

Single-Server Private Information Retrieval With Side Information Under Arbitrary Popularity Profiles

Alejandro Gomez-Leos and Anoosheh Heidarzadeh

Abstract—This paper introduces a generalization of the Private Information Retrieval with Side Information (PIR-SI) problem called Popularity-Aware PIR-SI (PA-PIR-SI). The PA-PIR-SI problem includes one or more remote servers storing copies of a dataset of K messages, and a user who knows M out of K messages—the identities of which are unknown to the server—as a prior side information, and wishes to retrieve one of the remaining $K - M$ messages. The goal of the user is to minimize the amount of information they must download from the server while revealing no information about the identity of the desired message. In contrast to PIR-SI, in PA-PIR-SI, the dataset messages are not assumed to be equally popular. That is, given the M side information messages, each of the remaining $K - M$ messages is not necessarily equally likely to be the message desired by the user. In this work, we focus on the single-server setting of PA-PIR-SI, and establish lower and upper bounds on the capacity of this setting—defined as the maximum possible achievable download rate. Our upper bound holds for any message popularity profile, and is the same as the capacity of single-server PIR-SI. We prove the lower bound by presenting a PA-PIR-SI scheme which takes a novel probabilistic approach—carefully designed based on the popularity profile—to integrate two existing PIR-SI schemes. The rate of our scheme is strictly higher than that of the only existing PIR-SI scheme applicable to the PA-PIR-SI setting.

I. INTRODUCTION

In the Private Information Retrieval (PIR) problem, a user wants to obtain one message belonging to a dataset of K messages with copies stored on a single (or multiple) remote server(s), while revealing no information about the identity of the desired message to the server(s). The goal of the user is to privately retrieve their desired message while downloading the minimum possible amount of information from the server(s). It was shown in [1] that in the single-server setting, the user must download the entire dataset in order to achieve the privacy requirement, whereas in the multi-server setting, the user can achieve a much higher download rate. While the maximum achievable download rate—referred to as *capacity*—of single-server PIR was characterized very early on, the capacity of multi-server PIR was left open until the seminal work by Sun and Jafar [2].

In recent years, several variations of PIR have been studied by the coding and information theory community. This includes multi-server PIR [3]–[13], single-server PIR with side information [14]–[23], multi-server PIR with side information [15], [24]–[31], multi-message PIR (MPIR) [32], [33], and MPIR with side information [34]–[39].

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (E-mail: {alexgomezleos, anoosheh}@tamu.edu).

In this work, we revisit the problem of single-server PIR with side information (PIR-SI) [15]. In PIR-SI, the user knows M out of K dataset messages—the identities of which are unknown to the server—as a prior side information, and wants to retrieve one other message without revealing the identity of the desired message to the server. As was shown in [15], the capacity of single-server PIR-SI is given by $\lceil K/(M+1) \rceil^{-1}$. This result hinges on the assumptions that (i) the M side information messages are chosen uniformly at random, and (ii) given these M messages, each of the remaining $K - M$ messages is equally likely to be the message required by the user. While the assumption (i) can be readily justified from the server's perspective, the assumption (ii) may not always be feasible in practice. This is because in many real-world scenarios, not all dataset messages are equally popular. In particular, recent studies show that the Zipf, Gamma, or Weibull distributions are more appropriate statistical models for online data access patterns as compared to the uniform distribution [40]–[42]. This implies the need for new PIR models which take into account the popularity of the dataset messages.

In [43], the authors characterize the capacity of PIR under any arbitrary popularity profile. To the best of our knowledge, there is, however, no prior result on the capacity of PIR-SI under any non-uniform popularity profile. Motivated by this, in this work, we introduce a generalization of the PIR-SI problem, referred to as *Popularity-Aware PIR-SI (PA-PIR-SI)*, which takes into account the popularity of the messages. In particular, the PA-PIR-SI problem reduces to the PIR-SI problem when all the messages are equally popular.

We focus on the single-server setting of the PA-PIR-SI problem, and for the ease of exposition, we assume that K and M are such that $M + 1$ divides K . We establish lower and upper bounds on the capacity of PA-PIR-SI in the single-server setting. In particular, we show that the capacity is upper bounded by $(M + 1)/K$. Note that this upper bound does not depend on the popularity profile, and is indeed the same as the capacity of PIR-SI under the uniform popularity profile when $M + 1$ divides K . To prove the upper bound, we rely on a mix of combinatorial and information-theoretic arguments. To derive a lower bound on the capacity, we propose a PA-PIR-SI scheme, referred to as *Randomized Code Selection (RCS)*, which takes into account the message popularity profile. The RCS scheme takes a novel probabilistic approach—carefully designed based on the popularity of the messages—for selecting between two existing PIR-SI schemes which were proposed in [15].

A motivating example—not presented here due to space constraints—can be found in the long version of this work, [44]. This example highlights the limitations of the existing PIR-SI schemes under a non-uniform popularity profile, and demonstrates how the RCS scheme can overcome these limitations. The RCS scheme is applicable for any arbitrary popularity profile, and achieves a rate strictly higher than $1/(K - M)$ —which is the rate of the only existing PIR-SI scheme applicable for non-uniform popularity profiles, i.e. the MDS Code scheme of [15]. In addition, our simulations for several commonly-used popularity profiles show that when compared to the rate $1/(K - M)$, the rate of the RCS scheme is much closer to the upper bound $(M + 1)/K$.

II. PROBLEM SETUP

We denote random variables by bold symbols, and denote a realization of a random variable by a regular symbol. For a positive integer i , we denote $\{1, 2, \dots, i\}$ by $[i]$. Moreover, for two positive integers $1 \leq i < j$, we denote $\{i, i + 1, \dots, j\}$ by $[i : j]$. For any set T , we denote by $[T]^N$ the set of all N -subsets of T , and denote $[T]^1$ by T for simplicity. We denote by \mathbb{F}_q a finite field of order q , and denote by \mathbb{F}_q^n the n -dimensional vector space over \mathbb{F}_q .

Consider a server that stores a dataset containing K messages X_1, \dots, X_K , where $X_i \in \mathbb{F}_q^n$ for all $i \in [K]$. We assume that the random variables $\mathbf{X}_1, \dots, \mathbf{X}_K$ are independent and uniformly distributed over \mathbb{F}_q^n . Thus, $H(\mathbf{X}_i) = B \triangleq n \log_2 q$ for all $i \in [K]$. For simplicity, we further denote $[K]$ by \mathcal{K} , and denote $\{X_i : i \in T\}$ by X_T for every $T \subseteq \mathcal{K}$.

Consider a user who has prior knowledge of M messages $X_S = \{X_i : i \in S\}$ for some $1 \leq M \leq K - 1$ and some $S \in [\mathcal{K}]^M$, and wishes to retrieve a single message X_W for some $W \in \mathcal{K} \setminus S$.¹ We refer to X_W as the *demand message*, X_S as the *side information messages*, W as the *demand index*, and S as the *side information index set*.

Similarly as in the original setting of PIR-SI [15], we assume that \mathbf{S} is distributed uniformly over $[\mathcal{K}]^M$. That is, the probability mass function (PMF) of \mathbf{S} is given by $p_{\mathbf{S}}(\mathbf{S}^*) \triangleq 1/(\binom{K}{M})$ for all $\mathbf{S}^* \in [\mathcal{K}]^M$. However, we do not assume that the conditional distribution of \mathbf{W} given \mathbf{S} is uniform as in [15]. Instead, we consider a more general setting that subsumes the setting being considered in [15]. For each $i \in \mathcal{K}$, we associate a *popularity* $\lambda_i > 0$ to the message X_i , where λ_i is assumed to be constant with respect to K (i.e., admitting new messages to the dataset does not change the popularity of the existing messages). For instance, λ_i can correspond to the average number of times that the message X_i is requested in a day, week, or month. Without loss of generality, we assume that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$. We denote the tuple $(\lambda_1, \dots, \lambda_K)$ by Λ , and refer to Λ as the *(message) popularity profile*. We also assume that the popularity profile Λ is known by both the user and the server. Note that [15] considers the special case of uniform popularity profile, i.e., $\lambda_1 = \lambda_2 = \dots = \lambda_K$. For simplicity, we denote $\sum_{i \in \mathcal{K} \setminus T} \lambda_i$ by $\lambda_{\bar{T}}$ for any $T \subseteq \mathcal{K}$.

¹We treat W as a singleton (i.e., a set of size 1), instead of an element of a set. Similarly, for the case of $M = 1$, we treat S as a singleton.

Given a popularity profile Λ , the conditional PMF of \mathbf{W} given \mathbf{S} is defined as

$$p_{\mathbf{W}|\mathbf{S}}(\mathbf{W}^*|\mathbf{S}^*) \triangleq \begin{cases} \frac{\lambda_{W^*}}{\lambda_{\bar{S}^*}} & \forall \mathbf{W}^* \in \mathcal{K}, \forall \mathbf{S}^* \in [\mathcal{K} \setminus \mathbf{W}^*]^M, \\ 0 & \text{otherwise.} \end{cases}$$

Note that for fixed \mathbf{S}^* , \mathbf{W} can realize any index W^* in $\mathcal{K} \setminus \mathbf{S}^*$, and the greater is the popularity λ_{W^*} , the higher is the probability of $\mathbf{W} = \mathbf{W}^*$.

Note that the joint PMF of \mathbf{W} and \mathbf{S} is given by

$$p_{\mathbf{W},\mathbf{S}}(\mathbf{W}^*, \mathbf{S}^*) = \begin{cases} \frac{1}{\binom{K}{M}} \frac{\lambda_{W^*}}{\lambda_{\bar{S}^*}} & \forall \mathbf{W}^* \in \mathcal{K}, \forall \mathbf{S}^* \in [\mathcal{K} \setminus \mathbf{W}^*]^M, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

and the PMF of \mathbf{W} is given by

$$p_{\mathbf{W}}(\mathbf{W}^*) = \frac{1}{\binom{K}{M}} \sum_{\mathbf{S}^* \in [\mathcal{K} \setminus \mathbf{W}^*]^M} \frac{\lambda_{W^*}}{\lambda_{\bar{S}^*}} \quad \forall \mathbf{W}^* \in \mathcal{K}. \quad (2)$$

We assume that the joint distribution of \mathbf{W} and \mathbf{S} is known to both the user and the server, whereas the realizations W and S are known only by the user and not the server.

Given the demand index W and the side information index set S , the user sends a query $Q^{[W,S]}$ which is a (potentially stochastic) function of W and S . The server responds with an answer $A^{[W,S]}$ which is a deterministic function of the user's query $Q^{[W,S]}$ and the messages X_1, \dots, X_K . That is, $H(A^{[W,S]} | Q^{[W,S]}, X_{\mathcal{K}}) = 0$. Note that the randomness in $Q^{[W,S]}$ is due to the randomness in the query construction, and the randomness in $A^{[W,S]}$ is due to the randomness in $Q^{[W,S]}$ and $X_{\mathcal{K}}$. For the ease of notation, we denote $Q^{[W,S]}$, $A^{[W,S]}$, $Q^{[W,S]}$, and $A^{[W,S]}$ by \mathbf{Q} , \mathbf{A} , \mathbf{Q} , and \mathbf{A} , respectively.

We require that the query \mathbf{Q} and the answer \mathbf{A} satisfy the following two conditions:

- 1) *Decodability*: Given \mathbf{Q} and X_S , the user must be able to decode the demand X_W from \mathbf{A} , i.e.,

$$H(X_W | \mathbf{A}, \mathbf{Q}, X_S) = 0.$$

- 2) *Privacy*: The server must not gain any information about the demand index W from the query \mathbf{Q} , i.e.,

$$\mathbb{P}(\mathbf{W} = \mathbf{W}^* | \mathbf{Q} = \mathbf{Q}) = \mathbb{P}(\mathbf{W} = \mathbf{W}^*) \quad \forall \mathbf{W}^* \in \mathcal{K}.$$

Given a popularity profile Λ , the problem is to design a protocol for generating $Q^{[W,S]}$ and $A^{[W,S]}$ for any realization (W, S) such that both the decodability and privacy conditions are satisfied. We refer to this problem as *single-server Popularity-Aware Private Information Retrieval with Side Information (PA-PIR-SI)*. Since we focus on the single-server setting, we often omit the term “single-server” for brevity.

We define the *rate* of a PA-PIR-SI protocol as the ratio of the expected amount of information required by the user, i.e., $\sum_{W^* \in \mathcal{K}} p_{\mathbf{W}}(\mathbf{W}^*) H(X_{W^*}) = B$, to the expected amount of information downloaded from the server, i.e., $\sum_{W^* \in \mathcal{K}} \sum_{\mathbf{S}^* \in [\mathcal{K} \setminus \mathbf{W}^*]^M} p_{\mathbf{W},\mathbf{S}}(\mathbf{W}^*, \mathbf{S}^*) H(A^{[W^*, \mathbf{S}^*]})$. For a given Λ , we define the *capacity* as the supremum of rates over all PA-PIR-SI protocols for the popularity profile Λ .

Our goal is to establish lower and upper bounds on the capacity of PA-PIR-SI for any arbitrary popularity profile.

III. MAIN RESULTS

In this section, we summarize our main results on the capacity of PA-PIR-SI.

Theorem 1. For PA-PIR-SI with K messages and M side information messages such that $M + 1$ is a divisor of K and strictly less than \sqrt{K} , under any popularity profile Λ , the capacity is upper bounded by R_{UB} defined as

$$\frac{M + 1}{K}, \quad (3)$$

and is lower bounded by R_{LB} defined as

$$\left(K - M - \left(K - M - \frac{K}{M + 1} \right) \times \Gamma_{\{1\}, [2:M+1]} \frac{p_{W,S}(\{1\}, [2:M+1])}{p_W(\{1\})} \binom{K-1}{M} \right)^{-1}, \quad (4)$$

where $\Gamma_{\{1\}, [2:M+1]}$ is given by

$$\min_{i \in [K-M:K]} \left\{ 1, \frac{p_{W,S}(\{i\}, [K-M:K] \setminus \{i\}) p_W(\{1\})}{p_{W,S}(\{1\}, [2:M+1]) p_W(\{i\})} \right\}, \quad (5)$$

and $p_{W,S}(\cdot, \cdot)$ and $p_W(\cdot)$ depend on the popularity profile Λ , and are defined as in (1) and (2), respectively.

The proof of the upper bound—which is omitted here due to the lack of space, and can be found in [44]—relies on a simple yet strong necessary condition imposed by the decodability and privacy requirements. To prove the lower bound, we build upon the existing PIR-SI schemes, and propose a PA-PIR-SI scheme that is applicable for arbitrary popularity profiles. The proposed scheme takes a randomized approach—carefully tailored to the popularity profile—to select between two different query construction techniques.

Remark 1. Note that the lower bound R_{LB} —which is the rate achieved by our scheme—is valid only for K and M such that $(M + 1) \mid K$ and $M + 1 < \sqrt{K}$, whereas the upper bound R_{UB} holds for all K and M . While our scheme can be modified so that it is applicable for all K and M , the modified scheme's description is lengthy and notation-heavy, and its analysis is tedious and involved. To avoid confusing the reader with technical details, in this work we present the simplest form of our scheme (i.e., for K and M satisfying the above conditions), and demonstrate its superiority over the MDS Code scheme of [15]—which is the only existing PIR-SI scheme applicable for arbitrary popularity profiles.

Remark 2. By the result of [43, Theorem 1] on the capacity of semantic PIR, the capacity of single-server PIR (without side information) under any arbitrary (uniform or non-uniform) popularity profile is $1/K$. That is, the privacy can be achieved only by downloading the entire dataset. While the capacity of single-server PA-PIR-SI remains open in general, the result of Theorem 1 shows that for any popularity profile, the capacity lies between $1/(K - M)$ and

$(M + 1)/K$, and hence, greater than $1/K$. This result extends our prior understanding of the role of side information in single-server PIR-SI under the uniform popularity profile, to arbitrary popularity profiles.

IV. ACHIEVABILITY SCHEME

In this section, we propose a PA-PIR-SI scheme for arbitrary popularity profiles. The proposed scheme, which we refer to as the *Randomized Code Selection (RCS) scheme*, is applicable for any parameters K and M such that $M + 1$ is a divisor of K and strictly less than \sqrt{K} , and any field size $q \geq K$. An illustrative example of the RCS scheme can be found in the long version of this work, [44].

Randomized Code Selection (RCS) Scheme: For any $W^* \in \mathcal{K}$ and $S^* \in [\mathcal{K} \setminus W^*]^M$, we define

$$\Gamma_{W^*, S^*} \triangleq \Gamma_{\{1\}, [2:M+1]} \frac{p_{W,S}(\{1\}, [2:M+1]) p_W(W^*)}{p_{W,S}(W^*, S^*) p_W(\{1\})}, \quad (6)$$

where $\Gamma_{\{1\}, [2:M+1]}$ is given by (5). Given the demand index W and the side information index set S , the user randomly selects the Partition-and-Code scheme of [15] with probability $\Gamma_{W,S}$, or the MDS Code scheme of [15] with probability $1 - \Gamma_{W,S}$, and follows the selected scheme as described below. In the following, we refer to the Partition-and-Code scheme and the MDS Code scheme as Scheme I and Scheme II, respectively.

Scheme I: The user partitions the message indices $1, \dots, K$ into $N \triangleq K/(M + 1)$ parts Q_1, \dots, Q_N , each of size $M + 1$, as outlined below. First, the user chooses an index $j^* \in [N]$ uniformly at random, and assigns the indices in $W \cup S$ to the part Q_{j^*} . The user then takes the remaining $K - (M + 1)$ message indices $\mathcal{K} \setminus (W \cup S)$, and randomly partitions them into the remaining $N - 1$ parts Q_j 's for $j \in [N] \setminus \{j^*\}$. Then, the user constructs the query $Q^{[W,S]} = \{Q_1, \dots, Q_N\}$, and sends it to the server.

Given $Q^{[W,S]}$, the server computes $A_j = \sum_{i \in Q_j} X_i$ for each $j \in [N]$. Then, the server constructs the answer $A^{[W,S]} = \{A_1, \dots, A_N\}$, and sends it back to the user.

Given $A^{[W,S]}$, the user recovers their demand message X_W by subtracting off the contribution of the side information messages X_S from A_{j^*} , i.e., $X_W = A_{j^*} - \sum_{i \in S} X_i$.

Scheme II: First, the user chooses K arbitrary (but distinct) elements $\omega_1, \dots, \omega_K$ from \mathbb{F}_q , and constructs $K - M$ vectors Q_1, \dots, Q_{K-M} , where $Q_j = [\omega_1^{j-1}, \dots, \omega_K^{j-1}]$ for each $j \in [K - M]$. Then, the user constructs the query $Q^{[W,S]} = \{Q_1, \dots, Q_{K-M}\}$, and sends it to the server.

Given $Q^{[W,S]}$, the server computes $A_j = \sum_{i=1}^K \omega_i^{j-1} X_i$ for each $j \in [K - M]$. The server then constructs the answer $A^{[W,S]} = \{A_1, \dots, A_{K-M}\}$, and sends it back to the user.

Given $A^{[W,S]}$, the user recovers their demand message X_W —along with all $K - (M + 1)$ messages $X_{\mathcal{K} \setminus (W \cup S)}$ —by subtracting off the contribution of the side information messages X_S from A_1, \dots, A_{K-M} , and solving the resulting system of $K - M$ equations with $K - M$ unknowns $X_{\mathcal{K} \setminus S}$.

A. Proof of Decodability and Privacy

Since both Schemes I and II satisfy the decodability condition, it should be obvious that the RCS scheme also satisfies this requirement. It thus remains to show that the RCS scheme also satisfies the privacy condition.

Consider a query constructed by the RCS scheme. When the query is formed by Scheme II, it should be obvious that the privacy condition is satisfied because Scheme II constructs the query independently of the realization (W, S) . In the following, we show that the privacy condition is also satisfied when the query is formed by Scheme I.

Recall that any query formed by Scheme I is a partition of \mathcal{K} with $N = K/(M+1)$ parts, each of size $M+1$. We denote by \mathcal{Q} the set of all such partitions. For each $Q \in \mathcal{Q}$, let Q_1, \dots, Q_N denote the N parts forming the partition Q .

Lemma 1. *For any query (partition) $Q \in \mathcal{Q}$, the privacy condition is satisfied if for any $i, j \in [N]$ and for any $W_i \in Q_i, W_j \in Q_j$, it holds that*

$$\Gamma_{W_j, S_j} = \Gamma_{W_i, S_i} \frac{p_{W, S}(W_i, S_i) p_W(W_j)}{p_{W, S}(W_j, S_j) p_W(W_i)}, \quad (7)$$

where $S_i = Q_i \setminus W_i$ and $S_j = Q_j \setminus W_j$.

Proof. The proof can be found in [44]. \square

By Lemma 1, the privacy requirement entails that the condition in (7) must hold for any two parts Q_i and Q_j in any partition Q . For the proof of privacy, it thus suffices to show that our choice of Γ_{W^*, S^*} in (6) satisfies (7).

Fix arbitrary $W^* \in \mathcal{K}$ and $S^* \in [\mathcal{K} \setminus W^*]^M$ such that $W^* \cup S^*$ is one of the parts in the partition Q . We consider the following cases separately: (i) $Q_i = [M+1]$ for some $i \in [N]$, and (ii) $Q_i \neq [M+1]$ for any $i \in [N]$.

First, consider the case (i). Taking $W_i = \{1\}$ and $S_i = [2 : M+1]$, the condition in (7) reduces to

$$\Gamma_{W^*, S^*} = \Gamma_{\{1\}, [2 : M+1]} \frac{p_{W, S}(\{1\}, [2 : M+1]) p_W(W^*)}{p_{W, S}(W^*, S^*) p_W(\{1\})},$$

which is consistent with our choice of Γ_{W^*, S^*} (cf. (6)).

Next, consider the case (ii). Recall that by assumption, $M+1 < \sqrt{K}$, i.e., $K = N(M+1) > (M+1)^2$, or equivalently, $N > M+1$. Since the partition Q consists of $N > M+1$ parts, by the pigeonhole principle, there exists some $k \in [N]$ such that Q_k and $[M+1]$ are disjoint. Let $Q^* \in \mathcal{Q}$ be an arbitrary partition such that both parts Q_k and $[M+1]$ belong to the partition Q^* . Recall that the privacy condition requires that for any given partition, the condition in (7) must hold for any two parts of that partition. Note also that Q_k and $[M+1]$ are two parts of the same partition Q^* . Let W_k be an arbitrary index in the part Q_k , and let $S_k = Q_k \setminus W_k$. Then, by (7), it is required that

$$\Gamma_{W_k, S_k} = \Gamma_{\{1\}, [2 : M+1]} \frac{p_{W, S}(\{1\}, [2 : M+1]) p_W(W_k)}{p_{W, S}(W_k, S_k) p_W(\{1\})}. \quad (8)$$

Note also that $W^* \cup S^*$ and $W_k \cup S_k$ are two parts of the partition Q . Thus, by (7), we require that

$$\Gamma_{W^*, S^*} = \Gamma_{W_k, S_k} \frac{p_{W, S}(W_k, S_k) p_W(W^*)}{p_{W, S}(W^*, S^*) p_W(W_k)}. \quad (9)$$

Combining (8) and (9), it follows that we must have

$$\Gamma_{W^*, S^*} = \Gamma_{\{1\}, [2 : M+1]} \frac{p_{W, S}(\{1\}, [2 : M+1]) p_W(W^*)}{p_{W, S}(W^*, S^*) p_W(\{1\})},$$

which coincides with our choice of Γ_{W^*, S^*} (cf. (6)). This completes the proof of privacy.

B. Proof of Achievable Rate

By construction, the server's answer to the user's query consists of $K/(M+1)$ (or $K-M$) linearly independent combinations of the messages X_1, \dots, X_K for Scheme I (or Scheme II). Since X_1, \dots, X_K are independent and uniformly distributed over \mathbb{F}_q^n , then $\mathbf{A}_1, \dots, \mathbf{A}_{K/(M+1)}$ (or $\mathbf{A}_1, \dots, \mathbf{A}_{K-M}$) are independent and uniformly distributed over \mathbb{F}_q^n . Thus, $H(\mathbf{A}^{[W, S]})$ is equal to $H(\mathbf{A}_1, \dots, \mathbf{A}_{K/(M+1)}) = (K/(M+1))B$ or $H(\mathbf{A}_1, \dots, \mathbf{A}_{K-M}) = (K-M)B$ for Scheme I or Scheme II, respectively. Using the joint PMF of \mathbf{W} and \mathbf{S} , the rate of the RCS scheme is given by

$$\left(\sum_{W^* \in \mathcal{K}, S^* \in [\mathcal{K} \setminus W^*]^M} p_{W, S}(W^*, S^*) \times \left[\Gamma_{W^*, S^*} \left(\frac{K}{M+1} \right) + (1 - \Gamma_{W^*, S^*}) (K-M) \right] \right)^{-1}. \quad (10)$$

Substituting for Γ_{W^*, S^*} using (6), we can rewrite (10) as

$$\left(K-M - \left(K-M - \frac{K}{M+1} \right) \times \Gamma_{\{1\}, [2 : M+1]} \frac{p_{W, S}(\{1\}, [2 : M+1])}{p_W(\{1\})} \binom{K-1}{M} \right)^{-1}, \quad (11)$$

which is the expression for R_{LB} in Theorem 1 (cf. (4)).

Since the probability Γ_{W^*, S^*} takes a value in the interval $[0, 1]$, it readily follows that $\Gamma_{\{1\}, [2 : M+1]}$ is lower bounded by 0, and upper bounded by

$$\min_{W^*, S^*} \left\{ 1, \frac{p_{W, S}(W^*, S^*) p_W(\{1\})}{p_{W, S}(\{1\}, [2 : M+1]) p_W(W^*)} \right\}, \quad (12)$$

where the minimization is over all $W^* \in \mathcal{K}$ and all $S^* \in [\mathcal{K} \setminus W^*]^M$. According to (11), for fixed K and M , the rate of the RCS scheme is an increasing function of $\Gamma_{\{1\}, [2 : M+1]}$, and hence, the rate is maximized when $\Gamma_{\{1\}, [2 : M+1]}$ is equal to (12). It remains to show that (12) and our choice of $\Gamma_{\{1\}, [2 : M+1]}$ in (5) are equal.

Instead of working directly with (12), it is more convenient to analyze the following minimization problem:

$$\min_{W^*, S^*} \left\{ \frac{p_{W, S}(\{1\}, [2 : M+1])}{p_W(\{1\})}, \frac{p_{W, S}(W^*, S^*)}{p_W(W^*)} \right\}, \quad (13)$$

where the minimization is over all $W^* \in \mathcal{K}$ and all $S^* \in [\mathcal{K} \setminus W^*]^M$. Note that (12) is equal to (13) times the constant term $p_{W,S}(\{1\})/p_{W,S}(\{1\}, [2 : M+1])$. By (1) and (2), we have

$$\frac{p_{W,S}(W^*, S^*)}{p_W(W^*)} = \frac{1}{\lambda_{S^*}} \left(\sum_{T \in [\mathcal{K} \setminus W^*]^M} \frac{1}{\lambda_T} \right)^{-1}. \quad (14)$$

For any given $W^* \in \mathcal{K}$, it is easy to see that (14) is minimized for $S^* \in [\mathcal{K} \setminus W^*]^M$ such that λ_{S^*} is maximum, or equivalently, $\lambda_{S^*} \triangleq \sum_{i \in S^*} \lambda_i$ is minimum. For any given W^* , we can determine S^* that minimizes λ_{S^*} as follows. Recall that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$ by assumption. We consider the following two cases separately: (i) $W^* \in [1 : K-M]$, and (ii) $W^* \in [K-M+1 : K]$. In the case (i), λ_{S^*} is minimized for $S^* = [K-M+1 : K]$. This is because the sum of the last M λ_i 's yields the minimum sum over all M -subsets of $\{\lambda_i : i \in \mathcal{K}\}$. In the case (ii), λ_{S^*} is minimized for $S^* = [K-M : K] \setminus W^*$. This is because W^* is one of the last $M+1$ indices in \mathcal{K} , and the M -subset S^* cannot contain W^* . Hence, we can rewrite (13) as

$$\min \left\{ \frac{p_{W,S}(\{1\}, [2 : M+1])}{p_W(\{1\})}, \min_{i \in [1:K-M]} \frac{p_{W,S}(\{i\}, [K-M+1 : K])}{p_W(\{i\})}, \min_{i \in [K-M+1:K]} \frac{p_{W,S}(\{i\}, [K-M : K] \setminus \{i\})}{p_W(\{i\})} \right\}. \quad (15)$$

Lemma 2. For any popularity profile $(\lambda_1, \dots, \lambda_K)$ such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K > 0$, it holds that

$$\begin{aligned} & \min_{i \in [1:K-M]} \frac{p_{W,S}(\{i\}, [K-M+1 : K])}{p_W(\{i\})} \\ &= \frac{p_{W,S}(\{K-M\}, [K-M+1 : K])}{p_W(\{K-M\})}. \end{aligned}$$

Proof. The proof can be found in [44]. \square

By the result of Lemma 2, the minimization problem in (15) can be simplified further as

$$\min \left\{ \frac{p_{W,S}(\{1\}, [2 : M+1])}{p_W(\{1\})}, \min_{i \in [K-M:K]} \frac{p_{W,S}(\{i\}, [K-M : K] \setminus \{i\})}{p_W(\{i\})} \right\},$$

or equivalently,

$$\min_{i \in [K-M:K]} \left\{ \frac{p_{W,S}(\{1\}, [2 : M+1])}{p_W(\{1\})}, \frac{p_{W,S}(\{i\}, [K-M : K] \setminus \{i\})}{p_W(\{i\})} \right\}. \quad (16)$$

Since (13), (15), and (16) are equal, and (12) is equal to (13) times $p_W(\{1\})/p_{W,S}(\{1\}, [2 : M+1])$, it then follows that (12) and (5) are equal, as was to be shown.

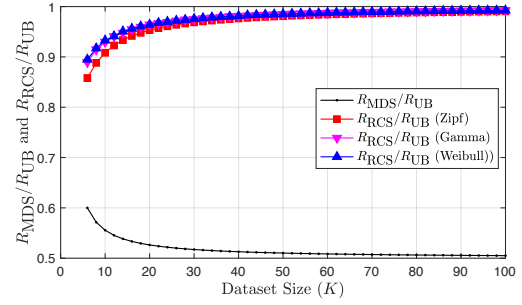


Fig. 1. The ratios R_{RCS}/R_{UB} and R_{MDS}/R_{UB} versus K , for $M = 1$ and different models for the popularity profile.

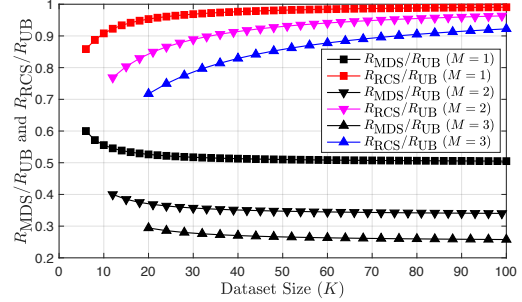


Fig. 2. The ratios R_{RCS}/R_{UB} and R_{MDS}/R_{UB} versus K , for different M and the Zipf model for the popularity profile.

V. SIMULATIONS

In this section, we compare the rate of the RCS scheme and that of the MDS Code scheme of [15], with respect to the capacity upper bound R_{UB} (see (3)). In the following, we denote the rates of the RCS scheme and the MDS Code scheme by R_{RCS} and R_{MDS} , respectively. Note that $R_{RCS} = R_{LB}$ (see (4)), and $R_{MDS} = 1/(K-M)$ (from [15]).

It is generally agreed that the Zipf, Gamma, and Weibull distributions are appropriate models for the popularity profile [40]–[42]. Motivated by this, in our simulations we consider popularity profiles generated according to each of these distributions. In addition, we consider very small values of M which are of significant practical importance.

Fig. 1 depicts R_{RCS}/R_{UB} and R_{MDS}/R_{UB} , for $M = 1$ and different K , where $\lambda_1, \dots, \lambda_K$ are sampled from each of the following distributions: (i) Zipf with parameters $N = 100$ and $s = 1$, (ii) Gamma with shape and scale parameters 0.62 and 31.22, respectively, and (iii) Weibull with shape and scale parameters 0.79 and 16.80, respectively. (The above parameters were chosen such that all three distributions have the same mean and the same variance.) As can be seen in Fig. 1, for a fixed distribution, as K increases, R_{RCS}/R_{UB} approaches 1, whereas R_{MDS}/R_{UB} approaches $1/2$.

Fig. 2 depicts R_{RCS}/R_{UB} and R_{MDS}/R_{UB} for $M \in \{1, 2, 3\}$ and different K , where $\lambda_1, \dots, \lambda_K$ are sampled from the Zipf distribution with parameters $N = 100$ and $s = 1$. In Fig. 2, one can observe that for each M , as K increases, R_{RCS}/R_{UB} approaches 1, while R_{MDS}/R_{UB} approaches $1/(M+1)$. It can also be seen that for fixed K (or M), the advantage of the RCS scheme over the MDS Code scheme is more pronounced as M (or K) increases.

REFERENCES

- [1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private Information Retrieval," in *IEEE Symposium on Foundations of Computer Science*, 1995, pp. 41–50.
- [2] H. Sun and S. A. Jafar, "The Capacity of Private Information Retrieval," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4075–4088, July 2017.
- [3] R. Tajeddine and S. El Rouayheb, "Robust Private Information Retrieval on Coded Data," in *IEEE International Symposium on Information Theory*, 2017.
- [4] R. Tajeddine, O. W. Gnille, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. E. Rouayheb, "Private Information Retrieval Schemes for Coded Data With Arbitrary Collusion Patterns," in *IEEE International Symposium on Information Theory*, June 2017, pp. 1908–1912.
- [5] H. Sun and S. A. Jafar, "The Capacity of Private Computation," *IEEE Trans. on Info. Theory*, vol. 65, no. 6, pp. 3880–3897, 2019.
- [6] K. Banawan and S. Ulukus, "The Capacity of Private Information Retrieval from Coded Databases," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1945–1956, March 2018.
- [7] C. Tian, H. Sun, and J. Chen, "Capacity-Achieving Private Information Retrieval Codes With Optimal Message Size and Upload Cost," *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 7613–7627, 2019.
- [8] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus, "The Capacity of Private Information Retrieval From Heterogeneous Uncoded Caching Databases," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3407–3416, 2020.
- [9] J. Lavauzelle, R. Tajeddine, R. Freij-Hollanti, and C. Hollanti, "Private Information Retrieval Schemes With Product-Matrix MBR Codes," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 441–450, 2021.
- [10] M. Shrivastava and P. Sarvepalli, "Capacity Achieving Uncoded PIR Protocol based on Combinatorial Designs," March 2021. [Online]. Available: arXiv:2103.09804
- [11] K. Banawan, A. Arafat, and S. Ulukus, "Timely Private Information Retrieval," *arXiv e-prints*, p. arXiv:2105.08713, May 2021.
- [12] R. Zhou, C. Tian, H. Sun, and J. S. Plank, "Two-Level Private Information Retrieval," in *IEEE International Symposium on Information Theory*, 2021, pp. 1919–1924.
- [13] Y. Lu, Z. Jia, and S. A. Jafar, "Double Blind T-Private Information Retrieval," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 428–440, 2021.
- [14] S. Kadhe, B. Garcia, A. Heidarzadeh, S. E. Rouayheb, and A. Sprintson, "Private Information Retrieval With Side Information: The Single Server Case," in *55th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2017, pp. 1099–1106.
- [15] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private Information Retrieval With Side Information," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2032–2043, 2020.
- [16] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "The Role of Coded Side Information in Single-Server Private Information Retrieval," *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 25–44, 2021.
- [17] —, "Capacity of Single-Server Single-Message Private Information Retrieval With Coded Side Information," in *IEEE Information Theory Workshop*, Nov 2018.
- [18] —, "Capacity of Single-Server Single-Message Private Information Retrieval With Private Coded Side Information," in *IEEE International Symposium on Information Theory*, July 2019, pp. 1662–1666.
- [19] S. Kadhe, A. Heidarzadeh, A. Sprintson, and O. O. Koyluoglu, "On an Equivalence Between Single-Server PIR With Side Information and Locally Recoverable Codes," in *IEEE Information Theory Workshop*, 2019.
- [20] S. Kadhe, A. Heidarzadeh, A. Sprintson, and O. O. Koyluoglu, "Single-Server Private Information Retrieval Schemes are Equivalent to Locally Recoverable Coding Schemes," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 391–402, 2021.
- [21] A. Heidarzadeh and A. Sprintson, "Single-Server Individually-Private Information Retrieval: A Combinatorial Approach," in *IEEE Information Theory Workshop*, 2021.
- [22] —, "The Role of Reusable and Single-Use Side Information in Private Information Retrieval," Jan 2022. [Online]. Available: arXiv:2201.11605
- [23] Y. Lu and S. A. Jafar, "On Single Server Private Information Retrieval With Private Coded Side Information," Feb 2022. [Online]. Available: arXiv:2202.07693
- [24] R. Tandon, "The Capacity of Cache Aided Private Information Retrieval," in *55th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2017, pp. 1078–1082.
- [25] Y. Wei, K. Banawan, and S. Ulukus, "Cache-Aided Private Information Retrieval With Partially Known Uncoded Prefetching: Fundamental Limits," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1126–1139, June 2018.
- [26] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental Limits of Cache-Aided Private Information Retrieval With Unknown and Uncoded Prefetching," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3215–3232, 2019.
- [27] F. Kazemi, E. Karimi, A. Heidarzadeh, and A. Sprintson, "Multi-Server Private Information Retrieval With Coded Side Information," in *Canadian Workshop on Information Theory*, 2019.
- [28] —, "Private Information Retrieval With Private Coded Side Information: The Multi-Server Case," in *57th Annual Allerton Conference on Communication, Control, and Computing*, 2019, pp. 1098–1104.
- [29] Z. Chen, Z. Wang, and S. A. Jafar, "The Capacity of T-Private Information Retrieval With Private Side Information," *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 4761–4773, 2020.
- [30] S. Li and M. Gastpar, "Converse for Multi-Server Single-Message PIR With Side Information," in *54th Annual Conference on Information Sciences and Systems*, 2020, pp. 1–6.
- [31] Murali Krishnan K. H. and J. Harshan, "XOR-Based Codes for Private Information Retrieval With Private Side Information," May 2021. [Online]. Available: arXiv:2105.05788
- [32] K. Banawan and S. Ulukus, "Multi-Message Private Information Retrieval: Capacity Results and Near-Optimal Schemes," *IEEE Transactions on Information Theory*, vol. 64, no. 10, pp. 6842–6862, Oct 2018.
- [33] K. Banawan and S. Ulukus, "Multi-Message Private Information Retrieval," in *IEEE International Symposium on Information Theory*, June 2017, pp. 1898–1902.
- [34] A. Heidarzadeh, S. Kadhe, B. Garcia, S. E. Rouayheb, and A. Sprintson, "On the Capacity of Single-Server Multi-Message Private Information Retrieval With Side Information," in *56th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2018.
- [35] S. Li and M. Gastpar, "Single-Server Multi-Message Private Information Retrieval With Side Information," in *56th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2018.
- [36] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-Message Private Information Retrieval With Private Side Information," in *IEEE Information Theory Workshop*, 2018.
- [37] A. Heidarzadeh, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "Single-Server Multi-Message Individually-Private Information Retrieval With Side Information," in *IEEE International Symposium on Information Theory*, July 2019, pp. 1042–1046.
- [38] F. Kazemi, E. Karimi, A. Heidarzadeh, and A. Sprintson, "Single-Server Single-Message Online Private Information Retrieval With Side Information," in *IEEE International Symposium on Information Theory*, July 2019, pp. 350–354.
- [39] A. Heidarzadeh and A. Sprintson, "The Linear Capacity of Single-Server Individually-Private Information Retrieval With Side Information," Feb 2022. [Online]. Available: arXiv:2202.12229
- [40] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1357–1370, Oct 2009.
- [41] X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of YouTube Videos," in *IEEE/ACM International Workshop on Quality of Service*, 2008.
- [42] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-Like Distributions: Evidence and Implications," in *Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '99)*, 1999.
- [43] S. Vithana, K. Banawan, and S. Ulukus, "Semantic Private Information Retrieval: Effects of Heterogeneous Message Sizes and Popularities," in *IEEE Global Communications Conference*, 2020, pp. 1–6.
- [44] A. Gomez-Leos and A. Heidarzadeh, "Single-Server Private Information Retrieval With Side Information Under Arbitrary Popularity Profiles," May 2022. [Online]. Available: arXiv:2205.06172