}}}$, where $X^{(m')}_0 = 0$ for all $m' \in [\mathsf{M}]$. Further, we see that the answer-length functions satisfy

$$L_l(\boldsymbol{A}_l) = \begin{cases} 0 & \text{if } \boldsymbol{q}_l = \boldsymbol{0}, \\ 1 & \text{otherwise.} \end{cases} \qquad (5)$$

This completes the construction of the (M, n) Scheme 1. The perfect retrievability of Scheme 1 can be verified by following the same argumentation as in [18, Sec. 3.2]. Moreover, using (5), the IR rate of the (M, n) Scheme 1, $\mathscr{C}_1$, can be shown to be

$$\mathsf{R}(\mathscr{C}_1) = \frac{n-1}{1 - P_{\boldsymbol{Q}_1}(\boldsymbol{0}) + n - 1}.$$

We also remark that if Scheme 1 uses a random strategy $\boldsymbol{S}$ with $\{S_j\}_{j=1}^{\mathsf{M}-1}$ i.i.d. according to Uniform($[0 : n-1]$), then it satisfies (2) and is equivalent to the PIR capacity-achieving scheme proposed in [18].

A. $(\mathsf{M}, 2)$ Scheme 1 With $\{S_j\}_{j=1}^{\mathsf{M}-1}$ I.I.D. According to Bernoulli(p)

The following result gives an achievable 4-tuple for Scheme 1 for the case of 2 servers and a random strategy $\boldsymbol{S} = (S_1, \ldots, S_{\mathsf{M}-1})$ with i.i.d. entries according to Bernoulli(p).

Theorem 2. *Consider $0 \leq p \leq \frac{1}{2}$. Then, the 4-tuple $\left(\mathsf{R}_1, \mathsf{U}_1, \Delta_1, \rho_1^{(\cdot)}\right)$,*

$$\mathsf{R}_1 = \left(1 - (1-p)^{\mathsf{M}-1} + 1\right)^{-1},$$

$$\mathsf{U}_1 = -\sum_{w=0}^{\mathsf{M}} \binom{\mathsf{M}}{w} f(w,p)\log_2 f(w,p),$$

$$
\begin{aligned}
\Delta_1 &= \textstyle\sum_{w=0}^{\mathsf{M}} w\binom{\mathsf{M}}{w} f(w,p),\\
\rho_1^{(\mathsf{MI})} &= \mathsf{U}_1/2 - (\mathsf{M}-1)\,\mathsf{H}_\mathsf{b}(p),\ \text{ and}\\
\rho_1^{(\mathsf{WIL})} &= \log_2 \mathsf{M} - \min_{w\in[0:\mathsf{M}]} \mathsf{H}(M_w),
\end{aligned}
$$

where $f(w,p) \triangleq \bigl[(\mathsf{M}-w)(1-p)^{\mathsf{M}-w-1}p^w + w(1-p)^{\mathsf{M}-w}p^{w-1}\bigr]/\mathsf{M}$ *and* M_w *has PMF*

$$
P_{M_w}(m') = \begin{cases} \frac{(1-p)^{\mathsf{M}-w-1}p^w}{\mathsf{M}f(w,p)} & \text{if } m' \in [\mathsf{M}-w],\\ \frac{(1-p)^{\mathsf{M}-w}p^{w-1}}{\mathsf{M}f(w,p)} & \text{if } m' \in [\mathsf{M}-w+1:\mathsf{M}], \end{cases}
$$

is achievable by the $(\mathsf{M},2)$ *Scheme 1 with* $\{S_j\}_{j=1}^{\mathsf{M}-1}$ *i.i.d. according to* $\mathrm{Bernoulli}(p)$.

B. Partition Scheme 1: Using Scheme 1 as a Subscheme

In Section III, the concept of adopting an existing $(\mathsf{M}/\eta, n)$ IR scheme to retrieve a file from a given partition is introduced. In this section, unlike (3), where the user sends different queries for different requested files among all partitions, we use a slightly more sophisticated way to construct a WPIR scheme by using Scheme 1 as a subscheme for every partition. We refer to this scheme as *partition Scheme 1* and denote it by $\mathscr{C}_1^{\mathsf{part}}$. In the following, we present the query generation and the answer construction.

1) Query Generation: We consider the j-th partition, $\mathcal{P}_j$, $j \in [\eta]$, containing all files of indices $(j-1)\mathsf{M}/\eta + 1, \dots, j\mathsf{M}/\eta$. Given a requested file with index $m = (j-1)\mathsf{M}/\eta + m' \in \mathcal{P}_j$, $m' \in [\mathsf{M}/\eta]$, we consider an $(\mathsf{M}/\eta, n)$ Scheme 1 as a subscheme for partition $\mathcal{P}_j$. The l-th query $\boldsymbol{q}_l \in \mathcal{Q}_l$, $l \in [n]$, is defined as

$$
\begin{aligned}
\boldsymbol{q}_l = \bigl(&\mathbf{0}_{1\times(j-1)\mathsf{M}/\eta}, s_1, \dots, s_{m'-1}, q_{l,(j-1)\mathsf{M}/\eta+m'},\\
&s_{m'}, \dots, s_{\mathsf{M}/\eta-1}, \mathbf{0}_{1\times(\eta-j)\mathsf{M}/\eta}\bigr),
\end{aligned}
$$

where $q_{l,(j-1)\mathsf{M}/\eta+m'} = \bigl(l-1-\sum_{j\in[\mathsf{M}/\eta-1]} s_j\bigr) \bmod n$. We remark that it is possible that the user sends the all-zero query $\boldsymbol{q}_l = \mathbf{0}$ to request different files among all partitions. In this way, since the uncertainty on the requested file is increased, it follows that the leakage of $\mathscr{C}_1^{\mathsf{part}}$ is slightly smaller than the leakage of the basic scheme. Moreover, the query alphabet size is not exactly the same for all servers, i.e., we have $|\mathcal{Q}_1| = 1 + \eta\bigl(n^{\mathsf{M}/\eta-1} - 1\bigr)$ and $|\mathcal{Q}_l| = \eta \cdot n^{\mathsf{M}/\eta-1}$ for $l \in [2:n]$.

2) Answer Construction: Similar to Scheme 1 in Section IV, the answer function φ_l maps query $\boldsymbol{q}_l$ into $\boldsymbol{A}_l = \varphi(\boldsymbol{q}_l, \boldsymbol{X}^{[\mathsf{M}]}) = X_{q_{l,1}}^{(1)} + \cdots + X_{q_{l,\mathsf{M}}}^{(\mathsf{M})}$, where $X_0^{(m')} = 0$ for all $m' \in [\mathsf{M}]$. Further, we see that $L_l(\boldsymbol{A}_l)$ satisfies (5).

C. $(\mathsf{M},2)$ *Partition Scheme 1 With* $\{S_j\}_{j=1}^{\mathsf{M}/\eta-1}$ *I.I.D. According to* $\mathrm{Bernoulli}(1/2)$

We focus again on the case of 2 servers. Since the servers can learn some information from which partition the requested file belongs to, in order to have a relatively small leakage of partition Scheme 1, it is reasonable to use Scheme 1 with $\{S_j\}_{j=1}^{\mathsf{M}/\eta-1}$ i.i.d. according to $\mathrm{Bernoulli}(1/2)$ as a subscheme (i.e., a PIR subscheme). We have the following result.

Theorem 3. *Let* M/η *be a positive integer with* $\eta \in [\mathsf{M}-1]$. *Then, the* 4*-tuple* $\bigl(\mathsf{R}_{1,\mathsf{P}}, \mathsf{U}_{1,\mathsf{P}}, \Delta_{1,\mathsf{P}}, \rho_{1,\mathsf{P}}^{(\cdot)}\bigr)$,

$$
\begin{aligned}
\mathsf{R}_{1,\mathsf{P}} &= \bigl(1 + 1/2 + \cdots + 1/2^{\mathsf{M}/\eta-1}\bigr)^{-1},\\
\mathsf{U}_{1,\mathsf{P}} &= 2\bigl[\mathsf{M}/\eta - 1 + \log_2 \eta\bigr] - \log_2 \eta/2^{\mathsf{M}/\eta-1},\\
\Delta_{1,\mathsf{P}} &= \mathsf{M}/\eta,\\
\rho_{1,\mathsf{P}}^{(\mathsf{MI})} &= \log_2 \eta - \log_2 \eta/2^{\mathsf{M}/\eta},\ \text{ and}\\
\rho_{1,\mathsf{P}}^{(\mathsf{WIL})} &= \log_2 \eta
\end{aligned}
$$

is achievable by the $(\mathsf{M},2)$ *partition Scheme 1 with the* $(\mathsf{M}/\eta,2)$ *Scheme 1 with* $\{S_j\}_{j=1}^{\mathsf{M}/\eta-1}$ *i.i.d. according to* $\mathrm{Bernoulli}(1/2)$ *as a subscheme.*

Let $\bigl(\tilde{\mathsf{R}}, \tilde{\mathsf{U}}, \tilde{\Delta}, 0\bigr)$ be the achievable 4-tuple of the $(\mathsf{M}/\eta, 2)$ Scheme 1 with $\{S_j\}_{j=1}^{\mathsf{M}/\eta-1}$ i.i.d. according to $\mathrm{Bernoulli}(1/2)$. It follows that $\mathsf{U}_{1,\mathsf{P}} = \tilde{\mathsf{U}} + 2\log_2\eta - \log_2\eta/2^{\mathsf{M}/\eta-1} < \mathsf{U}(\mathscr{C}^{\mathsf{basic}})$ and $\rho_{1,\mathsf{P}}^{(\mathsf{MI})} = \log_2\eta - \log_2\eta/2^{\mathsf{M}/\eta} < \rho^{(\mathsf{MI})}(\mathscr{C}^{\mathsf{basic}})$, while $\mathsf{R}_{1,\mathsf{P}}$, $\Delta_{1,\mathsf{P}}$, and $\rho_{1,\mathsf{P}}^{(\mathsf{WIL})}$ are identical to those of the basic scheme $\mathscr{C}^{\mathsf{basic}}$ in Section III (see the details in Theorem 1). Hence, in the numerical results section (see Fig. 2), the results of $\mathscr{C}^{\mathsf{basic}}$ are not presented.

V. Constant-Rate (M, n) Scheme 2

We propose an alternative WPIR scheme, referred to as *Scheme 2* and denoted by $\mathscr{C}_2$, based on the PIR scheme in [12, Lem. 4]. Scheme 2 is constructed as follows. Assume that $\beta = n-1$ and that the user requests file $\boldsymbol{X}^{(m)}$. The random strategy $\boldsymbol{S}$ takes the form of a vector $\boldsymbol{S} = (S_1, \dots, S_{\beta\mathsf{M}}) \in \mathrm{GF}(q)^{\beta\mathsf{M}}$ of length $\beta\mathsf{M}$. The query vector $\boldsymbol{Q}_l \in \mathcal{Q}_l = \mathrm{GF}(q)^{\beta\mathsf{M}}$, of length $\beta\mathsf{M}$, is obtained as $\boldsymbol{Q}_l = \phi(m, \boldsymbol{S}) = \boldsymbol{S} + \boldsymbol{v}_l$, where the vector $\boldsymbol{v}_l$ is deterministic and is completely determined by m. We refer the reader to [12, Sec. V] for details on the design of $\boldsymbol{v}_l$. The l-th server responds to its corresponding query with the answer $A_l \in \mathcal{A} = \mathrm{GF}(q)$ obtained as $A_l = \varphi(\boldsymbol{Q}_l, \boldsymbol{X}^{[\mathsf{M}]}) = \langle \boldsymbol{Q}_l, (X_1^{(1)}, \dots, X_\beta^{(1)}, \dots, X_\beta^{(\mathsf{M})}) \rangle$.

For the case where $\{S_j\}_{j=1}^{\beta\mathsf{M}}$ are i.i.d. according to $\mathrm{Uniform}(\mathrm{GF}(q))$, Scheme 2 achieves perfect privacy, and the scheme boils down to the PIR scheme in [12, Lem. 4]. Furthermore, similar to [12, Thm. 2], it can be shown that the scheme achieves perfect retrievability (see (1)), and since its answer-lengths are constant for all possible queries of each server, the IR rate R_2 of $\mathscr{C}_2$ is equal to $1 - 1/n$, irrespective of the information leakage $\rho^{(\cdot)}$.

A. $(\mathsf{M},2)$ *Scheme 2 With* $\{S_j\}_{j=1}^{\mathsf{M}}$ *I.I.D. According to* $\mathrm{Bernoulli}(p)$

Consider the binary field. We have the following result.

Theorem 4. *Consider* $0 \le p \le \frac{1}{2}$. *Then, the* 4*-tuple* $\bigl(1/2, \mathsf{U}_2, \Delta_2, \rho_2^{(\cdot)}\bigr)$,

$$
\begin{aligned}
\mathsf{U}_2 &= -\textstyle\sum_{w=0}^{\mathsf{M}} \binom{\mathsf{M}}{w} g(w,p)\log_2 g(w,p) + \mathsf{M}\,\mathsf{H}_\mathsf{b}(p),\\
\Delta_2 &= \textstyle\sum_{w=0}^{\mathsf{M}} w\binom{\mathsf{M}}{w}\bigl(g(w,p) + h(w,p)\bigr),\\
\rho_2^{(\mathsf{MI})} &= \mathsf{U}_2/2 - \mathsf{M}\,\mathsf{H}_\mathsf{b}(p),\ \text{ and}\\
\rho_2^{(\mathsf{WIL})} &= \log_2 \mathsf{M} - \min_{w\in[0:\mathsf{M}]} \mathsf{H}(M'_w),
\end{aligned}
$$

where $g(w,p) \triangleq \frac{1}{\mathsf{M}}\bigl[(\mathsf{M}-w)(1-p)^{\mathsf{M}-w-1}p^{w+1} + w(1-p)^{\mathsf{M}-w+1}p^{w-1}\bigr]$, $h(w,p) \triangleq (1-p)^{\mathsf{M}-w}p^w$, *and* M'_w *has PMF*

$$
P_{M'_w}(m') = \begin{cases} \frac{(1-p)^{\mathsf{M}-w-1}p^{w+1}}{\mathsf{M}g(w,p)} & \text{if } m' \in [\mathsf{M}-w],\\ \frac{(1-p)^{\mathsf{M}-w+1}p^{w-1}}{\mathsf{M}g(w,p)} & \text{if } m' \in [\mathsf{M}-w+1:\mathsf{M}], \end{cases}
$$

is achievable by the $(\mathsf{M},2)$ *Scheme 2 with* $\{S_j\}_{j=1}^{\mathsf{M}}$ *i.i.d. according to* $\mathrm{Bernoulli}(p)$.

In the following subsection, we analyze the $(\mathsf{M},2)$ Scheme 2 with a uniformly-distributed random strategy $\boldsymbol{S}$. Note that similarly to partition Scheme 1 in Section IV-C, we can also construct a partition scheme by using Scheme 2 as a subscheme for every partition. Since the analysis is almost the same as for partition Scheme 1, and the result for the $(\mathsf{M},2)$ partition Scheme 2 with $\{S_j\}_{j=1}^{\mathsf{M}/\eta}$ i.i.d. according to $\mathrm{Bernoulli}(1/2)$ is very close to the result in Theorem 3, we omit it.

B. $(\mathsf{M}, 2)$ *Scheme 2 With* $\boldsymbol{S}$ *Uniformly Distributed*

We consider the $(\mathsf{M}, 2)$ Scheme 2 with $\boldsymbol{S}$ uniformly distributed over all length-M binary vectors of weight w. In other words, $\boldsymbol{S} \sim \mathrm{Uniform}(\mathcal{B}_{w,\mathsf{M}})$, where $\mathcal{B}_{w,\mathsf{M}} \triangleq \{\boldsymbol{s} \in \{0,1\}^{\mathsf{M}} : w_{\mathsf{H}}(\boldsymbol{s}) = w\}$.

Theorem 5. *Consider* $0 \leq w \leq \mathsf{M}$. *Then, the* 4*-tuple* $(1/2, \mathsf{U}_{2,\mathsf{U}}, \Delta_{2,\mathsf{U}}, \rho^{(\cdot)}_{2,\mathsf{U}})$,

$$\begin{aligned}\mathsf{U}_{2,\mathsf{U}} &= \log_2 \tbinom{\mathsf{M}}{w} + y(w, \mathsf{M}),\\ \Delta_{2,\mathsf{U}} &= 1 + 2w(1 - 1/\mathsf{M}),\\ \rho^{(\mathrm{MI})}_{2,\mathsf{U}} &= \bigl(y(w, \mathsf{M}) - \log_2 \tbinom{\mathsf{M}}{w}\bigr)/2, \text{ and}\\ \rho^{(\mathrm{WIL})}_{2,\mathsf{U}} &= \log_2 \mathsf{M} - \min\{\log_2 (w + 1), \log_2 (\mathsf{M} - w + 1)\},\end{aligned}$$

where $y(w, \mathsf{M}) \triangleq \log_2 \binom{\mathsf{M}}{w} + \log_2 \mathsf{M} - \frac{\mathsf{M}-w}{\mathsf{M}} \log_2 (w + 1) - \frac{w}{\mathsf{M}} \log_2 (\mathsf{M} - w + 1)$, *is achievable by the* $(\mathsf{M}, 2)$ *Scheme 2 with* $\boldsymbol{S} \sim \mathrm{Uniform}(\mathcal{B}_{w,\mathsf{M}})$.

We remark that the analysis of the $(\mathsf{M}, 2)$ Scheme 1 with $\boldsymbol{S} \sim \mathrm{Uniform}(\mathcal{B}_{w,\mathsf{M}-1})$ can also be done by following the same approach as for Theorem 5.[2] However, since the resulting performance is much worse than those of the aforementioned WPIR schemes for the case of $n = 2$ servers, we omit the detailed analysis in this paper.

VI. Numerical Results

We consider the case of 2 servers and compare the achievable 4-tuples $\bigl(\mathsf{R}, \mathsf{U}, \Delta, \rho^{(\mathsf{MI})}\bigr)$ for the $(\mathsf{M}, 2)$ IR schemes proposed in Sections IV-A, IV-C, V-A, and V-B. For the sake of illustration, the information leakage $\rho^{(\mathsf{MI})}$ is normalized by $\log_2 \mathsf{M}$ bits, while the upload cost and access complexity are normalized by $2(\mathsf{M}-1)$ and M, respectively. $2(\mathsf{M}-1)$ and M are the upload cost and access complexity of the PIR capacity-achieving scheme presented in [18] for the case of 2 servers. The upload cost $2(\mathsf{M} - 1)$ is optimal among all so-called *decomposable* PIR capacity-achieving schemes [18].[3]

Fig. 2 illustrates the results of different WPIR schemes for the case of $\mathsf{M} = 32$ files and leakage metric $\rho^{(\mathsf{MI})}$. We can see that Scheme 1 yields the best performance. The IR rate of Scheme 2 with different $\boldsymbol{S}$ is always equal to $1/2$. The results of different WPIR schemes in terms of leakage metric $\rho^{(\mathsf{WIL})}$ will be provided in the extended version of the paper.

VII. Conclusion

We presented the first study of IR schemes with information leakage, which we refer to as WPIR schemes. We proposed two WPIR schemes based on two different PIR protocols and a family of schemes based on partitioning for the case of replication. By relaxing the perfect privacy requirement, we showed that the download rate, the upload cost, and the access complexity can be improved.

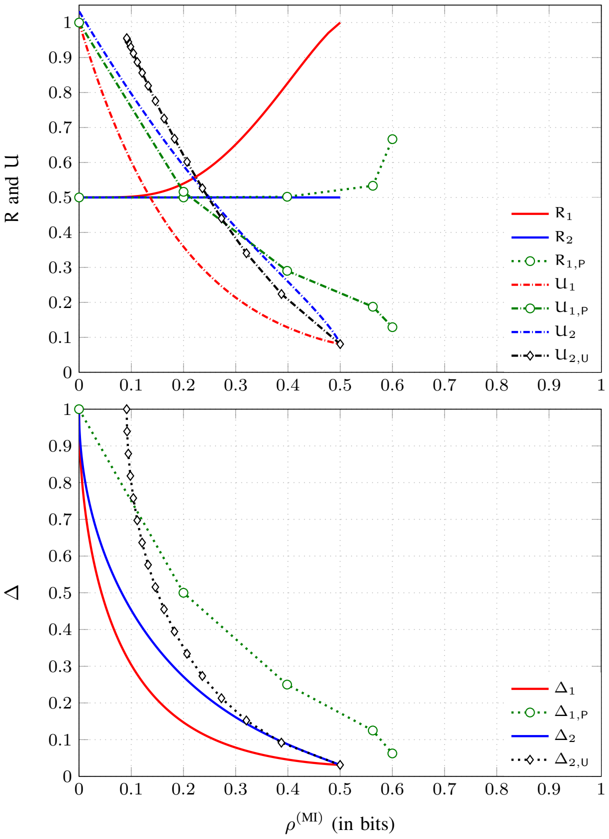

Fig. 2. R, U, and Δ of different WPIR schemes for $\mathsf{M} = 32$, as a function of $\rho^{(\mathsf{MI})}$. For $\mathsf{M} = 32$, $\mathsf{C}_{\mathsf{M},2}$ is almost equal to $1/2$.

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] 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. 43th Annu. IEEE Symp. Found. Comp. Sci. (FOCS)*, Vancouver, Canada, Nov. 16–19, 2002, pp. 261–270.

[3] Z. Dvir and S. Gopi, "2-server PIR with subpolynomial communication," *J. ACM*, vol. 63, no. 4, pp. 39:1–39:15, Nov. 2016.

[4] S. Yekhanin, "Towards 3-query locally decodable codes of subexponential length," *J. ACM*, vol. 55, no. 1, pp. 1:1–1:16, Feb. 2008.

[5] 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.

[6] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, China, Jun. 14–19, 2015, pp. 2852–2856.

[7] 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.

[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 robust private information retrieval with colluding databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.

[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, E. Rosnes, and A. Graell i Amat, "Private information retrieval in distributed storage systems using an arbitrary linear code," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 25–30, 2017, pp. 1421–1425.

[12] S. Kumar, H.-Y. Lin, E. Rosnes, and A. Graell i Amat, "Achieving maximum distance separable private information retrieval capacity with linear codes," 2019, to app. in *IEEE Trans. Inf. Theory*.

[13] Y. Zhang, E. Yaakobi, T. Etzion, and M. Schwartz, "On the access complexity of PIR schemes," to be pres. at *IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 7–12, 2019.

[14] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information: the single server case," in *Proc. 55th Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Oct. 3–6, 2017, pp. 1099–1106.

[15] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 57:1–57:38, Jul. 2018.

[16] B. Köpf and D. Basin, "An information-theoretic model for adaptive side-channel attacks," in *Proc. ACM Conf. Comput. & Commun. Secur.*, Alexandria, VA, USA, Oct. 29–Nov. 2, 2007, pp. 286–296.

[17] R. R. Toledo, G. Danezis, and I. Goldberg, "Lower-cost ϵ-private information retrieval," in *Proc. Privacy Enhancing Technol. Symp. (PETS)*, Darmstadt, Germany, Jul. 19-22, 2016, pp. 184 – 201.

[18] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," Aug. 2018, arXiv:1808.07536v2 [cs.IT].

[2] The scheme is not equal to that of [18] because of the difference in the vector space of the random strategy. The former involves all length-$(\mathsf{M} - 1)$ vectors of weight w, while the latter consists of all vectors of length $\mathsf{M} - 1$.

[3] Based on [18, Def. 2], all existing PIR schemes in the literature are decomposable.