On the Information Leakage in Private Information Retrieval Systems

Tao Guo[ID], *Member, IEEE*, Ruida Zhou, *Student Member, IEEE*, and Chao Tian[ID], *Senior Member, IEEE*

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

Index Terms—Privacy, information leakage, encryption.

I. INTRODUCTION

THE problem of *private information retrieval* (PIR) [1], [2] addresses the retrieval of one out of K messages from N replicated databases, without revealing the identity of the desired message to any individual database. A trivial solution is to simply retrieve all the messages, however, the download cost in this solution is unrealistic. Therefore, in the PIR problem, the goal is to find an efficient protocol, i.e., with the minimum download cost, to privately retrieve the desired message. The capacity of a PIR system is defined as the maximum number of bits of desired message that can be retrieved per bit of downloaded information, which was shown in [3] to be $C_{\text{PIR}} = \left(1 + 1/N + 1/N^2 + \cdots + 1/N^{K-1}\right)^{-1}$. There were also significant interests recently on various other variations of the PIR problem; see, e.g., [4]–[28].

The problem considered in this work is closely related to the *symmetric* private information retrieval (SPIR) problem [5], where "symmetric" refers to the fact that both user privacy and database privacy need to be preserved. Database privacy requires that the user obtains no information on other messages beyond the requested message; strictly speaking, this

is a security requirement rather than a privacy requirement, and we shall refer to it as such in the sequel. It is well understood that SPIR can also be viewed as a multi-database oblivious transfer problem [29]–[32]. It was shown in [5] that to ensure perfect security, the databases need to share *common randomness* (a common key) which is independent of the messages and only available to the databases. The common key is used for randomizing the answers such that no information about the non-desired messages is leaked. The capacity of SPIR was shown to be $C_{\text{SPIR}} = 1 - 1/N = \left(1 + 1/N + 1/N^2 + \cdots + 1/N^\infty\right)^{-1}$, as long as the amount of common randomness is at least $\frac{1}{N-1}$ bits per desired message bit.

In practice, the requirement of perfect security may be too stringent. For example, in the SPIR setting, it is likely to be acceptable for a user to obtain a single byte of information in a 1-hour video data (assuming the message is a video) that he does not request, i.e., leaking a very small amount of information [33]–[35]. This is the type of systems that we wish to understand in this work. Allowing a small amount of information leakage let us control the system security level in a finer grain manner. In this context, SPIR essentially requires strictly zero information leakage, while the classical PIR does not have a security constraint at all. Thus our goal is to understand the tradeoff between the download cost and the amount of information leakage in the regime other than these two extreme cases. From the perspective of the oblivious transfer, we are essentially considering a generalized multi-database oblivious transfer setting [32], where one side of the privacy requirements (that corresponding to the database privacy) is allowed to be less than perfect.

Two different notions of information leakage can be defined: the individual leakage is defined as the amount of information that the user can obtain about any individual non-desired message, and the total leakage as the amount of information that the user can obtain about all the non-desired messages. The former is similar to the weak security constraints in the literature [36]–[39], while the latter corresponds to the standard strong security constraints. In practice, the weak security requirement may be more suitable than the strong version in some applications. For example, when the messages are video sequences, the user should not obtain information about any individual video segment, but obtaining the binary XOR of two video sequences may not be an issue since it will not lead to a meaningfully decodable video sequence, and thus enforcing that the user cannot obtain this XOR may

be too stringent. Moreover, a protocol with a weak security constraint can potentially be implemented more efficiently in practical settings and may not require common keys.

The main result of this work is the characterizations of the optimal tradeoffs between the download cost and the amount of information leakage for both individual leakage and total leakage. By adapting the codes given in [4] and [5], we provide code constructions that can achieve these optimal tradeoffs; moreover, they are also shown to have the minimum message sizes. For the individual leakage case, the constructed codes do not require common randomness (unless there are only two messages). At the extreme case with perfect individual message security, the download cost is in fact the same as that with perfect total security. This is rather reassuring since it implies that the stronger security requirement does not induce any additional download cost.

A related problem, namely weakly private information retrieval, was studied in [40]–[42]. For weakly private information retrieval systems, a small amount of privacy leakage is allowed, i.e., the leakage is to the server and on the requested message index, and it differs from the setting we are considering where the information leakage is to the user and on the message content.

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

II. PRELIMINARIES

A. Problem Statement

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

In a *private information retrieval* (PIR) system, there are K independent messages $W_{1:K} = (W_1, W_2, \cdots, W_K)$, each of which is comprised of L i.i.d. symbols uniformly distributed over a finite alphabet $\mathcal{X}$. In the sequel, we will use $\log_{|\mathcal{X}|}$-ary units, instead of bits, as the information measure, unless specified otherwise. We have

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

$$H(W_1) = H(W_2) = \cdots = H(W_K) = L. \quad (2)$$

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

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

where $Q_n^{[k]} \in \mathcal{Q}_n$ for $n \in [1 : N]$. For $n \in [1 : N]$, the n-th query $Q_n^{[k]}$ is sent to the n-th database. Since we wish to protect the non-desired messages from information leakage,

the databases may need to share a common random key $S \in \mathcal{S}$ that is not accessible to the user, which induces the condition

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

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

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

which implies

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

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

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

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

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

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

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

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

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

In contrast to the perfect security of the non-desired messages $W_{\bar{k}} = (W_1, \cdots, W_{k-1}, W_{k+1}, \cdots, W_K)$ in [5], we allow information leakage in this work. Define the total leakage as $I(W_{\bar{k}}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F})$, which is the amount of information that the user can obtain about all the non-desired messages. Define the individual message leakage for message $k' \in [1 : K]$ ($k' \neq k$) as $I(W_{k'}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F})$, which is the amount of information that the user can obtain about the individual non-desired message-k'. The total leakage and the individual leakage in the systems are constrained as follows

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

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

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

In a PIR system, the message size is defined as the number of bits to represent each individual message, which is clearly $L \log_2 |\mathcal{X}|$ in our notation. The download cost is defined as

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

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

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

B. Analysis of Several Existing PIR Codes

Before presenting our main result, we provide the result of a simple analysis on the capacity-achieving PIR code given by Sun and Jafar [3], the capacity-achieving code proposed by Tian *et al.* [4], and the SPIR code [5]; these three classes of codes will be referred to as the SJ code, the TSC code, and the SPIR code, respectively.

- The SJ code [3]:

$$L = N^K; \quad D = \frac{N(N^K - 1)}{N - 1}; \tag{15}$$

$$\frac{1}{L} I(W_{\bar{k}}; A_{1:N}^{[k]}, \mathbf{F}) = \frac{1}{N - 1} \left(1 - \frac{1}{N^{K-1}} \right); \tag{16}$$

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

- The TSC code [4]:

$$L = N - 1; \quad D = \frac{N^K - 1}{N^{K-1}}; \tag{18}$$

$$\frac{1}{L} I(W_{\bar{k}}; A_{1:N}^{[k]}, \mathbf{F}) = \frac{1}{N - 1} \left(1 - \frac{1}{N^{K-1}} \right); \tag{19}$$

$$\frac{1}{L} I(W_{k'}; A_{1:N}^{[k]}, \mathbf{F}) = \frac{1}{N^{K-1}}, \quad \forall k' \neq k \in [1:K]. \tag{20}$$

- The SPIR code [5]:

$$L = N - 1; \quad D = N; \quad \rho = \frac{1}{N - 1}. \tag{21}$$

From (16), (19) and (17), (20), it is seen that the SJ code and TSC code have the same performance in terms of both the total leakage and the individual leakage. As an example, consider the case $N = 3$, $K = 3$, then $\frac{1}{L} I(W_{\bar{k}}; A_{1:N}^{[k]}, Q_{1:N}^{[k]}, \mathbf{F}) = \frac{4}{9}$ and $\frac{1}{L} I(W_{k'}; A_{1:N}^{[k]}, Q_{1:N}^{[k]}, \mathbf{F}) = \frac{1}{9}$ for both the SJ code and the TSC code.

III. MAIN RESULTS

A. Total Leakage

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

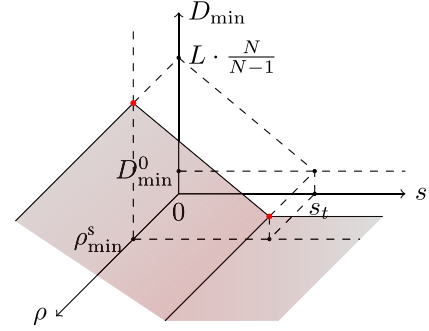$$D_{\min}^0 \triangleq L \cdot \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}} \right). \tag{22}$$

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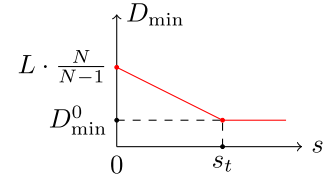

Fig. 2. Tradeoff curve between $D_{\min}$ and s.

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

$$\rho_{\min}^s \triangleq \frac{1}{N - 1} - \frac{N^{K-1}}{N^{K-1} - 1} \cdot s, \tag{23}$$

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

$$D_{\min} = \begin{cases} L \cdot \left(\dfrac{N}{N - 1} - \dfrac{1}{N^{K-1} - 1} \cdot s \right), & \text{if } 0 \leq s \leq s_t \\ D_{\min}^0, & \text{otherwise,} \end{cases} \tag{24}$$

where the threshold is defined by $s_t \triangleq \frac{1}{N-1} \left(1 - \frac{1}{N^{K-1}} \right)$ and $D_{\min}^0$ is defined in (22). If $\rho < \rho_{\min}^s$, then $D_{\min} = \infty$.

Proof: The proof is given in Section IV. □

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

In light of (23) and (24), we can view $D_{\min}$ as a bivariate function of ρ and s. For a given s, the dependency of $D_{\min}$ on ρ only appears at the threshold $\rho_{\min}^s$ so that i) $D_{\min} = \infty$ if $\rho < \rho_{\min}^s$; ii) $D_{\min} < \infty$ and $D_{\min}$ is independent of ρ if $\rho \geq \rho_{\min}^s$. For $\rho \geq \rho_{\min}^s$, the dependency of $D_{\min}$ on (ρ, s) is illustrated in Fig. 1. The dashed lines project the corner points to the axises that show their values. The value of $D_{\min}$ with respective to (ρ, s) is given by the shaded area. We see that for a given value of s, $D_{\min}$ is a constant and thus independent with ρ. The shaded area is projected onto the $D_{\min}$-s plane, which is drawn in Fig. 2.

The red line in Fig. 2 is the tradeoff curve between $D_{\min}$ and s for $\rho \geq \rho_{\min}^s$. Codes that achieve points on this optimal tradeoff curve will be referred to as *Pareto optimal* codes. We see from Fig. 2 that for $s = 0$, the problem reduces to SPIR [5] for which the capacity is $C_{\text{SPIR}} = \left. \frac{L}{D_{\min}} \right|_{s=0} = 1 - \frac{1}{N}$. We further observe that i) the threshold s_t is equal to the

normalized total leakage of the capacity-achieving SJ code and TSC code in (16) and (19); ii) for $s \geq s_t$, the capacity $C = \frac{L}{D_{\min}}\big|_{s=s_t} = 1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}}$ is equal to C_{PIR}.

B. Individual Leakage

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

Theorem 2: If the amount of common randomness satisfies $\rho \geq \rho_{\min}^{w}$, where

$$\rho_{\min}^{w} \triangleq \begin{cases} \dfrac{1}{N-1} - \dfrac{N}{N-1} \cdot w, & \text{if } K = 2 \\ 0, & \text{if } K \geq 3, \end{cases} \quad (25)$$

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

$$D_{\min} = \begin{cases} L \cdot \left(\dfrac{N}{N-1} - \dfrac{1}{N-1} w \right), & \text{if } 0 \leq w \leq \dfrac{1}{N^{K-1}} \\ D_{\min}^{0}, & \text{otherwise}, \end{cases} \quad (26)$$

where $D_{\min}^{0}$ is defined in (22). If $\rho < \rho_{\min}^{w}$, then $D_{\min} = \infty$.

Proof: The proof is given in Section V. □

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

Remark 3: For $K = 2$, the individual leakage is equal to the total leakage. The minimum download cost for the same value of individual and total leakage levels are the same, which can be easily verified from (24) and (26). Similarly, the minimum amount of common randomness are also equal to each other, i.e., $\rho_{\min}^{s} = \rho_{\min}^{w}$, which can be verified from (23) and (25). This can be seen from the coding scheme in Section V-A.

Remark 4: If the amount of total leakage is $s' \triangleq \left(1 + N + \cdots + N^{K-2}\right) \cdot w$, after a simple substitution of (24), the expression of minimum download cost in (24) is equal to that in (26). Therefore, for the same download cost, the minimum total leakage is always $(1 + N + \cdots + N^{K-2})$ times the minimum individual leakage.

From (25), it is seen that $\rho_{\min}^{w} = 0$ for $K \geq 3$, which is the case when one non-desired message can be used as the encryption key to protect another individual non-desired message. For $K = 2$, there is only one non-desired message, and this makes it impossible to have another non-desired messages as the encryption key. In this case, $\rho_{\min}^{w}$ is a linear function of w as shown in (25). For $\rho \geq \rho_{\min}^{w}$, the dependency of $D_{\min}$ on (ρ, w) is illustrated in Fig. 3 and Fig. 4.

The red line in Fig. 4 is the tradeoff curve between $D_{\min}$ and w for $\rho \geq \rho_{\min}^{w}$. Similar to the observations of Fig. 2, we see from Fig. 4 that for $w = 0$, the problem becomes weakly secure PIR (WS-PIR) which differs from SPIR only in the type of security. The capacity of WS-PIR can be obtained by calculating $\frac{L}{D_{\min}}$ at $w = 0$ from (26), which is given in the following corollary.

Corollary 1: If the amount of common randomness satisfies

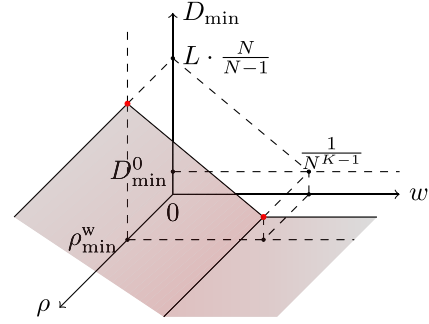$$\rho \geq \begin{cases} \dfrac{1}{N-1}, & \text{if } K = 2 \\ 0, & \text{if } K \geq 3, \end{cases} \quad (27)$$
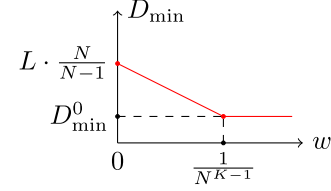

Fig. 3. The dependency of $D_{\min}$ on (ρ, w).

Fig. 4. Tradeoff curve between $D_{\min}$ and w.

the capacity of WS-PIR is given by

$$C_{\text{WS-PIR}} = 1 - \frac{1}{N}. \quad (28)$$

If (27) is not satisfied, the capacity $C_{\text{WS-PIR}} = 0$.

From this corollary, we observe that i) $C_{\text{WS-PIR}} = C_{\text{SPIR}}$, i.e., requiring strong security does not increase download cost compared to requiring weak security; ii) compared to weak security, we achieve strong security at the cost of a larger amount of common randomness. It is straightforward to verify that the optimal SPIR code in [5] is also optimal for WS-PIR. In Section V-A, we propose another optimal code for WS-PIR with $K \geq 3$, where the databases do not need to share common randomness.

Remark 5: Consider a specialization of s and w between the corner point in Fig. 2 and Fig. 4. For an integer $\alpha \geq K-1$, we write s and w in the form of

$$s = \frac{1}{N^{\alpha}} \cdot \frac{N^{K-1} - 1}{N - 1} \quad (29)$$

and

$$w = \frac{1}{N^{\alpha}}. \quad (30)$$

Denote the capacity of the PIR system for total and individual leakage constraints in (29) and (30) by C_S and C_W, respectively. Substituting (29) into Theorem 1 and (30) into Theorem 2, we obtain that for $\rho \geq \max\{\rho_{\min}^{s}, \rho_{\min}^{w}\}$,

$$C_S = C_W = \left(1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{\alpha}} \right)^{-1}, \quad (31)$$

which has a common form as $C_{\text{PIR}} = \left(1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}\right)^{-1}$ and $C_{\text{SPIR}} = C_{\text{WS-PIR}} = \left(1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{\infty}}\right)^{-1}$. This demonstrates how the capacity decreases from C_{PIR} to C_{SPIR} or $C_{\text{WS-PIR}}$ while strengthening the security requirement (i.e., decrease of s or w). For s and w

defined in (29) and (30), when we increase α, the capacity of the system decreases by adding a term of $\frac{1}{N^\alpha}$ in the denominator of the capacity expression. By increasing α from $K - 1$ to ∞, s (or w) decreases from the threshold s_t (or w_t) to zero which means perfect security, and the capacity finally decreases from $\left(1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}\right)^{-1}$ to $\left(1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^\infty}\right)^{-1} = 1 - \frac{1}{N}$.

C. Minimum Message Size

The message size $L \log_2 |\mathcal{X}|$ is also an important factor to consider in practice [4]. The minimum message size is highly dependent on the download cost. It was shown in [4] that the message size of capacity-achieving PIR code is greater than or equal to $N - 1$ if the code is uniformly decomposable. Before stating our results on the message size, we first define the notion of uniformly decomposable code, similar to that in [4].

A PIR code is called *decomposable*, if $\mathcal{Y}$ is a finite Abelian group, and for each $n \in [1 : N]$ and $q \in \mathcal{Q}_n$, the answer $A_n^{[k]} = \varphi_n(q, W_{1:K}, S)$ can be written in the form

$$\varphi_n(q, W_{1:K}, S) = \left(\varphi_{n,1}^{(q)}(W_{1:K}, S), \cdots, \varphi_{n,\ell_n}^{(q)}(W_{1:K}, S)\right) \quad (32)$$

where for $i \in \{1, 2, \cdots, \ell_n\}$,

$$\varphi_{n,i}^{(q)}(W_{1:K}, S) = \varphi_{n,i,1}^{(q)}(W_1) \oplus \cdots \oplus \varphi_{n,i,K}^{(q)}(W_K)$$
$$\oplus \varphi_{n,i,K+1}^{(q)}(S), \quad (33)$$

where $\oplus$ denotes addition in the finite group $\mathcal{Y}$, and each $\varphi_{n,i,k}^{(q)}$ ($k \in [1 : K]$) is a mapping $\mathcal{X}^L \to \mathcal{Y}$ and $\varphi_{n,i,K+1}^{(q)}$ is a mapping $\mathcal{S} \to \mathcal{Y}$.

Furthermore, a decomposable code is called *uniform* if any component function $\varphi_{n,i,k}^{(q)}$ in (33) either satisfies the conditions

$$\left|\left\{x \in \mathcal{X}^L : \varphi_{n,i,k}^{(q)}(x) = y\right\}\right| = \left|\left\{x \in \mathcal{X}^L : \varphi_{n,i,k}^{(q)}(x) = y'\right\}\right|,$$
$$y, y' \in \mathcal{Y}, \ k \in [1 : K], \quad (34)$$

and

$$\left|\left\{x \in \mathcal{S} : \varphi_{n,i,K+1}^{(q)}(x) = y\right\}\right| = \left|\left\{x \in \mathcal{S} : \varphi_{n,i,K+1}^{(q)}(x) = y'\right\}\right|,$$
$$y, y' \in \mathcal{Y}, \quad (35)$$

or it maps everything to the same value, i.e.,

$$\varphi_{n,i,k}^{(q)}(x) = \varphi_{n,i,k}^{(q)}(x'), \quad x, x' \in \begin{cases} \mathcal{X}^L, & \text{if } k \in [1 : K] \\ \mathcal{S}, & \text{if } k = K + 1. \end{cases} \quad (36)$$

The terminology "decomposable" refers to the property in (33) that each coded symbol is a summation of the component functions on each individual message. A uniformly decomposable code has an additional property that each mapping $\varphi_{n,i,k}^{(q)}$ either preserves the uniform probability distribution or induces a deterministic value. The notion of decomposable codes considerably generalizes the notion of linear codes. In particular, linear codes on finite fields are uniformly decomposable.

The following lemma and theorem are our main results on the minimum message size.

Lemma 1: The minimum message length of any Pareto optimal uniformly decomposable PIR code with either individual leakage constraint w or total leakage constraint s should be a multiple of $\log_{|\mathcal{X}|} |\mathcal{Y}|$.

Proof: The proof can be found in Appendix A. □

Remark 6: The lemma shows that the minimum L is a multiple of $\log_{|\mathcal{X}|} |\mathcal{Y}|$. However, the term $\log_{|\mathcal{X}|} |\mathcal{Y}|$ is not necessarily an integer.

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

Theorem 3: Except for the case of $K \geq 3, |\mathcal{X}| = N = 2$, the minimum message size of any Pareto optimal uniformly decomposable PIR codes achieving the minimum amount of common random randomness with either individual leakage constraint w or total leakage constraint s is $(N - 1) \log_2 |\mathcal{Y}|$; in particular, it achieves $N - 1$ when $|\mathcal{Y}| = |\mathcal{X}| = 2$.

Proof: Except for the case of $K \geq 3, |\mathcal{X}| = N = 2$, we have designed Pareto optimal codes achieving simultaneously the minimum amount of common random randomness and the minimum message size. Thus, we only need to prove the lower bound $L \log_2 |\mathcal{X}| \geq (N - 1) \log_2 |\mathcal{Y}|$, which is equivalent to $L \geq (N-1) \log_{|\mathcal{X}|} |\mathcal{Y}|$. The details can be found in Appendix B. □

Remark 7: It is remarkable that the minimum message size of the Pareto optimal uniformly decomposable PIR codes with leakage constraints w or s is equal to $N-1$, which is the same as that of optimal uniformly decomposable classical PIR codes without security constraints.

D. Relation to Shannon Cipher System

In Shannon's imperfect secrecy system in [34, pp. 71-72], X is the plain text, Y is the cipher text, and Z is the encryption key. Since X can be recovered from Y and Z, we have $H(X|Y, Z) = 0$. It was proved in [34] that this constraint implies

$$H(Z) \geq H(X) - I(X; Y). \quad (37)$$

This can be intuitively explained as follows: the key size is greater than or equal to amount of information that we need to protect, which is the difference between total amount of information $H(X)$ and the amount of information leakage $I(X; Y)$. The tradeoff between the minimum key size $H(Z)$ and the amount of information leakage $I(X; Y)$ is illustrated in Fig. 5(a).

Our result on the lower bound of common randomness in (23) of Theorem 1 presents the similar form as (37), which can be written as

$$H(S) \geq \frac{1}{N - 1}L - \frac{N^{K-1}}{N^{K-1} - 1}I(W_{\bar{k}}; A_{1:N}^{[k]}\mathbf{F}). \quad (38)$$

This can be interpreted as follows, where we focus on one particular answer. The average size of any individual answer should be at least $\frac{1}{N-1}L$, and the term $\frac{N^{K-1}}{N^{K-1}-1}I(W_{\bar{k}}; A_{1:N}^{[k]}\mathbf{F})$

(a) Tradeoff between $H(Z)$ and information leakage $I(X;Y)$.

(b) Tradeoff between $H(S)$ and total leakage $I(W_{\bar{k}}; A^{[k]}_{1:N}\mathbf{F})$.
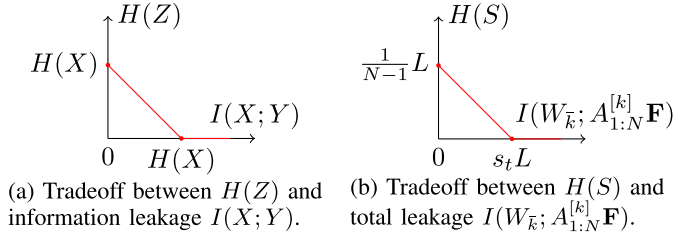
Fig. 5. Comparison of encryption key size between imperfect secrecy system and PIR system.

in (38) can roughly be viewed as the average amount of information leakage per answer. The amount of common randomness must be greater than or equal to the amount of information we need to protect, which is the difference between the total amount of information in an answer $\frac{1}{N-1}L$ and the information leakage of that answer $\frac{N^{K-1}}{N^{K-1}-1}I(W_{\bar{k}}; A^{[k]}_{1:N}\mathbf{F})$. The tradeoff between $H(S)$ and $I(W_{\bar{k}}; A^{[k]}_{1:N}\mathbf{F})$ is illustrated in Fig. 5(b). The tradeoff curve is the same as that in Fig. 5(a), and thus our result can also be interpreted as a generalization of Shannon's cipher system to multiple users with a privacy requirement.

IV. PROOF OF THEOREM 1

We prove Theorem 1 by showing the achievability, converse, and the minimum key size in the following.

A. Achievability

Following the analysis of Fig. 2, we only need to prove the theorem for $s \in [0, s_t]$. Consider a message length of $L = N - 1$. In addition to the query generation in (3), now the random key $\mathbf{F}$ generates one more indicator bit $F_0 \in \{0, 1\}$, according to probability $P(F_0 = 0) = \frac{N^{K-1}(N-1)s}{N^{K-1}-1} \in [0, 1]$ and $P(F_0 = 1) = 1 - \frac{N^{K-1}(N-1)s}{N^{K-1}-1}$. Note that the probability of F_0 is chosen according to the coding strategy so that the information leakage is within the constraint s.

For $x, y \in [1 : N]$, it is useful to define the operation $(x + y)_N$ by

$$(x + y)_N = \begin{cases} x + y, & \text{if } x + y \le N \\ x + y - N, & \text{if } x + y > N. \end{cases} \quad (39)$$

And similarly,

$$(x - y)_N = \begin{cases} x - y, & \text{if } x - y > 0 \\ x - y + N, & \text{if } x - y \le 0. \end{cases} \quad (40)$$

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

$$Q^{[k]}_n = \left(F_0, F_1, \cdots, F_{k-1}, (n - F^*)_N, F_k, \cdots, F_{K-1}\right),$$
$$n \in [1 : N], \quad k \in [1 : K]. \quad (41)$$

Since the query is a length-$(K + 1)$ vector, we can denote it by $Q^{[k]}_{n,0:K}$. Upon receiving the queries, the databases generate

the answers using the TSC code for $F_0 = 0$, and SPIR code for $F_0 = 1$. Specifically, the answer is

$$A^{[k]}_n = \begin{cases} W_{1,Q^{[k]}_{n,1}} \oplus \cdots \oplus W_{K,Q^{[k]}_{n,K}}, & \text{if } Q^{[k]}_{n,0} = 0 \\ W_{1,Q^{[k]}_{n,1}} \oplus \cdots \oplus W_{K,Q^{[k]}_{n,K}} \oplus S, & \text{if } Q^{[k]}_{n,0} = 1, \end{cases} \quad (42)$$

where S is the common randomness shared among all databases. The retrieval and privacy requirements are easily seen from the TSC and SPIR codes. Then we can check the code performances as follows.

1) Information leakage: in view of the performance of TSC code in (18) and (19), the overall normalized total leakage is

$$\frac{1}{L}I(W_{\bar{k}}; Q^{[k]}_{1:N}, A^{[k]}_{1:N})$$
$$= \frac{N^{K-1}(N-1)s}{N^{K-1}-1} \cdot \frac{1}{N-1}\left(1 - \frac{1}{N^{K-1}}\right)$$
$$+ \left[1 - \frac{N^{K-1}(N-1)s}{N^{K-1}-1}\right] \cdot 0 \quad (43)$$
$$= s, \quad (44)$$

which is the total leakage constraint.

2) Download cost: by (18) and (21), we have

$$D = \frac{N^{K-1}(N-1)s}{N^{K-1}-1} \cdot \frac{N^K - 1}{N^{K-1}}$$
$$+ \left[1 - \frac{N^{K-1}(N-1)s}{N^{K-1}-1}\right] \cdot N \quad (45)$$
$$= N - \frac{N-1}{N^{K-1}-1}s \quad (46)$$
$$= L \cdot \left(\frac{N}{N-1} - \frac{1}{N^{K-1}-1}s\right), \quad (47)$$

which is the minimum download cost in (24).

3) Amount of common randomness: since common randomness is only used in the SPIR code, we derive from (21) that

$$\rho = \frac{1 - \frac{N^{K-1}(N-1)s}{N^{K-1}-1}}{L} = \frac{1}{N-1} - \frac{N^{K-1}}{N^{K-1}-1}s, \quad (48)$$

which is equal to $\rho^s_{\min}$ in (23) of Theorem 1.

This proves the achievability of Theorem 1 for $s \in [0, s_t]$.

Remark 8: The code is obtained by combining TSC code and SPIR code with certain probabilities, which is accomplished by sending one more query bit as an indicator. This is different from time-sharing which uses sub-codes at a fraction of time. Our method i) takes advantage of the fact that the two codes have the same message size, and the message size of the combined code remains the same, which is $N - 1$ here; ii) achieves the desired leakage and download cost in expectation, while time-sharing achieves the same information leakage and download cost in a deterministic manner; iii) outperforms the time-sharing approach in the sense that time-sharing may require a large message size, which limits the applicability of the code, while the proposed scheme always uses a message size of $N - 1$.

B. Converse

Similar to the converse proof in [3], we first lower-bound the entropy of answers and messages iteratively by the message length L. Then the download cost D and information leakage s is proved to be greater than or equal to this lower bound, which gives the relation of D, s, and L. The converse in [3] lower-bounds D by L, while the converse proof here needs to lower-bound D in terms of both L and s. In a sense, this is a more general version of the proof in [3].

We first present an iterative lemma that will be used later.

Lemma 2: For all $k \in [1 : K]$,

$$NH(A^{[k]}_{1:N}, W_{1:k}|\mathbf{F}) \ge H(A^{[k+1]}_{1:N}, W_{1:k+1}|\mathbf{F}) + k(N-1)L.$$

Proof: See Appendix C. □

Applying Lemma 2 repeatedly for $k = 1, 2, \cdots, K-1$, we have

$$N^{K-1}H(A^{[1]}_{1:N}W_1|\mathbf{F})$$
$$\ge N^{K-2}H(A^{[2]}_{1:N}W_{1:2}|\mathbf{F}) + N^{K-2}(N-1)L \quad (49)$$
$$\ge N^{K-3}H(A^{[3]}_{1:N}W_{1:3}|\mathbf{F})$$
$$\quad + N^{K-2}(N-1)L + 2N^{K-3}(N-1)L \quad (50)$$
$$\vdots \quad (51)$$
$$\ge H(A^{[K]}_{1:N}W_{1:K}|\mathbf{F}) + \left(\sum_{i=1}^{K-1} iN^{K-1-i}\right)(N-1)L \quad (52)$$
$$= H(A^{[K]}_{1:N}W_{1:K}|\mathbf{F}) + \left(\sum_{i=1}^{K-1} N^i - (K-1)\right)L. \quad (53)$$

By symmetry, for any $k, k' \in [1 : K]$ and $k \ne k'$, we have

$$N^{K-1}H(A^{[k]}_{1:N}W_k|\mathbf{F}) \ge H(A^{[k']}_{1:N}W_{1:K}|\mathbf{F})$$
$$\quad + \left(\sum_{i=1}^{K-1} N^i - (K-1)\right)L. \quad (54)$$

To obtain a lower bound on download cost $D = \log_{|\mathcal{X}|}|\mathcal{Y}|\sum_{n=1}^{N}\mathbb{E}(\ell_n)$ (c.f. (13)), consider

$$D = \log_{|\mathcal{X}|}|\mathcal{Y}|\sum_{n=1}^{N}\mathbb{E}(\ell_n)$$
$$= \sum_{q_{1:N}\in\mathcal{Q}_{1:N}} \Pr(Q^{[k]}_{1:N} = q_{1:N})\sum_{n=1}^{N}\ell_n\log_{|\mathcal{X}|}|\mathcal{Y}| \quad (55)$$
$$\ge \sum_{q_{1:N}\in\mathcal{Q}_{1:N}} \Pr(Q^{[k]}_{1:N} = q_{1:N})\sum_{n=1}^{N}H(A_n^{(q_n)}) \quad (56)$$
$$\ge \sum_{q_{1:N}\in\mathcal{Q}_{1:N}} \Pr(Q^{[k]}_{1:N} = q_{1:N})H(A_1^{(q_1)}, \cdots, A_N^{(q_N)}) \quad (57)$$
$$= \sum_{q_{1:N}\in\mathcal{Q}_{1:N}} \Pr(Q^{[k]}_{1:N} = q_{1:N})H(A^{[k]}_{1:N}|Q^{[k]}_{1:N} = q_{1:N}) \quad (58)$$
$$= H(A^{[k]}_{1:N}|\mathbf{F}). \quad (59)$$

Then we derive the following,

$$\left(N^{K-1} - 1\right)D + I(W_{\bar{k}}; A^{[k]}_{1:N}Q^{[k]}_{1:N}\mathbf{F})$$
$$\ge \left(N^{K-1} - 1\right)H(A^{[k]}_{1:N}|\mathbf{F}) + I(W_{\bar{k}}; A^{[k]}_{1:N}|\mathbf{F}) \quad (60)$$
$$= \left(N^{K-1} - 1\right)H(A^{[k]}_{1:N}|\mathbf{F})$$
$$\quad + \left[H(W_{\bar{k}}) + H(A^{[k]}_{1:N}W_k|\mathbf{F}) - H(A^{[k]}_{1:N}W_{1:K}|\mathbf{F})\right] \quad (61)$$
$$= \left[(N^{K-1} - 1)H(A^{[k]}_{1:N}|\mathbf{F}) + H(A^{[k]}_{1:N}W_k|\mathbf{F})\right]$$
$$\quad + (K-1)L - H(A^{[k]}_{1:N}W_{1:K}|\mathbf{F}) \quad (62)$$
$$= N^{K-1}H(A^{[k]}_{1:N}W_k|\mathbf{F}) + (K-1)L - H(A^{[k]}_{1:N}W_{1:K}|\mathbf{F}) \quad (63)$$
$$\ge \left[H(A^{[k']}_{1:N}W_{1:K}|\mathbf{F}) + \left(\sum_{i=1}^{K-1} N^i - (K-1)\right)L\right]$$
$$\quad + (K-1)L - H(A^{[k]}_{1:N}W_{1:K}|\mathbf{F}) \quad (64)$$
$$= \left(\sum_{i=1}^{K-1} N^i\right)L + \left[H(A^{[k']}_{1:N}W_{1:K}|\mathbf{F}) - H(A^{[k]}_{1:N}W_{1:K}|\mathbf{F})\right], \quad (65)$$

where (63) follows from

$$H(A^{[k]}_{1:N}|\mathbf{F}) = H(A^{[k]}_{1:N}Q^{[k]}_{1:N}W_k|\mathbf{F}) = H(A^{[k]}_{1:N}W_k|\mathbf{F}), \quad (66)$$

and (64) follows from (54) for any $k' \in [1 : K]$ and $k' \ne k$. Similarly, by switching k and k' in (65), we obtain

$$\left(N^{K-1} - 1\right)D + I(W_{\bar{k}'}; A^{[k']}_{1:N}Q^{[k']}_{1:N}\mathbf{F})$$
$$\ge \left(\sum_{i=1}^{K-1} N^i\right)L + \left[H(A^{[k]}_{1:N}W_{1:K}|\mathbf{F}) - H(A^{[k']}_{1:N}W_{1:K}|\mathbf{F})\right]. \quad (67)$$

Summing up (65) and (67), we have

$$2\left(N^{K-1} - 1\right)D + I(W_{\bar{k}}; A^{[k]}_{1:N}\mathbf{F}) + I(W_{\bar{k}'}; A^{[k']}_{1:N}\mathbf{F})$$
$$\ge 2\left(\sum_{i=1}^{K-1} N^i\right)L. \quad (68)$$

Thus the lower bound on D is

$$D \ge \frac{N}{N-1}L$$
$$\quad - \frac{1}{N^{K-1} - 1}\frac{I(W_{\bar{k}}; A^{[k]}_{1:N}\mathbf{F}) + I(W_{\bar{k}'}; A^{[k']}_{1:N}\mathbf{F})}{2} \quad (69)$$
$$\ge L \cdot \left(\frac{N}{N-1} - \frac{1}{N^{K-1} - 1} \cdot s\right), \quad (70)$$

which matches the value of $D_{\min}$ for $s \in [0, s_t]$ in (24) of Theorem 1. For $s > s_t$, the converse follows from that of classical PIR. This completes the converse proof.

Remark 9: The proof steps for Lemma 2 are similar to those for Lemma 6 in [3], and (49)-(54) are similar to (62)-(67) in [3], which complete the converse proof therein by connecting D to L. The key difference between the proof here and that in [3] is in steps (60)-(70), which connects D to L and s jointly.

C. Amount of Common Randomness

To obtain the lower bound on the amount of common randomness, we first consider the following,

$$N^{K-1} I(W_{\bar{k}}; A_{1:N}^{[k]} Q_{1:N}^{[k]} \mathbf{F}) + (N^{K-1} - 1) H(S)$$

$$= N^{K-1} \left[H(W_{\bar{k}}) + H(A_{1:N}^{[k]} W_k | \mathbf{F}) - H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) \right]$$
$$+ (N^{K-1} - 1) H(S) \tag{71}$$

$$= N^{K-1} H(W_{\bar{k}}) + \left[N^{K-1} H(A_{1:N}^{[k]} W_k | \mathbf{F}) - H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) \right]$$
$$+ (N^{K-1} - 1) \left[H(S) - H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) \right] \tag{72}$$

$$\geq N^{K-1}(K-1)L - (N^{K-1} - 1) H(W_{1:K})$$
$$+ \left[N^{K-1} H(A_{1:N}^{[k]} W_k | \mathbf{F}) - H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) \right] \tag{73}$$

$$\geq (K - N^{K-1})L + \left(\sum_{i=1}^{K-1} N^i - (K-1) \right) L$$
$$+ \left[H(A_{1:N}^{[k']} W_{1:K} | \mathbf{F}) - H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) \right] \tag{74}$$

$$= \left(\sum_{i=0}^{K-2} N^i \right) L + \left[H(A_{1:N}^{[k']} W_{1:K} | \mathbf{F}) - H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) \right], \tag{75}$$

where (73) follows from

$$H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) - H(W_{1:K}) = H(A_{1:N}^{[k]} | W_{1:K} \mathbf{F}) \tag{76}$$
$$= H(A_{1:N}^{[k]} | W_{1:K} Q_{1:N}^{[k]} \mathbf{F}) - H(A_{1:N}^{[k]} | W_{1:K} Q_{1:N}^{[k]} \mathbf{F} S) \tag{77}$$
$$= I(S; A_{1:N}^{[k]} | W_{1:K} Q_{1:N}^{[k]} \mathbf{F}). \tag{78}$$

and

$$H(S) - I(S; A_{1:N}^{[k]} | W_{1:K} Q_{1:N}^{[k]} \mathbf{F}) \geq 0, \tag{79}$$

and (74) follows from (54) for any $k' \in [1 : K]$ and $k' \neq k$. Similarly, by switching k and k' in (75), we obtain

$$N^{K-1} I(W_{\bar{k'}}; A_{1:N}^{[k']} Q_{1:N}^{[k']} \mathbf{F}) + (N^{K-1} - 1) H(S)$$

$$\geq \left(\sum_{i=0}^{K-2} N^i \right) L + \left[H(A_{1:N}^{[k]} W_{1:K} | \mathbf{F}) - H(A_{1:N}^{[k']} W_{1:K} | \mathbf{F}) \right]. \tag{80}$$

Summing up (75) and (80), we have

$$2(N^{K-1} - 1) H(S) + N^{K-1} \left[I(W_{\bar{k}}; A_{1:N}^{[k]} \mathbf{F}) + I(W_{\bar{k'}}; A_{1:N}^{[k']} \mathbf{F}) \right]$$
$$\geq 2 \left(\sum_{i=0}^{K-2} N^i \right) L. \tag{81}$$

Then we have

$$\rho = \frac{H(S)}{L} \geq \frac{\sum_{i=0}^{K-2} N^i}{N^{K-1} - 1} L$$
$$- \frac{N^{K-1}}{N^{K-1} - 1} \frac{I(W_{\bar{k}}; A_{1:N}^{[k]} \mathbf{F}) + I(W_{\bar{k'}}; A_{1:N}^{[k']} \mathbf{F})}{2} \tag{82}$$
$$\geq \frac{1}{N-1} - \frac{N^{K-1}}{N^{K-1} - 1} \cdot s, \tag{83}$$

which is exactly the lower bound in (23) of Theorem 1.

V. PROOF OF THEOREM 2

A. Optimal Code for WS-PIR

Even though the SPIR code in [5] achieves the capacity of WS-PIR, it does not always achieve the minimum amount of common randomness for WS-PIR. From (27) in Lemma 1, we see that for $K = 2$, the minimum amount of common randomness of WS-PIR is equal to that of SPIR. This is because for $K = 2$, weak security is equivalent to strong security which means that WS-PIR is equivalent to SPIR. Thus, the SPIR code in [5] also achieves the minimum amount of common randomness of WS-PIR.

For $K \geq 3$, (27) indicates that the databases do not need to share common randomness. We consider the following three cases:

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

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

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

$$W_k = (W_{k,1}, W_{k,2} \cdots, W_{k,N-1}, W_{k,N}). \tag{84}$$

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

$$\mathbf{F} = (F_1, F_2, \cdots, F_{K-1}). \tag{85}$$

For $x, y \in [1 : N]$, the operations $(x + y)_N$ and $(x - y)_N$ are defined by (39) and (40). Let $F^* \triangleq \left(\sum_{i=1}^{K-1} F_i \right)_N$. For $k \in [1 : K]$ and $n \in [1 : N]$, the query is a deterministic function of the random key $\mathbf{F}$, defined as

$$Q_n^{[k]} = \left(F_1, F_2, \cdots, F_{k-1}, (n - F^*)_N, F_k, \cdots, F_{K-1} \right). \tag{86}$$

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

$$S = \sum_{k=1}^{K} \left(W_{k,1} \oplus W_{k,2} \oplus \cdots \oplus W_{k,N-1} \right). \tag{87}$$

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

$$A_n^{[k]} = W_{1,F_1} \oplus \cdots \oplus W_{k-1,F_{k-1}} \oplus W_{k,(n-F^*)_N}$$
$$\oplus W_{k+1,F_k} \oplus \cdots \oplus W_{K,F_{K-1}} \oplus S. \tag{88}$$

For simplicity, we define

$$B = W_{1,F_1} \oplus \cdots \oplus W_{k-1,F_{k-1}} \oplus W_{k+1,F_k} \oplus \cdots \oplus W_{K,F_{K-1}}, \tag{89}$$

where $\oplus$ denotes the addition operation in finite group $\mathcal{X}$ (this is an abuse of notation without ambiguity). Substituting (89) into (88), we have

$$A_n^{[k]} = W_{k,(n-F^*)_N} \oplus B \oplus S. \tag{90}$$

The user receives all the answers from databases, i.e., $A_1^{[k]}, A_2^{[k]}, \cdots, A_N^{[k]}$. Then we see that

$$W_{k,(n-F^*)_N} = A_n^{[k]} \ominus A_{F^*}^{[k]} = A_n^{[k]} \ominus (B \oplus S), \quad (91)$$

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

Since $\mathbf{F}$ is chosen uniformly from $\{0, 1\}^{K-1}$ and all the queries are deterministic functions of $\mathbf{F}$, we see that $Q_n^{[k]}$ is chosen uniformly from the query set for any $k \in [1 : K]$ and $n \in [1 : N]$. Thus $Q_n^{[k]}$ provides no information about the message index k, which ensures the privacy.

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

$$I(B \oplus S; W_{k'} W_k) = 0, \quad (92)$$

which implies

$$I(B \oplus S; W_k) = I(B \oplus S; W_{k'}|W_k) = 0. \quad (93)$$

Thus, we have

$$I(W_{k'}; B \oplus S, W_k) = I(W_{k'}; W_k) + I(W_{k'}; B \oplus S|W_k) = 0. \quad (94)$$

To see the security, we consider the following,

$$I(W_{k'}; A_{1:N}^{[k]} Q_{1:N}^{[k]} \mathbf{F}) = I(W_{k'}; A_{1:N}^{[k]} Q_{1:N}^{[k]} W_k \mathbf{F}) \quad (95)$$
$$= I(W_{k'}; A_{1:N}^{[k]} W_k \mathbf{F}) \quad (96)$$
$$= I(W_{k'}; B \oplus S, W_k|\mathbf{F}) \quad (97)$$
$$= 0, \quad (98)$$

where (95) follows from the recovery in (91), (97) follows from the expression of answer in (90) and $I(W_{k'}; \mathbf{F}) = 0$, and the last equality follows from (94) and $B \oplus S$ is a function of messages $W_{1:K}$ and the random key $\mathbf{F}$.

Since each answer consists of one symbol, the download cost is

$$D = N. \quad (99)$$

Since $L = N - 1$, the rate of the code is simply $R = \frac{N-1}{N} = 1 - \frac{1}{N}$ that matches the capacity in (28) of Lemma 1. Lastly, the databases do not share common randomness in addition to the messages themselves.

Case ii): It is easy to verify that the code also works for $|\mathcal{X}| = 2$ and $N \geq 3$. This can be seen intuitively as follows. Since $L = N - 1 \geq 2$, each message consists of at least two symbols. The encryption key S consists of all the message symbols. Since $W_{k,(n-F^*)_N} \oplus B$ contains at most one symbol from each message, it can cancel out at most one symbol from S. Then the answer $A_n^{[k]} = W_{k,(n-F^*)_N} \oplus B \oplus S$ (c.f. (90)) contains at least one symbol from each message. Since $K \geq 3$,

(92) and the subsequent equations also hold for the current case. This ensures the security of each individual message.

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

$$q_1^{(1)} = (2, 1, 1), \quad q_2^{(1)} = (2, 2, 1) \quad (100)$$

and

$$q_1^{(2)} = (1, 1, 2), \quad q_2^{(2)} = (1, 2, 2). \quad (101)$$

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

$$B^{(1)} \oplus S = W_{1,2}^{(1)} \oplus W_{3,1}^{(1)} \oplus S = W_1^{(1)} \oplus W_1^{(2)} \oplus W_3^{(2)}, \quad (102)$$
$$B^{(2)} \oplus S = W_{1,1}^{(2)} \oplus W_{3,2}^{(2)} \oplus S = W_1^{(1)} \oplus W_3^{(1)} \oplus W_3^{(2)}. \quad (103)$$

We can easily see that $(B^{(1)} \oplus S, B^{(2)} \oplus S)$ provides no information about either W_1 or W_3. Then using similar arguments as (92)-(98), we can obtain the weak security.

Remark 10: The SPIR code in [5] is an optimal code for WS-PIR that achieves capacity. However, it is important to point out that the method of using the sum of message symbols as encryption keys does not work for SPIR, since it leaks partial of the total information of the whole message set $W_{1:K}$.

B. Achievability of Pareto Optimal Points

Following the analysis of Fig. 4 after Lemma 1, we only need to prove the theorem for $w \in [0, w_t]$. Since for $K = 2$, the WS-PIR code reduces to SPIR code and the achievability is proved in Section IV-A, we consider only $K \geq 3$ here.

Similar to Section IV-A, consider a message length of $L = N - 1$. The random key $\mathbf{F}$ generates one more indicator bit $F_0 \in \{0, 1\}$ according to probability $P(F_0 = 0) = wN^{K-1} \in [0, 1]$ and $P(F_0 = 1) = 1 - wN^{K-1}$, which is chosen to control the individual leakage within the constraint w. The queries and answers are generated similarly as that in Section IV-A. The only difference is to replace S by the encryption key using (87). We then check the following.

1) Information leakage: in view of the performance of TSC code in (18) and (20), the overall normalized total leakage is

$$\frac{1}{L} I(W_{\bar{k}}; Q_{1:N}^{[k]}, A_{1:N}^{[k]}, \mathbf{F})$$
$$= wN^{K-1} \cdot \frac{1}{N^{K-1}} + (1 - wN^{K-1}) \cdot 0 \quad (104)$$
$$= w, \quad (105)$$

which is the individual leakage constraint.

2) Download cost: by (18) and (99), we have

$$D = wN^{K-1} \cdot \frac{N^K - 1}{N^{K-1}} + (1 - wN^{K-1}) \cdot N \quad (106)$$

$$= N - w \quad (107)$$

$$= L \cdot \left(\frac{N}{N-1} - \frac{1}{N-1} w \right), \quad (108)$$

which is the minimum download cost in (26) of Theorem 2.

3) Amount of common randomness: i) for $K = 2$, the key size in (48) reduces to $\frac{1}{N-1} - \frac{N}{N-1} \cdot w$; ii) for $K \geq 3$, there is no common randomness needed and $\rho = 0$. Thus the amount of common randomness of the code is equal to $\rho_{\min}^s$ in (23) of Theorem 1.

This proves the achievability of Theorem 2 for $w \in [0, w_t]$.

Remark 11: We notice that the linear combination coefficients as well as the other code performances here are the same as that of the linear combination between TSC and SPIR codes in Section IV-A. The reason is $C_{\text{WS-PIR}} = C_{\text{PIR}}$ and the WS-PIR code in Section V-A has the same message length and download cost as SPIR code in [5].

Remark 12: The achievability proof of Theorem 1 in Section IV-A is a probabilistic combination of two existing codes, i.e., the TSC code in [4] and SPIR code in [5]. However, the achievability proof of Theorem 2 in Section V-B uses a new WS-PIR code proposed in Section V-A that does not require common randomness.

C. Converse

The converse proof is very similar to that in Section IV-B except that the iteration is replaced by a single relation between the entropy of answers for desired and non-desired messages.

For any $k, k' \in [1 : K]$ and $k \neq k'$, we first derive the following which is similar to the proof of Lemma 2.

$$NH(A_{1:N}^{[k]} W_k | \mathbf{F})$$

$$\geq \sum_{n=1}^{N} H(A_n^{[k]} W_k | \mathbf{F}) \quad (109)$$

$$= \sum_{n=1}^{N} H(A_n^{[k]} W_k | Q_n^{[k]}) \quad (110)$$

$$= \sum_{n=1}^{N} H(A_n^{[k']} W_k | Q_n^{[k']}) \quad (111)$$

$$= \sum_{n=1}^{N} H(A_n^{[k']} W_k | \mathbf{F}) \quad (112)$$

$$= \sum_{n=1}^{N} H(A_n^{[k']} | W_k \mathbf{F}) + N H(W_k) \quad (113)$$

$$\geq \sum_{n=1}^{N} H(A_n^{[k']} | W_k \mathbf{F} A_{1:n-1}^{[k']}) + NL \quad (114)$$

$$= H(A_{1:N}^{[k']} | W_k \mathbf{F}) + NL \quad (115)$$

$$= H(A_{1:N}^{[k']} W_k | \mathbf{F}) + (N - 1)L \quad (116)$$

$$= H(A_{1:N}^{[k']} Q_{1:N}^{[k']} W_k W_{k'} | \mathbf{F}) + (N - 1)L \quad (117)$$

$$= H(A_{1:N}^{[k']} W_k W_{k'} | \mathbf{F}) + (N - 1)L, \quad (118)$$

where (110) and (112) follow from the Markov chain $(A_n^{[k]}, W_{[1:K]}) \rightarrow Q_n^{[k]} \rightarrow \mathbf{F}$, and (111) follows from identical distribution constraint in (10). Thus, we have

$$NH(A_{1:N}^{[k]} W_k | \mathbf{F}) - H(A_{1:N}^{[k']} W_k W_{k'} | \mathbf{F}) \geq (N - 1)L. \quad (119)$$

To provide a lower bound on the download cost D (c.f. (13)), we derive the following

$$(N-1)D + I(W_{k'}; A_{1:N}^{[k]} Q_{1:N}^{[k]} \mathbf{F})$$

$$\geq (N-1)H(A_{1:N}^{[k]} | \mathbf{F}) + I(W_{k'}; A_{1:N}^{[k]} | \mathbf{F}) \quad (120)$$

$$= \left[H(W_{k'}) + H(A_{1:N}^{[k]} W_k | \mathbf{F}) - H(A_{1:N}^{[k]} W_k W_{k'} | \mathbf{F}) \right]$$

$$\quad + (N-1)H(A_{1:N}^{[k]} W_k | \mathbf{F}) \quad (121)$$

$$= NH(A_{1:N}^{[k]} W_k | \mathbf{F}) - H(A_{1:N}^{[k]} W_k W_{k'} | \mathbf{F}) + L \quad (122)$$

$$\geq NL + \left[H(A_{1:N}^{[k']} W_k W_{k'} | \mathbf{F}) - H(A_{1:N}^{[k]} W_k W_{k'} | \mathbf{F}) \right] \quad (123)$$

where (120) follows from (59), (121) follows from (3) and (7), and (123) follows from (119). By switching k and k' in (123), we obtain

$$(N-1)D + I(W_k; A_{1:N}^{[k']} Q_{1:N}^{[k']} \mathbf{F})$$

$$\geq NL + \left[H(A_{1:N}^{[k]} W_k W_{k'} | \mathbf{F}) - H(A_{1:N}^{[k']} W_k W_{k'} | \mathbf{F}) \right]. \quad (124)$$

Summing up (123) and (124), we have

$$2(N-1)D + \left[I(W_{k'}; A_{1:N}^{[k]} \mathbf{F}) + I(W_k; A_{1:N}^{[k']} \mathbf{F}) \right] \geq 2NL.$$

Thus, the lower bound on D is

$$D \geq \frac{N}{N-1}L - \frac{1}{N-1} \frac{I(W_{k'}; A_{1:N}^{[k]} \mathbf{F}) + I(W_k; A_{1:N}^{[k']} \mathbf{F})}{2}$$

$$\geq L \cdot \left(\frac{N}{N-1} - \frac{1}{N-1} w \right) \quad (125)$$

which matches the value of $D_{\min}$ for $w \in [0, w_t]$ in (26) of Theorem 2. For $w > w_t$, the converse follows from that of classical PIR. This completes the converse proof.

D. Minimum Key Size

For $K = 2$, individual leakage is equal to total leakage, which means that PIR system with the same value of total leakage constraint s and individual leakage constraint w are equivalent. This gives the minimum amount of common randomness in (25) of Theorem 2. For $K \geq 3$, the lower bound in (25) is zero, which is trivial.

VI. CONCLUSION

In this work, we studied the PIR problem with total and individual leakage constraints, respectively. Our main contribution was the characterization of the tradeoffs between the minimum download cost $D_{\min}$ and the total leakage constraint s and individual leakage constraint w. The minimum amount of common randomness with respect to s and w was also characterized. Remarkably, for $K \geq 3$, the databases do not need to share common randomness in order to achieve the

individual leakage constraint w. For $s \in [0, s_t]$, it was shown that the linear combination of TSC code and SPIR code is Pareto optimal (achieving the whole tradeoff curve). For $w \in [0, w_t]$, an optimal code that achieves the capacity of WS-PIR without shared common randomness was proposed ($K \geq 3$), and its linear combination with the classical PIR code was proved Pareto optimal. The proposed Pareto optimal codes for both individual and total leakage were proved to have a minimum message size of $N - 1$.

APPENDIX A
PROOF OF LEMMA 1

Throughout the proof, we use L to denote the minimum message length for Pareto optimal uniformly decomposable codes. Since we have coding scheme of message length $(N - 1) \log_{|\mathcal{X}|} |\mathcal{Y}|$, the minimum message length should be $L \leq (N - 1) \log_{|\mathcal{X}|} |\mathcal{Y}|$. If $L = (N - 1) \log_{|\mathcal{X}|} |\mathcal{Y}|$, we are done. Otherwise, $L < (N - 1) \log_{|\mathcal{X}|} |\mathcal{Y}|$.

In the following, we prove the lemma by contradiction and suppose that $\frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}$ is not an integer. We only consider the individual leakage case, since by Remark 4 the same results also hold for the total leakage case. From (26) in Theorem 2, the average number of downloaded symbols in $\mathcal{Y}$ for Pareto optimal codes should be

$$D_{\mathcal{Y}} = \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} \left(\frac{N}{N-1} - \frac{w}{N-1} \right) = \frac{N-w}{N-1} \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}.$$
(126)

Since $L < (N - 1) \log_{|\mathcal{X}|} |\mathcal{Y}|$ implies $0 < \frac{N-w}{N-1} \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} - \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} < 1$, the only possible realization of queries that have a downloaded number of symbols $d \leq D_{\mathcal{Y}}$ are those q inducing $d = \lceil \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} \rceil$ downloaded symbols. It is easy to see that

$$L < \lceil \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} \rceil \cdot \log_{|\mathcal{X}|} |\mathcal{Y}| < (L + 1).$$
(127)

Since the equality in (120) is equivalent to equality in (59), we see from Appendix B that independence property **P**1 holds for both total and individual leakage cases. By **P**1 and uniform distribution of answer symbols, we obtain

$$H(A_{1:N}^{(q)}) = \lceil \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} \rceil \cdot \log_{|\mathcal{X}|} |\mathcal{Y}|.$$
(128)

Since the whole answer set can recover W_k, we have $H(W_k | A_{1:N}^{(q)}) = 0$, and thus $H(A_{1:N}^{(q)} | W_k) = H(A_{1:N}^{(q)}) - H(W_k)$. Then from (127) and (128), we can bound $H(A_{1:N}^{(q)} | W_k)$ by

$$0 < H(A_{1:N}^{(q)} | W_k) < \log_{|\mathcal{X}|} |\mathcal{Y}|.$$
(129)

Next, we derive another bound for $H(A_{1:N}^{(q)} | W_k)$, which leads to a contradiction. By the definition of uniform code in (34)-(36), denote the non-deterministic symbols of $A_{1:N}^{(q)}$ by $A_{n,i}^{(q)} = \varphi_{n,i}^{(q)}$. For $n \in [1 : N]$ and $i \in [1 : \ell_n]$, define a random variable by $U_{n,i}^{(q)} = A_{n,i}^{(q)} \ominus \varphi_{n,i,k}^{(q)}(W_k)$, where $\ominus$ is

the subtraction in the Abelian group $\mathcal{X}$ as defined after (91). Then, we have

$$H(A_{1:N}^{(q)} | W_k) = H(U_{1,1}^{(q)}, \cdots, U_{N,\ell_N}^{(q)}).$$
(130)

We see that $U_{n,i}^{(q)}$ is either deterministic or uniformly distributed over $\mathcal{Y}$, thus $H(U_{n,i}^{(q)}) = 0$ or $\log_{|\mathcal{X}|} |\mathcal{Y}|$. If $H(U_{n,i}^{(q)}) = 0$ for all (n, i), we have $H(U_{1,1}^{(q)}, \cdots, U_{N,\ell_N}^{(q)}) = 0$. Otherwise, there is at least one pair (n, i) such that $H(U_{n,i}^{(q)}) = \log_{|\mathcal{X}|} |\mathcal{Y}|$, which implies that $H(U_{1,1}^{(q)}, \cdots, U_{N,\ell_N}^{(q)}) \geq \log_{|\mathcal{X}|} |\mathcal{Y}|$. Thus, we see from (130) that

$$H(A_{1:N}^{(q)} | W_k) \begin{cases} = 0 \\ \geq \log_{|\mathcal{X}|} |\mathcal{Y}|. \end{cases}$$
(131)

Now, (131) contradicts with (129). Then the supposition that $\frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}$ is non-integer is false, and thus L is a multiple of $\log_{|\mathcal{X}|} |\mathcal{Y}|$.

APPENDIX B
PROOF OF THEOREM 3

A. Total Leakage Case

In [4], three properties P1-P3 were used to prove the minimum message size. Here, we verify that P1-P3 are also true for Pareto optimal uniformly decomposable PIR codes with total leakage constraint s, and then the minimum message size can be obtained accordingly. We first define some notations before presenting the properties. For notational simplicity, we write $W_{K+1} = S$ and then $W_{1:K+1} = (W_{1:K}, S)$ and $W_{\bar{k}} = (W_1, \cdots, W_{k-1}, W_{k+1}, \cdots, W_K, S)$. For $n \in [1 : N]$ and $q \in \mathcal{Q}_n$, the answer for decomposable codes can be simply written in matrix form as

$$A_n^{[k]} = \varphi_n(Q_n^{[k]} = q, W_{1:K}, S) = W_{1:K+1} \cdot G_n^{(q)},$$
(132)

where $W_{1:K+1}$ is viewed as a length-$(K + 1)$ vector, and $G_n^{(q)}$ is a matrix of dimension $(K + 1) \times l_n$ whose elements $G_{n,i,k}^{(q)}$ are functions $\mathcal{X}^L (\text{or } \mathcal{S}) \to \mathcal{Y}$ with the "·" operation defined as $W_k \cdot G_{n,i,k}^{(q)} \triangleq \varphi_{n,i,k}^{(q)}(W_k)$. Let $q_{1:N}$ be a set of queries such that $\Pr(Q_{1:N}^{[k]} = q_{1:N}) > 0$. For any $\mathcal{A} \subseteq [1 : K]$, let $G_{n|\mathcal{A}}^{(q)}$ be the submatrix of $G_n^{(q)}$ with rows corresponding to $\{W_i, i \in \mathcal{A}\}$ removed. For Pareto optimal uniformly decomposable PIR codes, the three properties are stated as follows.

P1. The N random variables $A_1^{(q_1)}, A_2^{(q_2)}, \cdots, A_N^{(q_N)}$ are mutually independent, where $A_n^{(q_n)}$ is the n-th answer when $Q_n^{[k]} = q_n$.

P2. The N random variables $W_{\bar{k}} \cdot G_{1|k}^{(q_1)}, W_{\bar{k}} \cdot G_{2|k}^{(q_2)}, \cdots, W_{\bar{k}} \cdot G_{N|k}^{(q_N)}$ are deterministic of each other.

P3. The N random variables $W_k \cdot G_{1|\bar{k}}^{(q_1)}, W_k \cdot G_{2|\bar{k}}^{(q_2)}, \cdots, W_k \cdot G_{N|\bar{k}}^{(q_N)}$ are mutually independent.

Then the properties can be verified as follows.

  i) It is easily seen from the proof of **P**1 in [4] that the equality of (59) in the current paper holds if and only if the equality in Lemma 1 of [4] holds.

ii) When (161) is equality, we obtain for $k = 1$ that

$$H(A_{1:N}^{[1]}|W_1 A_n^{[1]}\mathbf{F}) = 0, \tag{133}$$

which is exactly the condition that proves **P2** in [4].

iii) The inequality in (166) is exactly the same as (d) in the proof of Lemma 2 in [4], which induces **P3**.

The above observations ensure that the properties **P1**-**P3** are also true for Pareto optimal uniformly decomposable PIR codes with total leakage constraint s. Then we can use the argument in [4] to lower bound the message length as $L \geq (N-1)\log_{|\mathcal{X}|}|\mathcal{Y}|$.

B. Individual Leakage Case

For the individual leakage case, the following lemma provides a new method of proving the minimum message size. We can also apply the same method to the total leakage case, where the analysis of total leakage may be more complicated. The details for verifying this is omitted here.

Lemma 3: The message length of Pareto optimal uniformly decomposable PIR codes with individual leakage constraint w is lower bounded by $L \geq (N-1)\log_{|\mathcal{X}|}|\mathcal{Y}|$.

Proof: For $w > \frac{1}{N^{K-1}}$, the problem is equivalent to classical PIR. Thus, we only consider $w \in [0, \frac{1}{N^{K-1}}]$, for which we prove the lemma by contradiction. By Lemma 1, we suppose there exist Pareto optimal uniformly decomposable codes such that

$$L \leq (N-2)\log_{|\mathcal{X}|}|\mathcal{Y}|. \tag{134}$$

Fix a $k \in [1 : K]$. For each query realization $Q_{1:N}^{[k]} = q_{1:N}$, let d be the corresponding total number of downloaded symbols in $\mathcal{Y}$, which can be written as $d = \sum_{n=1}^{N} \ell_n(q_n)$. Because of recovery requirement of W_k, we have

$$d \geq \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}. \tag{135}$$

For nonnegative integer m, let $\mathcal{Q}^{(m)}$ be the set of query realizations defined by

$$\mathcal{Q}^{(m)} \triangleq \left\{ q_{1:N} : d = \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} + m \right\}. \tag{136}$$

Define a probability $P^{(m)}$ by

$$P^{(m)} = \sum_{q_{1:N} \in \mathcal{Q}^{(m)}} \Pr(q_{1:N}). \tag{137}$$

From (126), the average number of downloaded symbols in $\mathcal{Y}$ for Pareto optimal codes is

$$D_{\mathcal{Y}} = \frac{N-w}{N-1} \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}, \tag{138}$$

which, by the supposition (134) and the range of $w \in [0, \frac{1}{N^{K-1}}]$, satisfies $0 < D_{\mathcal{Y}} - \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} < 1$, and thus $D_{\mathcal{Y}}$ is not an integer. In light of the lower bound of d in (135), the only possible realization of queries that induce $d < D_{\mathcal{Y}}$

are $q_{1:N} \in \mathcal{Q}^{(0)}$, and the corresponding downloaded number of symbols is $d = \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}$. Then we have

$$D_{\mathcal{Y}} = \sum_m P^{(m)} \left(\frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} + m \right) \tag{139}$$

$$= \left(\sum_m P^{(m)} \right) \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} + \sum_{m \geq 1} m P^{(m)} \tag{140}$$

$$= \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} + \sum_{m \geq 1} m P^{(m)} \tag{141}$$

$$\geq \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} + \sum_{m \geq 1} P^{(m)} \tag{142}$$

$$= \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} + (1 - P^{(0)}), \tag{143}$$

which by (138) implies that

$$P^{(0)} \geq 1 - \frac{1-w}{N-1} \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}. \tag{144}$$

In the following, we will derive a lower bound for information leakage that leads to a contradiction with the individual leakage constraints. For $q_{1:N} \in \mathcal{Q}^{(0)}$, we have

$$d = \ell_1 + \ell_2 + \cdots + \ell_N = \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}. \tag{145}$$

which implies $\ell_n \in \left[0 : \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}\right]$ for any $n \in [1 : N]$. Then for any $n \in [1 : N]$, define a parameter by

$$T_n^{[k]} \triangleq \sum_{i=1}^{\frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}} i \cdot \Pr\left\{ q_{1:N} \in \mathcal{Q}^{(0)}, \ell_n = i \right\}, \tag{146}$$

which can be roughly viewed as the average length of the answer when $q_{1:N} \in \mathcal{Q}^{(0)}$. Then it is intuitive that $\sum_{n=1}^{N} T_n^{[k]} = \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} \cdot P^{(0)}$, which implies that there exists an $n \in [1 : N]$ such that

$$T_n^{[k]} \geq \frac{\frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}}{N} \cdot P^{(0)}. \tag{147}$$

In the following, we consider this particular n. Then by (144), we have

$$T_n^{[k]} \geq \frac{\frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|}}{N} \left(1 - \frac{1-w}{N-1} \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} \right). \tag{148}$$

For $q_{1:N} \in \mathcal{Q}^{(0)}$, since each answer symbol is either deterministic or uniformly distributed over $\mathcal{Y}$, we obtain from (145) that

$$H(A_{1:N}^{(q_{1:N})}) \leq \frac{L}{\log_{|\mathcal{X}|}|\mathcal{Y}|} \cdot \log_{|\mathcal{X}|}|\mathcal{Y}| = L = H(W_k). \tag{149}$$

This together with the recovery requirement of W_k, i.e., $H(W_k|A_{1:N}^{(q_{1:N})}) = 0$, imply that $H(A_{1:N}^{(q_{1:N})}) = H(W_k)$. Then for the queries $q_{1:N} \in \mathcal{Q}^{(0)}$, the answers $A_{1:N}^{(q_{1:N})}$ and the message W_k are deterministic of each other, and all the answer symbols are non-deterministic.

Let $A_n^{(q_n)} = \left(A_{n,1}^{(q_n)}, A_{n,2}^{(q_n)}, \cdots, A_{n,\ell_n}^{(q_n)} \right)$ where each component $A_{n,j}^{(q_n)} \in \mathcal{Y}$. Then we have

$$I(W_k; A_{n,j}^{(q_n)}) = H(A_{n,j}^{(q_n)}) = \log_{|\mathcal{X}|} |\mathcal{Y}|, \quad \forall j \in [1 : \ell_n]. \quad (150)$$

Define a set of query realizations of $Q_n^{[k]}$ by $\mathcal{Q}_n^{(0)} \triangleq \left\{ q_n : q_{1:N} \in \mathcal{Q}^{(0)} \right\}$. Since for $q_{1:N} \notin \mathcal{Q}^{(0)}$, the n-th query can also be in $\mathcal{Q}_n^{(0)}$, we have

$$\Pr\{q_n \in \mathcal{Q}_n^{(0)}\} \geq \Pr\{q_{1:N} \in \mathcal{Q}^{(0)}\}. \quad (151)$$

Next, we use (148) to derive an lower bound for individual leakage that will lead to a contradiction. Consider $k' \neq k$. Similarly, we define $T_n^{[k']}$ by

$$T_n^{[k']} \triangleq \sum_{i=1}^{\frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}} i \cdot \Pr\left\{ q_n \in \mathcal{Q}_n^{(0)}, \ell_n = i \right\}. \quad (152)$$

Then the leakage of W_k from the answers for k' can be bounded as follows,

$$I(W_k; A_{1:N}^{[k']} Q_{1:N}^{[k']}) \geq I(W_k; A_n^{[k']} Q_n^{[k']}) \geq T_n^{[k']} \cdot \log_{|\mathcal{X}|} |\mathcal{Y}|, \quad (153)$$

where the last inequality follows from (150). Because of identical distribution of $Q_n^{[k]}$ and $Q_n^{[k']}$ in (10), we obtain from (151) that

$$\Pr\left\{ Q_n^{[k']} \in \mathcal{Q}_n^{(0)} \right\} = \Pr\left\{ Q_n^{[k]} \in \mathcal{Q}_n^{(0)} \right\} \geq \Pr\left\{ Q_{1:N}^{[k]} \in \mathcal{Q}^{(0)} \right\},$$

which implies that

$$T_n^{[k']} \geq T_n^{[k]}. \quad (154)$$

By (153), (154), and the lower bound in (148), we obtain

$$I(W_k; A_{1:N}^{[k']} Q_{1:N}^{[k']}) \geq \frac{L}{N} \left(1 - \frac{1-w}{N-1} \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} \right). \quad (155)$$

In the following, we derive the contradiction by arguing

$$\frac{L}{N} \left(1 - \frac{1-w}{N-1} \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|} \right) > wL, \quad (156)$$

which can be verified equivalent to

$$w < \frac{(N-1) - \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}}{(N^2 - N) - \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}}. \quad (157)$$

It is easy to see that the right-hand side is a decreasing function of L. Since by assumption $L \leq (N-2)\frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}$, we have

$$\frac{(N-1) - \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}}{(N^2 - N) - \frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}} \geq \frac{(N-1) - (N-2)}{(N^2 - N) - (N-2)}$$

$$= \frac{1}{N^2 - 2(N-1)}. \quad (158)$$

For $K \geq 3$ and $N \geq 2$, we can see that $\frac{1}{N^2 - 2(N-1)} > \frac{1}{N^{K-1}}$. Thus, for any $w \in [0, \frac{1}{N^{K-1}}]$, we have

$$\frac{1}{N^2 - 2(N-1)} > w. \quad (159)$$

From (158) and (159), we obtain (157), and thus the argument (156) is true. Then (155) and (156) imply that

$$I(W_k; A_{1:N}^{[k']} Q_{1:N}^{[k']}) > wL, \quad (160)$$

which contradicts the individual leakage constraint definition in (12). Thus, the supposition at beginning that $L \leq (N-2)\frac{L}{\log_{|\mathcal{X}|} |\mathcal{Y}|}$ is false, and this proves the lemma. $\square$

APPENDIX C
PROOF OF LEMMA 2

To prove the lemma, we consider the following.

$$NH(A_{1:N}^{[k]}, W_{1:k}|\mathbf{F})$$

$$\geq \sum_{n=1}^{N} H(A_n^{[k]}, W_{1:k}|\mathbf{F}) \quad (161)$$

$$= \sum_{n=1}^{N} H(A_n^{[k]}, W_{1:k}|Q_n^{[k]}) \quad (162)$$

$$= \sum_{n=1}^{N} H(A_n^{[k+1]}, W_{1:k}|Q_n^{[k+1]}) \quad (163)$$

$$= \sum_{n=1}^{N} H(A_n^{[k+1]}, W_{1:k}|\mathbf{F}) \quad (164)$$

$$= \sum_{n=1}^{N} H(A_n^{[k+1]}|W_{1:k}, \mathbf{F}) + NH(W_{1:k}) \quad (165)$$

$$\geq \sum_{n=1}^{N} H(A_n^{[k+1]}|W_{1:k}, \mathbf{F} A_{1:n-1}^{[k+1]}) + NH(W_{1:k}) \quad (166)$$

$$= H(A_{1:N}^{[k+1]}|W_{1:k}, \mathbf{F}) + NH(W_{1:k}) \quad (167)$$

$$= H(A_{1:N}^{[k+1]}, W_{1:k}|\mathbf{F}) + (N-1)H(W_{1:k}) \quad (168)$$

$$= H(A_{1:N}^{[k+1]}, Q_{1:N}^{[k+1]}, W_{1:k+1}|\mathbf{F}) + (N-1)H(W_{1:k}) \quad (169)$$

$$= H(A_{1:N}^{[k+1]}, W_{1:k+1}|\mathbf{F}) + k(N-1)L, \quad (170)$$

where (162) and (164) follow from the Markov chain $(A_n^{[k]}, W_{[1:K]}) \to Q_n^{[k]} \to \mathbf{F}$, and (163) follows from identical distribution constraint in (10). The lemma is proved.

REFERENCES

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

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

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

[4] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7613–7627, Nov. 2019.

[5] 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.

[6] 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.

[7] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 2920–2932, Dec. 2017.

[8] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.

[9] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by freij-hollanti et al.," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 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] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[12] K. Banawan and S. Ulukus, "The capacity of private information retrieval from byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[13] Q. Wang and M. Skoglund, "Symmetric private information retrieval for MDS coded distributed storage," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

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

[16] Q. Wang and M. Skoglund, "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.

[17] M. K. Kiskani and H. R. Sadjadpour, "Secure and private information retrieval (SAPIR) in cloud storage systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12302–12312, Dec. 2018.

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

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

[20] 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.

[21] Y. Zhang, X. Wang, H. Wei, and G. Ge, "On private information retrieval array codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5565–5573, Sep. 2019.

[22] K. Banawan and S. Ulukus, "Asymmetry hurts: Private information retrieval under asymmetric traffic constraints," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7628–7645, Nov. 2019.

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

[24] 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 Geometry*, vol. 1, no. 1, pp. 647–664, Jan. 2017.

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

[26] S. Kumar, E. Rosnes, and A. G. i Amat, "Private information retrieval in distributed storage systems using an arbitrary linear code," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1421–1425.

[27] S. Kumar, H.-Y. Lin, E. Rosnes, and A. G. 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.

[28] 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*, to be published.

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

[30] S. Even, O. Goldreich, and A. Lempel, "A randomized protocol for signing contracts," *Commun. ACM*, vol. 28, no. 6, pp. 637–647, 1985.

[31] Y. Ishai, M. Prabhakaran, and A. Sahai, "Founding cryptography on oblivious transfer–efficiently," in *Proc. Annu. Int. Cryptol. Conf.*, Berlin, Germany: Springer, 2008, pp. 572–591.

[32] M. Naor and B. Pinkas, "Distributed oblivious transfer," in *Proc. Int. Conf. Theory Appl. Cryptol. Inf. Secur.*, Berlin, Germany: Springer, 2000, pp. 205–219.

[33] H. Yamamoto, "On secret sharing communication systems with two or three channels," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 3, pp. 387–393, May 1986.

[34] R. W. Yeung, *Information Theory and Network Coding*. New York, NY, USA: Springer, 2008.

[35] T. Guo, X. Guang, and K. Shum, "Symmetric multilevel imperfect secret sharing," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Dec. 2018, pp. 1–5.

[36] K. Bhattad and K. R. Narayanan, "Weakly secure network coding," in *Proc. NetCod*, Riva del Garda, Italy, Apr. 2005, p. 104.

[37] M. Yan and A. Sprintson, "Weakly secure network coding for wireless cooperative data exchange," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Kathmandu, Nepal, Dec. 2011, pp. 1–5.

[38] T. Guo, C. Tian, T. Liu, and R. W. Yeung, "Weakly secure symmetric multilevel diversity coding," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Sweden, Aug. 2019, pp. 1–5.

[39] Y. Chen, O. O. Koyluoglu, and A. J. H. Vinck, "On secure communication over the multiple access channel," in *Proc. Int. Symp. Inf. Theory Its Appl. (ISITA)*, Monterey, CA, USA, Oct. 2016, pp. 350–354.

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

[41] 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. 2019, pp. 1262–1266.

[42] Z. Jia, "On the capacity of weakly-private information retrieval," M.S. thesis, Dept. Elect. Eng., Univ. California, Irvine, CA, USA, 2019.

Tao Guo (Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University in 2013 and the Ph.D. degree from the Department of Information Engineering, The Chinese University of Hong Kong, in 2018. He is currently a Post-Doctoral Research Associate with the Department of Electrical and Computer Engineering, Texas A&M University. His research interests include information theory and it applications to security and privacy, multi-user source coding, and coding for distributed storage systems.

Ruida Zhou (Student Member, IEEE) received the B.S. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Texas A&M University at College Station, College Station, TX, USA. His research interests include information theory and statistical learning.

Chao Tian (Senior Member, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2000, and the M.S. and Ph.D. degrees in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2003 and 2005, respectively.

He was a Post-Doctoral Researcher with the École Polytechnique Fédérale de Lausanne (EPFL) from 2005 to 2007, a Member of Technical Staff–Research at AT&T Labs–Research, NJ, USA, from 2007 to 2014, and an Associate Professor with the Department of Electrical Engineering and Computer Science, The University of Tennessee, Knoxville, from 2014 to 2017. He joined the Department of Electrical and Computer Engineering, Texas A&M University, in 2017. His research interests include data storage systems, multi-user information theory, joint source-channel coding, signal processing, and compute algorithms.

Dr. Tian received the Liu Memorial Award at Cornell University in 2004 and the AT&T Key Contributor Award in 2010, 2011, and 2013. His authored or coauthored articles received the 2014 IEEE ComSoc DSTC Data Storage Best Paper Award and the 2017 IEEE Jack Keil Wolf ISIT Student Paper Award. He was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS from 2012 to 2014. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY.