The Capacity of Symmetric Private Information Retrieval with Private Noisy Side Information

Hassan ZivariFard† and Rémi A. Chou‡ and Xiaodong Wang†

†Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
‡Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA
E-mails:{hz2863, xw2008}@columbia.edu, remi.chou@uta.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619115

***Abstract*— Noiseless private side information does not reduce the download cost in Symmetric Private Information Retrieval (SPIR) unless the client knows all but one file. While this is a pessimistic result, we explore in this paper whether noisy private side information available at the client helps decrease the download cost in the context of SPIR with colluding and replicated servers. Specifically, we assume that the client possesses noisy side information about each stored file, which is obtained by passing each file through one of D possible discrete memoryless test channels. The statistics of the test channels are known by the client and by all the servers, but the mapping $\mathcal{M}$ between the files and the test channels is unknown to the servers. We study this problem under two privacy metrics. Under the first metric, the client wants to preserve the privacy of its file selection and the mapping $\mathcal{M}$, and the servers want to preserve the privacy of all the non-selected files. Under the second metric, the client is willing to reveal the index of the test channel that is associated with its desired file. For both privacy metrics, we derive the optimal common randomness and download cost. Our setup generalizes SPIR with colluding servers and SPIR with private noiseless side information. Unlike noiseless side information, our results demonstrate that noisy side information can reduce the download cost, even when the client does not have noiseless knowledge of all but one file.**

I. Introduction

Private Information Retrieval (PIR) consists in privately downloading a file from one or multiple distributed servers and was first introduced in [1], [2]. More recently, the PIR problem has been reformulated by using information-theoretic measures in [3]. The PIR problem only considers the privacy of the file selection with respect to the servers and ignores the privacy of the non-selected files with respect to the client. However, it is often desired that the client does not learn anything beyond the selected file. The resulting setting is referred to as SPIR to emphasize the symmetry of privacy requirements of the client and the servers.

SPIR is also known as oblivious transfer [4], which is a fundamental primitive that generated considerable interest due to its applications in cryptography [5]–[7]. Specifically, SPIR has been shown to be an essential building block of many problems that involve symmetric privacy requirements among participating parties, such as in secure multiparty computation or private set intersection [8], [9].

A. Motivation

It is well known [10, Theorem 2] that private noiseless side information does not help to decrease the download cost in SPIR, except in the special case where the client has noiseless knowledge of all but one file. While this is a pessimistic result, in this paper, we investigate whether noisy side information can help decrease the download cost. Specifically, in this paper, we assume that the client has noisy side information about each file, which contrasts with previous work (reviewed in the next section), where only noiseless side information about the files is considered. The client could have acquired this noisy side information in several ways. For example, the user could have acquired a noisy version of the files opportunistically from other users in its network, overheard them from a wireless noisy broadcast channel, or downloaded them from other servers. Note that the availability of noisy side information encompasses having obtained parts of the files, or even entire files, in a noiseless manner. The noise could also be the result of storing the files for a long period of time.

B. Overview of the problem studied in this paper

In this paper, we study the SPIR problem where a client wishes to retrieve one of the K files that are replicated in N servers and T of these servers may collude. As reviewed in the next section, so far, only SPIR with noiseless side information, where the client possesses a subset of the files, has been studied in the literature. By contrast, in our setting, the client has a noisy version of each file which is modeled by passing each file through a discrete memoryless test channel. We assume that there are $D \leq K$ different test channels whose statistics are public knowledge and known by the client and the servers.[1] We denote the mapping between the files and the test channels by $\mathcal{M}$. We study this problem under two different scenarios. In both of these scenarios, the servers want the client to learn no information about the files that have not been requested. For the first scenario, the client wants to keep the index of the desired file and the entire mapping $\mathcal{M}$ secret from the servers. Note that when the client has a subset of the files as side information [10], [11], then not keeping private the mapping between the test channels and the files may reveal to the servers the indices of the files that are available as side information at the client. For the second scenario, the client aims to keep the index of the desired file and the mapping $\mathcal{M}$ secret from the servers, but the client is willing to reveal the index of the test channel that is associated with the desired file, i.e., $\mathcal{M}(Z)$, where Z is the index of the selected file. For both scenarios, we derive the optimal normalized download

The work of H. ZivariFard and X. Wang is supported in part by the U.S. Office of Naval Research (ONR) under grant N000142412212 and the work of R. Chou is supported in part by NSF grant CCF-2401373.

[1]We assume that the statistics of the D test channels are public information. Note that the client has side information for each of the K files, which can potentially be $\emptyset$, consequently, to model the side information available at the client no more than K test channels are needed.

cost and show that the second privacy metric can lead to a lower download cost, for example when the desired file is included in the side information and the client does not need to download anything.

C. State of the art related to our model

We restrict our discussion to the SPIR models most related to our setting. The capacity of the SPIR problem is derived in [12], which shows that the capacity of the SPIR is smaller than the PIR capacity. Indeed, the SPIR problem has one additional security constraint compared to the PIR problem. Subsequently, this problem was extended to various scenarios, as reviewed in detail in [13]. For instance, SPIR from MDS coded database with non-colluding and colluding servers is studied in [14]. The SPIR with colluding servers in the presence of Byzantine adversaries and eavesdroppers is studied in [15]. The SPIR problem when the client knows part of the common randomness shared between the servers is studied in [16]. The SPIR problem with private noiseless side information, where the client knows a subset of the files in a noiseless manner and wants to keep the identity of these files private, is studied in [10]. More specifically, [10] shows that private noiseless side information does not help to increase the capacity of the SPIR problem, compared to the case where the client has no side information unless the client has noiseless knowledge of all but one file.

Note that the related PIR problem has also been studied when side information is available at the client. In particular, the PIR problem with private and non-private noiseless side information is studied in [10], [11], [17]–[24]. Additionally, the PIR problem with private noisy side information in which the client has a noisy version of each file is studied in [25], [26].

D. Main differences between our setting and previous settings

In our SPIR setting, the client has access to private noisy side information about each file. Note that the side information in the SPIR problem studied in [10] is noiseless, in the sense that the side information at the client corresponds to a subset of the files and the client knows which files are perfectly available as side information. By contrast to [10], in this paper, side information is noisy and, for instance, if the files are binary and the test channels are Binary Symmetric Channels (BSCs), then the client does not know which information bits have been flipped by the BSCs and which ones have not been flipped. Note that since we consider noisy side information that is generated by passing the files through some Discrete Memoryless Channels (DMCs), our problem setup recovers all previous works by considering test channels that are Binary Erasure Channels (BECs). Indeed, passing a file through a BEC with parameter 0 means that the side information is equal to the input file, and passing a file through a BEC with parameter 1 means that no side information is available for this file. For example, when there is only one BEC with parameter 1, then our setting recovers the SPIR problem with colluding servers in [14]. If we assume that there are two BECs with parameters 0 and 1, then our setting recovers the SPIR problem with colluding servers and private noiseless side information [10, Section II.B].

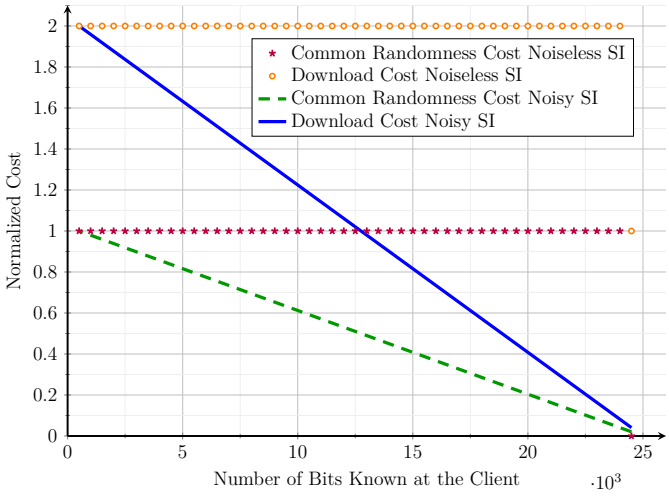

Fig. 1. The optimal normalized download cost and the optimal normalized common randomness cost when $K = 50$, $N = 2$, $T = 1$ and the files' length is $n = 500$, plotted as a function of the number of bits that are known as side information at the client. The plot for the noiseless side information case is obtained from the results in [10, Theorem 2], when the side information is a subset of the files. The plot for the noisy side information case is obtained as the output of a BEC with parameter zero for one file and with parameter ϵ_2 varying between $\frac{1}{K-1}$ and 1 for all the other files.

Recall that, except for the special case where the client has noiseless knowledge of all but one file, noiseless side information does not help to reduce the download cost for the SPIR problem, compared to the case where the client has no side information [10, Theorem 2]. By considering a more general SPIR problem setup, where noisy side information about each stored file is available at the client, our results demonstrate that this noisy side information can be leveraged to decrease the download cost. For example, the optimal normalized download cost of the T-colluding SPIR problem without any side information when all the files have the same length, say n bits, is $\left(1-\frac{T}{N}\right)^{-1}$ and the servers need common randomness with rate $\frac{T}{N-T}$ [14], which is also, the optimal normalized download cost of the T-colluding SPIR problem when the client knows $K-2$ files as side information [10, Theorem 2]. Hence, having $(K-2)\times n$ bits of side information about the files does not help to decrease the download cost. By contrast, when the side information about each file is a noisy version of that file with erasure probability ϵ_1 or ϵ_2, where $\epsilon_1 < \epsilon_2 < 1$, our results show that the optimal normalized download cost is $\left(1-\frac{T}{N}\right)^{-1}\epsilon_2$ and the servers need common randomness with rate $\frac{T}{N-T}\epsilon_2$. As an example, let $\epsilon_1 = 0$ and only one file is passed through this binary erasure test channel. Then, our results show that the noisy side information can lower the normalized download cost from $\left(1-\frac{T}{N}\right)^{-1}$ to $\left(1-\frac{T}{N}\right)^{-1}\epsilon_2$, even though the client only has $n + (1-\epsilon_2)\times(K-1)\times n$ bits of noisy side information instead of $(K-2)\times n$ bits of noiseless side information. We illustrate this example in Fig. 1. In this figure, the x-axis represents the number of bits that the client knows. In the noiseless side information case, the unit is the length of one file such that the first points on the left for the download cost and the common randomness cost correspond to the case where the client knows exactly one file and the

last point corresponds to the case where the client knows $K-1$ files. For the noisy side information case, the client knows $n+(1-\epsilon_2)(K-1)n \in [n,(K-1)n]$ bits with ϵ_2 varying between $\frac{1}{K-1}$ and 1. As seen in Fig. 1, the download cost for the noisy side information is always lower than the download cost for the noiseless side information. Also, the common randomness cost for the noisy side information is always lower than the common randomness cost for noiseless side information, except for the special case where the client knows $K-1$ files as side information. This demonstrates that, unlike noiseless side information, noisy side information can help decrease the download cost and the common randomness cost. Our main results generalize this example to arbitrary noisy side information beyond the erasure model. We also study the case where the client may reveal the index of the test channel associated with its desired file, and show that this can further reduce the download cost.

E. Notation

Let $\mathbb{N}_*$ be the set of positive natural numbers, and $\mathbb{R}_+$ be the set of positive real numbers. For any $a,b \in \mathbb{R}_+$, define $[a:b] \triangleq [\lfloor a \rfloor, \lceil b \rceil] \cap \mathbb{N}_*$ and $[a] \triangleq [1:a]$. Random variables are denoted by capital letters and their realizations by lowercase letters. Vectors and matrices are denoted by boldface letters, e.g., $\mathbf{X}^n = [X_1, X_2, \cdots, X_n]$. For a set of indices $\mathcal{I} \subset \mathbb{N}_*$, $\mathbf{X}_{\mathcal{I}}$ denotes $(X_i)_{i\in\mathcal{I}}$. $\mathbb{E}_X[\cdot]$ is the expectation with respect to the random variable X. The cardinality of a set is denoted by $|\cdot|$. For a mapping $\boldsymbol{\mathcal{M}} : \mathcal{X} \to \mathcal{Y}$, the preimage of $y \in \mathcal{Y}$ by $\boldsymbol{\mathcal{M}}$ is denoted as $\boldsymbol{\mathcal{M}}^{-1}(y) \triangleq \{x \in \mathcal{X} : \boldsymbol{\mathcal{M}}(x) = y\}$. For $D \in \mathbb{N}_*$ and a mapping $\boldsymbol{\mathcal{M}} : [D] \to \mathbb{R}_+$, we represent the domain and co-domain of $\boldsymbol{\mathcal{M}}$ as a matrix of dimension $2 \times D$ as $\boldsymbol{\mathcal{M}} = \left(\begin{smallmatrix} 1 & 2 & \cdots & D \\ \mathcal{M}(1) & \mathcal{M}(2) & \cdots & \mathcal{M}(D) \end{smallmatrix}\right)$.

II. Problem Statement

Consider a client and N servers, where up to T of these N servers may collude, and each server has a copy of K files of length n. Additionally, consider a set of D test channels, whose transition probabilities are known to the client and the servers, and whose outputs take value in finite alphabets. We assume that the client has noisy side information about all the K files in the sense that each file is passed through one of the D test channels, and the output of this test channel is available at the client. The mapping $\boldsymbol{\mathcal{M}}$ between the files and the test channels is not known at the servers. The objective of the client is to retrieve one of the files such that the index of this file and the mapping $\boldsymbol{\mathcal{M}}$ are kept secret from the servers.

A. Definitions

We now define the problem formally and provide some intuitions about our problem setup.

Definition 1. *Consider $K, n, N, T, D \in \mathbb{N}_*$, $(d_i)_{i\in[D]} \in \mathbb{N}_*^D$ such that $\sum_{i=1}^D d_i = K$, and D distinct test channels $\left(C^{(i)}\right)_{i\in[D]}$, with $C^{(i)} \triangleq (\mathcal{X}, P_{Y|X}^{(i)}, \mathcal{Y}_i)$, where $\mathcal{X}$ and $\mathcal{Y}_i$, $i \in [K]$, are finite alphabets. Without loss of generality, assume that $H(X|Y_i) \le H(X|Y_j)$, for $i,j \in [D]$ such that $i<j$, where X is uniformly distributed over $\mathcal{X}$ and Y_i and Y_j are the outputs of $C^{(i)}$ and $C^{(j)}$, respectively, when X is the input.*

A SPIR protocol with private noisy side information and parameters $\left(K, n, N, T, D, (d_i)_{i\in[D]}, \left(C^{(i)}\right)_{i\in[D]}\right)$ consists of,

- *N servers where up to T of them may collude;*
- *K independent random sequences $\mathbf{X}_{[K]}^n \triangleq (\mathbf{X}_i^n)_{i\in[K]}$ each of length n that are uniformly distributed over $\mathcal{X}^n$, which represent K files replicated at each of the N servers;*
- *common randomness $\mathbf{U} \in \boldsymbol{\mathcal{U}}$ shared between the servers but not the client; we also call normalized randomness cost the following quantity $R_0 \triangleq \frac{H(\mathbf{U})}{\max_{i\in[K]} H(X_i)} = \frac{H(\mathbf{U})}{n}$;*
- *local randomness $\mathbf{L} \in \boldsymbol{\mathcal{L}}$ at the client;*
- *D distinct test channels $\left(C^{(i)}\right)_{i\in[D]}$;*
- *a mapping $\boldsymbol{\mathcal{M}}$ chosen at random from the set $\mathfrak{M} \triangleq \left\{\boldsymbol{\mathcal{M}} : [K] \to [D] : \forall i \in [D], \left|\boldsymbol{\mathcal{M}}^{-1}(i)\right| = d_i\right\}$; this mapping is only known at the client and not at the servers;*
- *for each file $\mathbf{X}_i^n$, where $i \in [K]$, the client has access to a noisy version of $\mathbf{X}_i^n$, denoted by $Y_{i,\boldsymbol{\mathcal{M}}(i)}^n$, which is the output of the test channel $\mathrm{C}^{(\boldsymbol{\mathcal{M}}(i))}$ when $\mathbf{X}_i^n$ is the input;*
- *the random variable Z is uniformly distributed over $[K]$ and represents the index of the file that the client wishes to retrieve, i.e., the client wants to retrieve the file $\mathbf{X}_Z^n$.*
- *for $i \in [N]$, a query function $\mathcal{F}_i : [K] \times \boldsymbol{\mathcal{L}} \times \mathfrak{M} \times \mathcal{Y}_{[K]}^n \to \boldsymbol{\mathcal{Q}}_i$, where $\boldsymbol{\mathcal{Q}}_i$ is a finite alphabet;*
- *for $i \in [N]$, an answer function $\mathcal{E}_i : \boldsymbol{\mathcal{Q}}_i \times \boldsymbol{\mathcal{U}} \times \mathcal{X}^{nK} \to \left[2^{nR(\mathbf{Q}_i)}\right]$;*
- *a decoding function $\mathcal{D} : [K] \times \boldsymbol{\mathcal{L}} \times \mathfrak{M} \times \left[2^{n\sum_{i=1}^N R(\mathbf{Q}_i)}\right] \times \mathcal{Y}^{nK} \to \mathcal{X}^n$;*

and operates as follows,

1) *the client creates the queries $\mathbf{Q}_i \triangleq \mathcal{F}_i\left(Z, \mathbf{L}, \boldsymbol{\mathcal{M}}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n\right)$, where $\mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n \triangleq \left(Y_{i,\boldsymbol{\mathcal{M}}(i)}^n\right)_{i\in[K]}$, and sends it to Server $i \in [N]$; with $\mathbf{Q}_{[N]} \triangleq (\mathbf{Q}_i)_{i\in[N]}$, note that $H\left(\mathbf{Q}_{[N]} \middle| \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n, \mathbf{L}, \boldsymbol{\mathcal{M}}, Z\right) = 0$;*
2) *then, for all $i \in [N]$, Server i creates the answer $\mathbf{A}_i \triangleq \mathcal{E}_i\left(\mathbf{Q}_i, U, \mathbf{X}_{[K]}^n\right)$ and sends it to the client; note that $H\left(\mathbf{A}_i \middle| \mathbf{Q}_i, \mathbf{U}, \mathbf{X}_{[K]}^n\right) = 0, \quad \forall i \in [N]$;*
3) *finally, the client computes an estimate of $\mathbf{X}_Z^n$ as $\widehat{\mathbf{X}}_Z^n \triangleq \mathcal{D}\left(Z, \mathbf{L}, \boldsymbol{\mathcal{M}}, \mathbf{A}_{[N]}, \mathbf{Y}_{[K],\boldsymbol{\mathcal{M}}}^n\right)$, where $\mathbf{A}_{[N]} \triangleq (\mathbf{A}_i)_{i\in[N]}$.*

Hence, the probability of error at the client is, $P_e \triangleq \limsup_{n\to\infty} \mathbb{P}\left[\widehat{\mathbf{X}}_Z^n \neq \mathbf{X}_Z^n\right]$. $R\left(\mathbf{Q}_{[N]}\right) \triangleq \sum_{i=1}^N R\left(\mathbf{Q}_i\right)$ is the normalized download cost of the SPIR protocol and is random with respect to $\mathbf{Q}_{[N]}$, which makes the protocol a variable length coding scheme. We also define the expected normalized download cost of the protocol as $R \triangleq \mathbb{E}_{\mathbf{Q}_{[N]}}[R\left(\mathbf{Q}_{[N]}\right)]$. Note that the random variables $(Z, \mathbf{L}, \boldsymbol{\mathcal{M}}, \mathbf{U}, \mathbf{X}_{[K]}^n)$ are independent.

We then consider two privacy metrics in Definitions 2 and 3. For both metrics, all the files in the servers except the desired file must remain secret from the client. However, for the first metric, the index of the desired file Z and the mapping $\boldsymbol{\mathcal{M}}$ must remain private from the servers, whereas, for the second metric, the index $\boldsymbol{\mathcal{M}}(Z)$ is allowed to be revealed to the

servers. As discussed next in Section III, these two privacy metrics recover, as special cases, the SPIR settings previously studied in the literature.

Definition 2 (Cost region $\mathrm{C}_{\text{SPIR-PNSI}}$ for undisclosed side information statistics of the desired file)**.** *An expected normalized download cost R and normalized randomness cost R_0, are achievable with undisclosed side information statistics of the desired file, if there exist SPIR protocols with private noisy side information such that, for any set $\mathcal{T} \subseteq [N]$ of colluding servers such that $|\mathcal{T}| = T$, we have*

$$P_e = 0, \tag{1a}$$
$$I\big(\mathbf{Q}_{\mathcal{T}}, \mathbf{A}_{\mathcal{T}}, \mathbf{X}^n_{[K]}, \mathbf{U}; Z, \boldsymbol{\mathcal{M}}\big) = 0, \tag{1b}$$
$$I\big(\mathbf{X}^n_{[K]\setminus\{Z\}}; \mathbf{A}_{[N]} \big| Z, \mathbf{L}, \boldsymbol{\mathcal{M}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\big) = 0. \tag{1c}$$

The closure of the set of achievable cost pairs (R, R_0) is referred to as the SPIR optimal cost region with private noisy side information and undisclosed side information statistics of the desired file, and is denoted by $\mathrm{C}_{\text{SPIR-PNSI}}$.

Definition 3 (Cost region $\mathrm{C}^*_{\text{SPIR-PNSI}}$ for disclosed side information statistics of the desired file)**.** *An expected normalized download cost R and normalized randomness cost R_0, are achievable with disclosed side information statistics of the desired file, if there exist SPIR protocols with private noisy side information such that, for any set $\mathcal{T} \subseteq [N]$ of colluding servers such that $|\mathcal{T}| = T$, we have*

$$P_e = 0, \tag{2a}$$
$$I\big(\mathbf{Q}_{\mathcal{T}}, \mathbf{A}_{\mathcal{T}}, \mathbf{X}^n_{[K]}, \mathbf{U}; Z, \boldsymbol{\mathcal{M}} \big| \boldsymbol{\mathcal{M}}(Z)\big) = 0, \tag{2b}$$
$$I\big(\mathbf{X}^n_{[K]\setminus\{Z\}}; \mathbf{A}_{[N]} \big| Z, \mathbf{L}, \boldsymbol{\mathcal{M}}, \mathbf{Y}^n_{[K],\boldsymbol{\mathcal{M}}}\big) = 0. \tag{2c}$$

The closure of the set of achievable cost pairs (R, R_0) is referred to as the SPIR optimal cost region with private noisy side information and disclosed side information statistics of the desired file, and is denoted by $\mathrm{C}^*_{\text{SPIR-PNSI}}$. Note that under the privacy metric (2b)*, we have $H\big(\mathbf{A}_i \big| \mathbf{Q}_i, \mathbf{U}, \mathbf{X}^n_{[K]}, \boldsymbol{\mathcal{M}}(Z)\big) = 0, \quad \forall i \in [N]$.*

Observe that the privacy constraints in Definition 2 implies the privacy constraints in Definition 3, i.e., (1b)⇒(2b), and we will show that revealing the index of the test channel associated with the desired file $\mathbf{X}^n_Z$ can result in a strictly lower download rate.

III. Main Results

In the following, we assume that at least one file is not noiselessly known by the client, otherwise, the client would know the entire database. We also define the following event $\mathcal{E} \triangleq \{(K-1)$ files are noiselessly known at the client$\}$, which, with our notation, means that $d_1 = K-1$ and $C^{(1)}$ is a BEC with parameter zero. Finally, we define the following indicator function

$$\mathbb{1}\{\mathcal{E}^c\} \triangleq \begin{cases} 1 & \text{if Event } \mathcal{E} \text{ is false} \\ 0 & \text{if Event } \mathcal{E} \text{ is true} \end{cases}. \tag{3}$$

Theorem 1. *Consider $K \geq 2$ files that are replicated in $N \geq 2$ servers, where up to T of them may collude. Then,*

$$\mathrm{C}_{\text{SPIR-PNSI}} = \left\{ \begin{array}{l} (R, R_0): \\ R \geq \big(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\big)^{-1} H(X_1|Y_{1,D}) \\ R_0 \geq \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N-T} H(X_1|Y_{1,D}) \end{array} \right\}.$$

Proof. The achievability proof of Theorem 1 is based on source coding and the achievability scheme in [14, Section VI.B]. As discussed in Definition 1, we assume that the test channels are ordered in the sense that $H(X|Y_i) \leq H(X|Y_j)$, for $i, j \in [D]$ such that $i < j$, where X is distributed uniformly over $\mathcal{X}$ and Y_i and Y_j are the outputs of the test channels $C^{(i)}$ and $C^{(j)}$, respectively, when the channel input is X. In our scheme, the servers store the source code of all the files in a new database denoted by $\tilde{\mathbf{X}}^n_{[K],D}$, which allows the recovery of $\mathbf{X}^n_{[K]}$ when the available side information at the decoder is obtained by passing all the files through the noisiest test channel, i.e., $C^{(D)}$. Then, the client retrieves a source coded version of the desired file from $\tilde{\mathbf{X}}^n_{[K],D}$ by using the scheme in [14, Section VI.B]. According to the Slepian-Wolf theorem [27], [28], by accessing the side information and the source-coded version of the desired file the client will be able to retrieve the desired file. Our converse proof shows that this scheme is optimal. Details for the achievability and converse proofs of Theorem 1 can be found in [29, Section IV]. □

Remark 1 (Index of the random variables)**.** *Since all the files are generated according to the uniform distribution over $\mathcal{X}^n$, one can change the index of X_1 and $Y_{1,D}$ in Theorem 1 to X_i and $Y_{i,D}$, for $i \in [K]$.*

Remark 2. *Note that the optimal normalized download cost for PIR with noisy side information in [26, Theorem 1] depends on the conditional entropy of all the test channels, whereas our optimal result in Theorem 1 only depends on the conditional entropy of the noisiest test channel.*

Corollary 1 (Binary erasure test channels)**.** *Consider $K \geq 2$ files that are replicated in $N \geq 2$ servers and up to T of these servers may collude. Furthermore, let the test channels be BECs with parameters $(\epsilon_j)_{j\in[D]} \in [0,1]^D$ such that $\epsilon_j < \epsilon_{j'}$ for $j, j' \in \mathbb{N}_+$ and $j < j'$. Then,*

$$\mathrm{C}_{\text{SPIR-PNSI}} = \left\{ \begin{array}{l} (R, R_0): \\ R \geq \big(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\big)^{-1} \epsilon_D \\ R_0 \geq \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N-T}\epsilon_D \end{array} \right\}.$$

Example 1 (SPIR without side information)**.** *If we set $D = 1$ and $\epsilon_1 = 1$ in Corollary 1, which corresponds to the case where there is no side information about any files, then the optimal cost region in Corollary 1 reduces to [14, Theorem 2], i.e.,*

$$\mathrm{C}_{\text{SPIR-PNSI}} = \left\{ \begin{array}{l} (R, R_0): \\ R \geq \big(1 - \frac{T}{N}\big)^{-1} \\ R_0 \geq \frac{T}{N-T} \end{array} \right\}.$$

Example 2 (SPIR with noiseless side information)**.** *If we set $D = 2$, $\epsilon_1 = 0$, and $\epsilon_2 = 1$ in Corollary 1, which corresponds to the case where the client has access to d_1 files in a noiseless*

manner as side information and has no side information about the remaining $d_2 = K - d_1$ files, then $\mathrm{C}_{\text{SPIR-PNSI}}$ in Corollary 1 reduces to [10, Theorem 2], i.e.,

$$\mathrm{C}_{\text{SPIR-PNSI}} = \left\{ \begin{matrix} (R, R_0): \\ R \geq \left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} \\ R_0 \geq \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N-T} \end{matrix} \right\}.$$

Theorem 2. *Consider $K \geq 2$ files that are replicated in $N \geq 2$ servers where up to T of them may collude. Then, the optimal cost region $\mathrm{C}^*_{\text{SPIR-PNSI}}$ of the SPIR with private noisy side information and disclosed side information statistics of the desired file is*

$$\mathrm{C}^*_{\text{SPIR-PNSI}} = \left\{ \begin{matrix} (R, R_0): \\ R \geq \mathbb{E}_V[R(V)] = \frac{1}{K}\left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} \sum_{v=1}^{D} d_v H(X_1|Y_{1,v}) \\ R_0 \geq \mathbb{E}_V[R_0(V)] = \mathbb{1}\{\mathcal{E}^c\}\frac{T}{K(N-T)} \sum_{v=1}^{D} d_v H(X_1|Y_{1,v}) \end{matrix} \right\} \tag{4a}$$

where,

$$R(V) \triangleq \left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} H(X_1|Y_{1,V}), \tag{4b}$$

$$R_0(V) \triangleq \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N-T} H(X_1|Y_{1,V}), \tag{4c}$$

and V is distributed according to $\mathbb{P}[V = v] \triangleq \frac{d_v}{K}$, for $v \in [D]$.

Proof. Similar to the achievability proof of Theorem 1, the achievability proof of Theorem 2 is based on source coding and the achievability scheme in [14, Section VI.B]. Specifically, we use the same achievability scheme as that of Theorem 1 by generating the source code of the files when the side information of the files is generated according to $C^{(\boldsymbol{\mathcal{M}}(Z))}$ instead of $C^{(D)}$. However, the converse proof of Theorem 2 does not follow from the converse proof of Theorem 1. One of the main differences between the two converse proofs is that when the index Z of the desired file and the index of the test channel that is associated with the desired file are fixed, i.e., $\boldsymbol{\mathcal{M}}(Z) = d$ for some $d \in [D]$ and $Z = z$ for some $z \in [K]$, changing the index z into another index $z' \neq z$ may not be possible as one may have $\boldsymbol{\mathcal{M}}(z') \neq d$, which complexifies the proof. The details of the achievability and converse proofs are available in [29, Section V]. □

Corollary 2 (Binary erasure test channels)**.** *Consider $K \geq 2$ files that are replicated in $N \geq 2$ servers and up to T of these servers may collude. Furthermore, let the test channels be BECs with parameters $(\epsilon_j)_{j\in[D]} \in [0,1]^D$ such that $\epsilon_j < \epsilon_{j'}$ for $j, j' \in \mathbb{N}_+$ and $j < j'$. Then,*

$$\mathrm{C}^*_{\text{SPIR-PNSI}} = \left\{ \begin{matrix} (R, R_0): \\ R \geq \mathbb{E}_V[R(V)] = \frac{1}{K}\left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} \sum_{v=1}^{D} d_v \epsilon_v \\ R_0 \geq \mathbb{E}_V[R_0(V)] = \mathbb{1}\{\mathcal{E}^c\}\frac{T}{K(N-T)} \sum_{v=1}^{D} d_v \epsilon_v \end{matrix} \right\},$$

where

$$R(V) \triangleq \left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} \epsilon_V,\ R_0(V) \triangleq \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N-T}\epsilon_V,$$

and V is distributed according to $\mathbb{P}[V = v] \triangleq \frac{d_v}{K}$, for $v \in [D]$.

Example 3 (SPIR with noiseless side information)**.** *If we set $D = 2$, $\epsilon_1 = 0$, and $\epsilon_2 = 1$ in Corollary 2, which corresponds to the case where the client has access to d_1 files in a noiseless manner as side information and has no side information about the remaining $d_2 = K - d_1$ files, then when $\boldsymbol{\mathcal{M}}(Z) = 1$ the optimal download cost $R(V)$ in Corollary 2 is zero. When $\boldsymbol{\mathcal{M}}(Z) = 2$ the optimal download cost in Corollary 2 reduces to [10, Theorem 2], i.e.,*

$$\left\{ \begin{matrix} (R, R_0): \\ R \geq \left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} \\ R_0 \geq \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N-T} \end{matrix} \right\}.$$

Example 4 (When there is only one test channel)**.** *If $D = 1$, and therefore $d_1 = K$, then the optimal download cost in Theorem 1 and Theorem 2 reduces to*

$$\mathrm{C}^*_{\text{SPIR-PNSI}} = \left\{ \begin{matrix} (R, R_0): \\ R \geq \left(1 - \frac{T}{N}\right)^{-1} H(X_1|Y_{1,1}) \\ R_0 \geq \frac{T}{N-T} H(X_1|Y_{1,1}) \end{matrix} \right\}.$$

Remark 3 (Comparing Theorems 1 and 2)**.** *The difference between the normalized download cost in Theorem 1 and Theorem 2 is*

$$\left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} \left(H(X_1|Y_{1,D}) - \frac{1}{K}\sum_{v=1}^{D} d_v H(X_1|Y_{1,v})\right)$$
$$= \frac{1}{K}\left(1 - \mathbb{1}\{\mathcal{E}^c\}\frac{T}{N}\right)^{-1} \sum_{v=1}^{D} d_v \big(H(X_1|Y_{1,D}) - H(X_1|Y_{1,v})\big)$$

Therefore, the optimal normalized download cost in Theorem 2 is always smaller than or equal to the optimal normalized download cost in Theorem 1. Similarly, the difference between the optimal common randomness cost in Theorem 1 and Theorem 2 is

$$\mathbb{1}\{\mathcal{E}^c\}\frac{T}{K(N-T)} \sum_{v=1}^{D} d_v \Big(H(X_1|Y_{1,D}) - H(X_1|Y_{1,v})\Big),$$

which shows that the optimal common randomness cost in Theorem 2 is always smaller than or equal to the optimal common randomness cost in Theorem 1. Hence, $\mathrm{C}_{\text{SPIR-PNSI}} \subseteq \mathrm{C}^*_{\text{SPIR-PNSI}}$, which shows that disclosing the index of the test channel associated with the desired file leads to a smaller download cost and a smaller common randomness cost.

Similar to the above, one can show that the difference between (4b) *and the optimal download cost in Theorem 1, and the difference between* (4c) *and the optimal common randomness cost in Theorem 1, are non-negative and increase as the index of the test channel associated with the desired file, i.e., $\boldsymbol{\mathcal{M}}(Z)$, decreases.*

REFERENCES

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. 36th Annu. Symp. Found. Comput. Sci.*, 1995, pp. 41–50.

[2] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *Journal of the ACM*, vol. 45, no. 6, pp. 965–981, Nov. 1998.

[3] 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.

[4] M. O. Rabin, "How to exchange secrets with oblivious transfer," *Aiken Comput. Lab, Harvard Univ., Cambridge, MA, USA, Tech. Rep., TR-81*, May 1981.

[5] Y. Ishai, E. Prabhakaran, and A. Sahai, "Founding cryptography on oblivious transfer–efficiency," in *Proc. Annu. Int. in Criptol. Conf.*, Berlin, Germany: Springer, 2008, pp. 572–591.

[6] R. Ahlswede and I. Csiszár, "On oblivious transfer capacity," in *Information Theory, Combinatorics, and Search Theory*, Springer, 2013, pp. 145–166.

[7] A. C. A. Nascimento and A. Winter, "On oblivious-transfer capacity of noisy resources," *IEEE Trans. Inf. Theory*, vol. 54, no. 6, pp. 2572–2581, Jun. 2008.

[8] Z. Wang, K. Banawan, and S. Ulukus, "Multi-party private set intersection: An information-theoretic approach," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 366–379, Mar. 2021.

[9] ——, "Private set intersection: A multi-message symmetric private information retrieval perspective," *IEEE Trans. Inf. Theory*, vol. 68, no. 3, pp. 2001–2019, Mar. 2022.

[10] Z. Chen, Z. Wang, and S. A. Jafar, "The capacity of T-private information retrieval with private side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4761–4773, Aug. 2020.

[11] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2032–2043, Apr. 2020.

[12] 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.

[13] S. Ulukus, S. Avestimehr, S. A. Gastpar, M Jafar, R. Tandon, and C. Tian, "Private retrieval, computing, and learning: Recent progress and future challenges," *IEEE J. Sel. Areas Inf. Theory*, vol. 40, no. 3, pp. 729–748, Mar. 2022.

[14] Q. Wang and M. Skoglund, "Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5160–5175, Aug. 2019.

[15] ——, "On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3183–3197, May 2019.

[16] Z. Wang and S. Ulukus, "Symmetric private information retrieval at the private information retrieval rate," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 2, pp. 350–361, Jun. 2022.

[17] R. Tandon, "The capacity of cache aided private information retrieval," in *Proc. 55th Annu. Allerton Conf. on Commun., Control, and Comput. (Allerton)*, Monticello, IL, Oct. 2017, pp. 1078–1082.

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

[19] Y.-P. Wei and S. Ulukus, "The capacity of private information retrieval with private side information under storage constraints," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2023–2031, Apr. 2020.

[20] S. Li and M. Gastpar, "Converse for multi-server single-message PIR with side information," in *Proc. 54th Annual Conf. on Info. Sciences and Systems (CISS)*, 2020, pp. 1–6.

[21] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "The role of coded side information in single-server private information retrieval," *IEEE Trans. Inf. Theory*, vol. 67, no. 1, pp. 25–44, Jan. 2021.

[22] M. Jafari Siavoshani, S. P. Shariatpanahi, and M. A. Maddah-Ali, "Private information retrieval for a multi-message scenario with private side information," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3235–3244, May 2021.

[23] B. Herren, A. Arafa, and K. Banawan, "Download cost of private updating," in *Proc. IEEE Int. Conf. on Comm. (ICC)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6.

[24] Y. Lu and S. A. Jafar, "On single server private information retrieval with private coded side information," *IEEE Trans. Inf. Theory*, vol. 69, no. 5, pp. 3263–3284, May 2023.

[25] H. ZivariFard and R. A. Chou, "Private information retrieval when private noisy side information is available," in *Proc. IEEE Int. Symp. on Info. Theory (ISIT)*, Taipei, Taiwan, Jul. 2023, pp. 1–6.

[26] ——, "Private information retrieval with private noisy side information," *IEEE Trans. Inf. Theory*, vol. 70, no. 4, pp. 2886–2902, Apr. 2024.

[27] D. Slepian and J. K. Wolf, "A coding theorem for multiple access channels with correlated sources," *Bell System Technical Journal*, vol. 52, no. 7, pp. 1037–1076, Sep. 1973.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., 2006.

[29] H. ZivariFard, R. A. Chou, and X. Wang, "Private noisy side information helps to increase the capacity of SPIR," *submitted to IEEE Trans. Inf. Theory*, Nov. 2023.