The Capacity of 4-Star-Graph PIR

Yuhang Yao and Syed A. Jafar
University of California, Irvine
Center for Pervasive Communication and Computing (CPCC)
Email: {yuhangy5, syed}@uci.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206729

***Abstract*—Introduced by Sadeh et al., the K-star-graph private information retrieval (PIR) problem, so-labeled because the storage graph is a star-graph with K leaf nodes, is comprised of K messages that are stored separately (one-each) at K dedicated servers, and a universal server that stores all K messages, for a total of $K+1$ servers. While it is one of the simplest PIR settings to describe, the capacity C_K of K-star-graph PIR is open for $K \geq 4$. We study the critical $K = 4$ setting, for which prior work establishes the bounds $2/5 \leq C_4 \leq 3/7$. As our main contribution, we characterize the exact capacity of 4-star-graph PIR as $C_4 = 5/12$, thus improving upon both the prior lower-bound as well as the prior upper-bound. The main technical challenge resides in the new converse bound, whose non-trivial structure is deduced indirectly from the achievable schemes that emerge from the study of a finer tradeoff between the download costs from the dedicated servers versus the universal server. A sharp characterization of this tradeoff is also obtained for $K = 4$.**

I. Introduction

There is much interest in the capacity of various Private Information Retrieval (PIR) settings [1] motivated not only by growing privacy concerns, but also, and perhaps even more so, by the fundamental connections of PIR to other ideas — locally decodable codes and interference alignment to name a few — that provide opportunities for cross-cutting insights. What makes problems like PIR attractive for information theoretic analysis is that they are simple to describe (allowing broader connections), but challenging to solve (allowing deeper insights) — a trait also evident in the celebrated index coding problem [2], similarly simple-to-describe and broadly insightful. Following this thought, a particularly appealing PIR setting is K-star-graph PIR, introduced by Sadeh et al. in [3].

The name 'K-star-graph PIR' reflects that the storage graph[1] is a star-graph with K leaf nodes. As illustrated in Fig. 1, the setting is comprised of K messages, $\mathsf{W}_1, \cdots, \mathsf{W}_K$, say of size L bits each, that are stored separately (one-each) at K dedicated servers, Server $1, \cdots$, Server K, respectively, and a universal server, Server 0, that stores all K messages, for a total of $N = K+1$ servers.[2] A user generates a private and uniform index $\theta \in \{1, 2, \cdots, K\}$, and wishes to retrieve the desired message W_θ with the smallest total download cost possible, without revealing any information about θ to any server. The total download cost $\Delta_\Sigma \triangleq \Delta_0 + \Delta_1 + \cdots + \Delta_K = \Delta_0 + K\Delta$, where Δ_k is the download cost from Server k, $k \in \{0, 1, \cdots, K\}$, and we assume (without loss of generality) that the download costs from the dedicated servers are balanced, i.e., $\Delta_1 = \Delta_2 = \cdots = \Delta_K \triangleq \Delta$. The maximum number of desired message bits that can be retrieved per bit of total download is the capacity C_K of K-star-graph PIR. The capacity C_K is characterized approximately for large K by Sadeh et al. in [3] as $\Theta(1/\sqrt{K})$, and exactly for $K \leq 3$. However, in spite of the descriptive simplicity of K-star-graph PIR, a precise capacity characterization is difficult to obtain for $K \geq 4$. The capacity C_4 for the critical case of 4-star-graph PIR, where we have 4 messages stored among 5 servers, is our focus in this work.

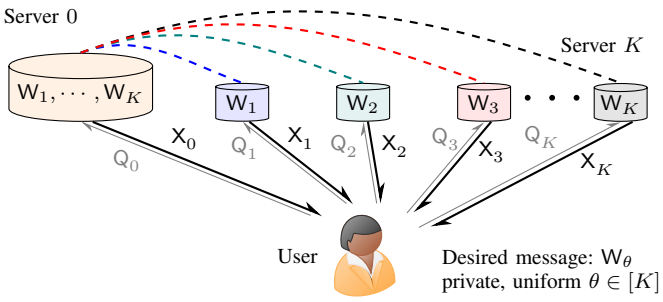

Fig. 1. K-star-graph PIR with messages $\mathsf{W}_1, \cdots, \mathsf{W}_K$ stored among $K+1$ servers according to the K-star storage-graph illustrated with dashed-edges.

To establish a basic understanding, let us start with a simpler setting for which the capacity is known, i.e., 3-star-graph PIR. For notational simplicity, let $\mathsf{A}, \mathsf{B}, \mathsf{C}$ represent the 3 messages $\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3$, respectively. Also let $\theta = a, b, c$ denote the demand being $\mathsf{A}, \mathsf{B}, \mathsf{C}$, respectively. Consider the following coding scheme, where $L = 3$ bits of each message are encoded together, e.g., $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3)$. Let $(\mathsf{a}_1, \mathsf{a}_2, \mathsf{a}_3), (\mathsf{b}_1, \mathsf{b}_2, \mathsf{b}_3), (\mathsf{c}_1, \mathsf{c}_2, \mathsf{c}_3)$ be three i.i.d. uniform permutations of $(1, 2, 3)$ generated privately by the user. Let the download from Server 0 be

$$\mathsf{X}_0 = \left[\mathsf{A}_{\mathsf{a}_1} + \mathsf{B}_{\mathsf{b}_1}, \quad \mathsf{A}_{\mathsf{a}_2} + \mathsf{C}_{\mathsf{c}_1}, \quad \mathsf{B}_{\mathsf{b}_2} + \mathsf{C}_{\mathsf{c}_2}\right], \tag{1}$$

which is trivially private as it does not depend on θ. For each possible desired message $\mathsf{W}_\theta \in \{\mathsf{A}, \mathsf{B}, \mathsf{C}\}$, the downloads from the dedicated servers are specified in the form of the tuples $(\mathsf{W}_\theta, \mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3)$ as $(\mathsf{A}, \mathsf{A}_{\mathsf{a}_3}, \mathsf{B}_{\mathsf{b}_1}, \mathsf{C}_{\mathsf{c}_1})$, $(\mathsf{B}, \mathsf{A}_{\mathsf{a}_1}, \mathsf{B}_{\mathsf{b}_3}, \mathsf{C}_{\mathsf{c}_2})$, and $(\mathsf{C}, \mathsf{A}_{\mathsf{a}_2}, \mathsf{B}_{\mathsf{b}_2}, \mathsf{C}_{\mathsf{c}_3})$. It is readily verified that in every case the

[1] A storage graph has the servers as vertices, and an edge (hyperedge) for each message, identifying servers that store that message (cf. [4]–[6]).

[2] While we identify the K-star-graph with the number of messages K, i.e., the number of leaf-nodes in the star-graph, note that reference [3] identifies the same graph as $S_N = S_{K+1}$, based on the number of servers $N = K+1$, i.e., the total number of nodes in the star graph. Thus, our 4-star-graph corresponds to S_5 in [3]. We avoid the S_N notation because it is interpreted differently by different authors (e.g., see Wikipedia entry on star graphs).

user is able to retrieve all 3 bits of the desired message from the downloads. The scheme is private because regardless of θ the download from each of Servers $1, 2, 3$ is equally likely to be any of the 3 bits of the message stored at that server. The rate achieved by the scheme is $L/\Delta_\Sigma = 3/6 = 1/2$ since $L = 3$ bits of the desired message are retrieved at a total download cost of $\Delta_\Sigma = \Delta_0 + \Delta_1 + \Delta_2 + \Delta_3 = 3+1+1+1 = 6$ bits. In fact the rate $1/2$ is optimal, i.e., the capacity $C_3 = 1/2$. As a rough intuitive sketch, the converse is based on the following reasoning. Since the query to Server 0 is independent of θ, it follows that there exist some queries $(\overline{Q}_0^a, \overline{Q}_1^a, \overline{Q}_2^a, \overline{Q}_3^a)$ and $(\widetilde{Q}_0^b, \widetilde{Q}_1^b, \widetilde{Q}_2^b, \widetilde{Q}_3^b)$ such that from the corresponding downloads $(\overline{\mathsf{X}}_0^a, \overline{\mathsf{X}}_1^a, \overline{\mathsf{X}}_2^a, \overline{\mathsf{X}}_3^a)$ the user can recover A, from $(\widetilde{\mathsf{X}}_0^b, \widetilde{\mathsf{X}}_1^b, \widetilde{\mathsf{X}}_2^b, \widetilde{\mathsf{X}}_3^b)$ the user can recover B, and $\overline{Q}_0^a = \widetilde{Q}_0^b$ so that $\overline{\mathsf{X}}_0^a = \widetilde{\mathsf{X}}_0^b$. Now, we claim that from $\mathsf{C}, \overline{\mathsf{X}}_0^a, \overline{\mathsf{X}}_1^a, \overline{\mathsf{X}}_2^a, \widetilde{\mathsf{X}}_2^b$ the user can recover all 3 messages $\mathsf{A}, \mathsf{B}, \mathsf{C}$. This claim is justified as follows. The user is already given C. Because the user is given $\overline{\mathsf{X}}_0^a, \overline{\mathsf{X}}_1^a, \overline{\mathsf{X}}_2^a$, and can generate $\overline{\mathsf{X}}_3^a$ locally from C, she must be able to recover A. Then, having both A and C at this point, the user can obtain locally anything it needs from Server 1 and Server 3, i.e., servers that store only A and C. In particular, the user can locally obtain $\widetilde{\mathsf{X}}_1^b$ from A, and $\widetilde{\mathsf{X}}_3^b$ from C, and is given both $\widetilde{\mathsf{X}}_0^b = \overline{\mathsf{X}}_0^a$, and $\widetilde{\mathsf{X}}_2^b$. This allows the user to recover B, thus completing the recovery of all three messages $\mathsf{A}, \mathsf{B}, \mathsf{C}$. Now, by the general property $H(f(\mathsf{X})) \leq H(\mathsf{X})$, it follows that $3L = H(\mathsf{A}, \mathsf{B}, \mathsf{C}) \leq H(\mathsf{C}, \overline{\mathsf{X}}_0^a, \overline{\mathsf{X}}_1^a, \overline{\mathsf{X}}_2^a, \widetilde{\mathsf{X}}_2^b) \leq H(\mathsf{C}) + \Delta_0 + \Delta_1 + 2\Delta_2 = L + \Delta_0 + 3\Delta = L + \Delta_\Sigma$, which gives us $3L \leq L + \Delta_\Sigma$, and therefore we obtain the desired converse bound $\Delta_\Sigma/L \geq 2$.

Now consider the open problem that is the focus of this work, i.e., 4-star-graph PIR. Sadeh et al. establish in [3] that the rate $2/5$ is achievable, so we have $C_5 \geq 2/5$. Generalization of the converse reasoning from [3] as sketched above, yields the bound $C_4 \leq 3/7$. Thus we have a gap, $2/5 \leq C_4 \leq 3/7$. Closing this gap is the immediate goal of this work, with the long-term hope that the new ideas that emerge in the process may help with future generalizations, e.g., leading to C_K for $K > 4$, and subsequently to converse bounds for related interference alignment schemes that arise in other contexts (discussed in Section V). As our main contribution we settle the precise capacity of 4-star-graph PIR as $C_4 = 5/12$. Thus, the capacity strictly improves upon both the prior lower-bound and the prior upper-bound. The main technical challenge is manifested in a new converse bound, $3\Delta_0 + 14\Delta \geq 8L$, that has a remarkably distinct structure from previous converse bounds. The structure of this new bound is deduced indirectly from the achievable schemes that emerge from the study of a finer tradeoff between the download costs from the dedicated servers (Δ) versus the download cost from the universal server (Δ_0). A sharp characterization of this tradeoff is also obtained for $K = 4$.

Notation: For a positive integer K, $[K]$ denotes the set of positive integers $\{1, 2, \ldots, K\}$. $\binom{n}{r} \triangleq \frac{n!}{r!(n-r)!}$ with the convention that $\binom{n}{r} = 0$ if $n < r$ or $r < 0$. Sans-serif letters are used to denote random variables.

II. Problem Statement: K-star-graph PIR

K Messages, $K+1$ Servers: There are K independent messages $\mathsf{W}_1, \mathsf{W}_2, \ldots, \mathsf{W}_K$. Each message is a stream of $\mathbb{F}_2$ symbols. For $\ell \in \mathbb{N}$, $\mathsf{W}_k(\ell)$ denotes the ℓ^{th} symbol of $\mathsf{W}_k, k \in [K]$. Each symbol is drawn i.i.d. uniform. There are K dedicated servers that store one message each, with Server k storing only W_k, $k \in [K]$. There is one universal server, called Server 0 that stores all K messages.

User's demand θ, local randomness Z: The demand $\theta \in [K]$ is generated privately and uniformly by the user. Additional local randomness Z may be generated privately by the user, independent of θ. The user wishes to retrieve W_θ.

Coding scheme: A coding scheme $\mathcal{C}$ for K-star-graph PIR is formally represented as a tuple $\mathcal{C} = (L, \mathsf{Z}, \mu_0, \mu_1, \cdots, \mu_K, \phi_0, \phi_1, \cdots, \phi_K, \psi)$ where $L \in \mathbb{N}$ is the number of bits of each message to be encoded together, $\mathsf{Z} \in \mathcal{Z}$ is the local randomness generated privately by the user (independent of θ), $\mu_0, \cdots, \mu_K$ are functions that generate queries sent by the user to the servers, $\phi_0, \cdots, \phi_K$ are functions that generate answers returned by the servers to the user, and ψ is the final decoding function at the user. In order to be considered *feasible*, a coding scheme must satisfy correctness and privacy constraints. Let us denote the set of feasible coding schemes for K-star-graph PIR by $\mathcal{C}_K$. These functions and constraints are described next.

Queries: The $K+1$ functions $\mu_k : [K] \times \mathcal{Z} \to \mathcal{Q}_k$, $\forall k \in \{0, 1, \cdots, K\}$, map (θ, Z) to the $K+1$ queries $\mathsf{Q}_0, \mathsf{Q}_1, \cdots, \mathsf{Q}_K$, respectively, i.e.,

$$\mathsf{Q}_k = \mu_k(\theta, \mathsf{Z}), \qquad \forall k \in \{0, 1, \cdots, K\}. \tag{2}$$

The query functions must satisfy the following which requires that the query to any server must be independent of the desired message index θ, i.e.,

$$I(\theta; \mathsf{Q}_k) = 0, \qquad \forall k \in \{0, 1, \cdots, K\}. \tag{3}$$

Answers from Servers: The $K+1$ functions, $\phi_0 : [(\mathbb{F}_2^L)^K] \times \mathcal{Q}_0 \to \mathcal{X}_0$, and $\phi_k : \mathbb{F}_2^L \times \mathcal{Q}_k \to \mathcal{X}_k, \forall k \in [K]$, map the stored message(s) and the query at each server to the answer that the server returns to the user, i.e.,

$$\mathsf{X}_0 = \phi_0(\mathsf{W}_1, \mathsf{W}_2, \ldots, \mathsf{W}_K, \ \mathsf{Q}_0), \tag{4}$$

$$\mathsf{X}_k = \phi_k(\mathsf{W}_k, \mathsf{Q}_k), \qquad k \in [K]. \tag{5}$$

Decoding by the User: The decoding function,

$$\psi : \mathcal{X}_0 \times \mathcal{X}_1 \times \ldots \times \mathcal{X}_K \times [K] \times \mathcal{Z} \to \mathbb{F}_2^L \tag{6}$$

allows the user to retrieve the desired message W_θ,

$$\mathsf{W}_\theta = \psi(\mathsf{X}_0, \mathsf{X}_1, \ldots, \mathsf{X}_K, \theta, \mathsf{Z}). \tag{7}$$

Download Cost: The download cost (in bits) from Server k is defined as[3],

$$\Delta_k \triangleq \log_2 |\mathcal{X}_k|, \qquad \forall k \in \{0, 1, \cdots, K\}. \tag{8}$$

[3]Since each alphabet set $\mathcal{X}_k$ is deterministic, our focus in this work is limited to the maximum (instead of average) download costs across queries.

Rates, Capacity: The rate of a feasible coding scheme $\mathcal{C} = (L, \mathsf{Z}, \mu_0, \mu_1, \cdots, \mu_K, \phi_0, \phi_1, \cdots, \phi_K, \psi) \in \mathcal{C}_K$ is defined as,

$$R(\mathcal{C}) = \frac{L}{\Delta_0 + \Delta_1 + \ldots + \Delta_K}. \tag{9}$$

The capacity of K-star-graph PIR is defined as

$$C_K \triangleq \sup_{\mathcal{C} \in \mathcal{C}_K} R(\mathcal{C}). \tag{10}$$

Balanced Costs: Let $\overline{\mathcal{C}}_K$ be the set of all feasible schemes with *balanced* costs from the dedicated servers, i.e.,

$$|\mathcal{X}_1| = |\mathcal{X}_2| = \cdots = |\mathcal{X}_K| \triangleq |\mathcal{X}| \tag{11}$$

$$\Delta_1 = \Delta_2 = \cdots = \Delta_K \triangleq \Delta. \tag{12}$$

For capacity, it follows from [7, Thm 4] that there is no loss of generality in restricting the coding schemes to $\overline{\mathcal{C}}_K$, i.e.,

$$C_K = \sup_{\mathcal{C} \in \mathcal{C}_K} R(\mathcal{C}) = \sup_{\mathcal{C}' \in \overline{\mathcal{C}}_K} R(\mathcal{C}'). \tag{13}$$

This is because any feasible coding scheme $\mathcal{C} = (L, \mathsf{Z}, \mu_0, \mu_1, \cdots, \mu_K, \phi_0, \phi_1, \cdots, \phi_K, \psi) \in \mathcal{C}_K$ that does not have balanced costs, can be mapped to a feasible scheme $\mathcal{C}' \in \mathcal{C}_K$ with balanced costs, while preserving the rate of the original scheme, i.e., $R(\mathcal{C}') = R(\mathcal{C})$, by a straightforward equal '*time-sharing*' of schemes (cf. [7, Thm 4]) $\mathcal{C}_\pi = (L, \mathsf{Z}_\pi, \mu_0, \mu_{\pi(1)}, \cdots, \mu_{\pi(K)}, \phi_0, \phi_{\pi(1)}, \cdots, \phi_{\pi(K)}, \psi) \in \mathcal{C}_K$, for all permutations $\pi : [K] \to [K]$, with independent local randomness Z_π for each $\mathcal{C}_\pi$. In fact, it suffices to consider only cyclic permutations, i.e., all $\pi \in \{\pi_1, \cdots, \pi_K\}$ where $\pi_i(k) = 1 + ((k+i) \mod K), \forall i, k \in [K]$. There are K cyclic permutations, producing K such schemes, for a combined scheme $\mathcal{C}'$ that considers $L' = KL$ bit messages, and has a download cost $\Delta'_0 = K\Delta_0$, $\Delta'_k = \Delta_{\pi_1(k)} + \cdots + \Delta_{\pi_K(k)} = \Delta_1 + \cdots + \Delta_K = \Delta'$ for all $k \in [K]$. The rate of the new scheme $\mathcal{C}'$ is,

$$R(\mathcal{C}') = \frac{L'}{\Delta'_0 + \Delta'_1 + \cdots + \Delta'_K} \tag{14}$$

$$= \frac{KL}{K\Delta_0 + K(\Delta_1 + \cdots + \Delta_K)} \tag{15}$$

$$= \frac{L}{\Delta_0 + \Delta_1 + \cdots + \Delta_K} = R(\mathcal{C}) \tag{16}$$

i.e., same as the original scheme $\mathcal{C}$, but the new scheme $\mathcal{C}'$ has balanced download costs from the dedicated servers. The correctness of $\mathcal{C}'$ follows directly from the correctness of $\mathcal{C}$. The privacy of $\mathcal{C}'$ follows from the reasoning that the queries to a server in each $\mathcal{C}_{\pi_i}$ are individually independent of θ (because the original scheme $\mathcal{C}$ is private), and conditioned on θ the queries to the same server in the different $\mathcal{C}_{\pi_i}$ schemes are independent because Z_{π_i} are independent. From this it follows[4] that the combined-query in $\mathcal{C}'$ is independent of θ.

[4]For example, consider queries to Server 1 for schemes $\mathcal{C}_{\pi_1}, \mathcal{C}_{\pi_2}$, namely $\mathsf{Q}_1^{\pi_1}, \mathsf{Q}_2^{\pi_2}$. We have $I(\theta; \mathsf{Q}_1^{\pi_1}) = I(\theta; \mathsf{Q}_1^{\pi_2}) = 0$, $I(\mathsf{Q}_1^{\pi_1}; \mathsf{Q}_1^{\pi_2} \mid \theta) = 0$. This implies that $H(\mathsf{Q}_1^{\pi_1}, \mathsf{Q}_1^{\pi_2} \mid \theta) = H(\mathsf{Q}_1^{\pi_1} \mid \theta) + H(\mathsf{Q}_1^{\pi_2} \mid \theta) - I(\mathsf{Q}_1^{\pi_1}; \mathsf{Q}_1^{\pi_2} \mid \theta) = H(\mathsf{Q}_1^{\pi_1}) + H(\mathsf{Q}_1^{\pi_2}) \geq H(\mathsf{Q}_1^{\pi_1}, \mathsf{Q}_1^{\pi_2})$. Since conditioning cannot increase entropy, we have $H(\mathsf{Q}_1^{\pi_1}, \mathsf{Q}_1^{\pi_2} \mid \theta) = H(\mathsf{Q}_1^{\pi_1}, \mathsf{Q}_1^{\pi_2})$, i.e., θ is independent of the combined query $(\mathsf{Q}_1^{\pi_1}, \mathsf{Q}_1^{\pi_2})$ to Server 1. The reasoning extends to any server, and to the combination of all $\mathcal{C}_{\pi_i}$, i.e., $\mathcal{C}'$.

Normalized Feasible Cost Region: Let us define the normalized feasible cost region as the set of tuples $(\Delta_0/L, \Delta/L)$ that are achievable within ϵ for all positive epsilon.

$$\mathcal{D}_K^* \triangleq \left\{ (\overline{\Delta}_0, \overline{\Delta}) \in \mathbb{R}_2^+ : \begin{array}{c} \forall \epsilon > 0, \exists\, \mathcal{C} \in \overline{\mathcal{C}}_K, \text{ s.t.} \\ \Delta_0/L \geq \overline{\Delta}_0 - \epsilon, \\ \Delta/L \geq \overline{\Delta} - \epsilon. \end{array} \right\} \tag{17}$$

III. Capacity and $(\Delta_0/L, \Delta/L)$ Tradeoff

The converse arguments of Sadeh et al. [3], along the lines summarized in the introduction, yield the bounds,

$$\Delta_0 + (1 + 2 + \cdots + t)\,\Delta \geq tL, \qquad \forall t \leq K, \tag{18}$$

for all $K \geq 1$. It is not difficult to show that these bounds characterize the normalized feasible cost region $\mathcal{D}_K^*$ for $K \leq 3$. The boundary of $\mathcal{D}_K^*$ is is shown in the figures below for $K = 2$ and $K = 3$. Points above the boundary are feasible, and points below the boundary (shaded region) are not feasible.

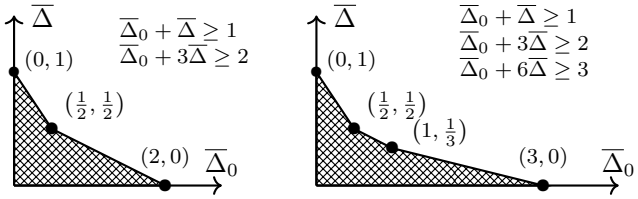

Fig. 2. Normalized feasible cost regions $\mathcal{D}_2^*$ (left) and $\mathcal{D}_3^*$ (right). (*Optimal*)

For $K = 4$, the converse bounds in [3] yield the constraints $\Delta_0 + \Delta \geq L$, $\Delta_0 + 3\Delta \geq 2L$, $\Delta_0 + 6\Delta \geq 3L$ and $\Delta_0 + 10\Delta \geq 4L$. Minimizing the total download cost $\Delta_\Sigma = \Delta_0 + 4\Delta$ subject to these constraints yields the converse bound $\Delta_\Sigma \geq 7L/3$, i.e., $C_4 \leq 3/7$.

On the other hand, the best achievable rate from [3] is $2/(K+1) = 2/5$, so we have the bound $C_4 \geq 2/5$. Let us denote $\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \mathsf{W}_4$ as $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ (respectively) for ease of exposition, and recall that messages have length L bits, e.g., $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2, \cdots, \mathsf{A}_L)$. Let $(\mathsf{a}_1, \cdots, \mathsf{a}_L)$, $(\mathsf{b}_1, \cdots, \mathsf{b}_L)$, $(\mathsf{c}_1, \cdots, \mathsf{c}_L)$ and $(\mathsf{d}_1, \cdots, \mathsf{d}_L)$ be i.i.d. uniform permutations of $(1, 2, \cdots, L)$ generated privately by the user. The key idea of achievable schemes in [3] as well as the generalizations in this work, is the following: The download from the universal server, Server 0, is comprised of sums of symbols of subsets of messages. The download from the dedicated server that stores the desired message, is comprised of those symbols that do not appear in X_0. The download from each dedicated server that stores an undesired message is comprised of precisely those symbols of that message which appear in linear combination with the desired message symbols in X_0. Thus, downloads from undesired messages are used for interference cancellation in X_0, while those from the desired message retrieve missing symbols. The privately permuted indices $\mathsf{a}_i, \mathsf{b}_i, \mathsf{c}_i, \mathsf{d}_i$ guarantee privacy, as the download from each dedicated server is uniform over the message symbols. Following this principle, there are two schemes in [3], each achieving the rate $2/5$, as follows.

Scheme 1 [3]: $(\Delta_0, \Delta) = (1, 1), L = 2$

The download from Server 0,

$$\mathsf{X}_0 = \left[\mathsf{A}_{a_1} + \mathsf{B}_{b_1} + \mathsf{C}_{c_1} + \mathsf{D}_{d_1}\right]. \tag{19}$$

In terms of the tuples $(\mathsf{W}_\theta, \mathsf{X}_1, \mathsf{X}_2, \mathsf{X}_3, \mathsf{X}_4)$ the scheme is summarized as $(\mathsf{A}, \mathsf{A}_{a_2}, \mathsf{B}_{b_1}, \mathsf{C}_{c_1}, \mathsf{D}_{d_1})$, $(\mathsf{B}, \mathsf{A}_{a_1}, \mathsf{B}_{b_2}, \mathsf{C}_{c_1}, \mathsf{D}_{d_1})$, $(\mathsf{C}, \mathsf{A}_{a_1}, \mathsf{B}_{b_1}, \mathsf{C}_{c_2}, \mathsf{D}_{d_1})$, and $(\mathsf{D}, \mathsf{A}_{a_1}, \mathsf{B}_{b_1}, \mathsf{C}_{c_1}, \mathsf{D}_{d_2})$. Since the user retrieves the $L = 2$ bits of the desired messages with the download cost $\Delta_\Sigma = 5$ bits, the rate is $2/5$.

Scheme 2 [3]: $(\Delta_0, \Delta) = (6, 1), L = 4$

The download from Server 0 is expressed compactly as,

$$\mathsf{X}_0 = \begin{bmatrix} \mathsf{A}_{a_1} + \mathsf{B}_{b_1}, & \mathsf{A}_{a_2} + \mathsf{C}_{c_1}, & \mathsf{A}_{a_3} + \mathsf{D}_{d_1} \\ \mathsf{B}_{b_2} + \mathsf{C}_{c_2}, & \mathsf{B}_{b_3} + \mathsf{D}_{d_2}, & \mathsf{C}_{c_3} + \mathsf{D}_{d_3} \end{bmatrix} \tag{20}$$

According to the general idea described earlier in this section, for example if B is the desired message, the user downloads $\mathsf{A}_{a_1}, \mathsf{B}_{b_4}, \mathsf{C}_{c_2}, \mathsf{D}_{d_2}$ from Servers $1, 2, 3, 4$, respectively. The rate is $L/(\Delta_0 + 4\Delta) = 4/(6 + 4) = 2/5$ for this scheme as well.

Our main result appears in the following theorem.

Theorem 1. *The normalized feasible cost region $\mathcal{D}_4^*$ for 4-star-graph PIR is the set of all $(\overline{\Delta}_0, \overline{\Delta}) \in \mathbb{R}_2^+$, such that,*

$$\overline{\Delta}_0 + \overline{\Delta} \geq 1, \tag{21}$$

$$\overline{\Delta}_0 + 3\overline{\Delta} \geq 2, \tag{22}$$

$$\overline{\Delta}_0 + 10\overline{\Delta} \geq 4, \tag{23}$$

$$3\overline{\Delta}_0 + 14\overline{\Delta} \geq 8. \tag{24}$$

That is, $(\overline{\Delta}_0, \overline{\Delta})$ tuples that satisfy (21)-(24) *are feasible, and $(\overline{\Delta}_0, \overline{\Delta})$ tuples that do not satisfy* (21)-(24) *are not feasible. Furthermore, the capacity of 4-star-graph PIR, $C_4 = 5/12$.*

We note that the inequalities (21)-(23) are already established as converse bounds in [3]. The key to Theorem 1 is the bound (24), that is conspicuously distinct from the remaining bounds, and is the main reason that 4-star-graph PIR is considerably more challenging. The form of this bound becomes intuitive if we consider first the achievability perspective. Inspired by Scheme 1 and Scheme 2, let us present a new scheme, Scheme 3, that is capacity achieving.

Scheme 3 (Capacity-achieving): $(\Delta_0, \Delta) = (4, 2), L = 5$

The download from Server 0 is expressed as,

$$\mathsf{X}_0 = \begin{bmatrix} \mathsf{A}_{a_1} + \mathsf{B}_{b_1} + \mathsf{C}_{c_1}, & \mathsf{A}_{a_2} + \mathsf{B}_{b_2} + \mathsf{D}_{d_1} \\ \mathsf{A}_{a_3} + \mathsf{C}_{c_2} + \mathsf{D}_{d_2}, & \mathsf{B}_{b_3} + \mathsf{C}_{c_3} + \mathsf{D}_{d_3} \end{bmatrix}. \tag{25}$$

The scheme follows the same principle as Scheme 1 and Scheme 2, e.g., if C is the desired message, the user downloads $\mathsf{A}_{a_1}, \mathsf{A}_{a_3}, \mathsf{B}_{b_1}, \mathsf{B}_{b_3}, \mathsf{C}_{c_4}, \mathsf{C}_{c_5}, \mathsf{D}_{d_2}, \mathsf{D}_{d_3}$ from Servers $1, 2, 3, 4$, respectively. The rate is $L/(\Delta_0 + 4\Delta) = 5/(4 + 8) = 5/12$ for this scheme. This proves that $C_4 \geq 5/12$.

Considering the $(\overline{\Delta}_0, \overline{\Delta})$ tuples for the three schemes, Scheme 1 achieves $(1/2, 1/2)$, Scheme 2 achieves $(3/2, 1/4)$, and Scheme 3 achieves $(4/5, 2/5)$. The tuples $(0, 1)$ and $(4, 0)$ are trivial. By time-sharing (cf. [7, Thm. 4], [8, Lem. 1])) we achieve the boundary of the entire $\mathcal{D}_K^*$ region specified by the bounds (21)-(24), and shown in Fig. 3. This completes the proof of achievability for Theorem 1.

The proof of (24) was found by using symmetry and insights from alternative caching formulations in conjunction with the CAI software [9]. The proof is presented separately in the next section. From (24), we have that (22)$\times 2$+(24) $\implies \Delta_\Sigma/L = \overline{\Delta}_0 + 4\overline{\Delta} \geq 12/5 \implies C_4 \leq 5/12$. Therefore, the capacity $C_4 = 5/12$.

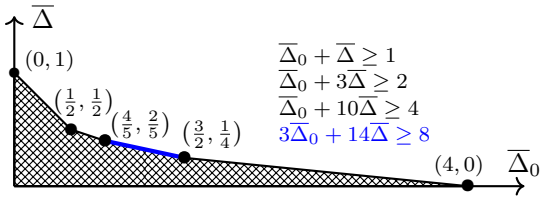

Fig. 3. Normalized feasible cost region $\mathcal{D}_4^*$ for $K = 4$. (*Optimal*)

From Fig. 3 we note that the boundary of the $\mathcal{D}_4^*$ region is comprised of line segments corresponding to the bounds (21)-(24). In particular, the highlighted blue segment in Fig. 3 corresponds precisely to the desired new bound (24).

Generalized Schemes for K Messages: $t \in \{1, \ldots, K\}$, $(\Delta_0, \Delta) = \left(\binom{K}{t}, \binom{K-2}{t-2}\right), L = \binom{K-1}{t-1} + \binom{K-2}{t-2}$

We generalize the idea to the setting with K messages. Let $t \in [K]$. For the scheme parameterized by t, the download from Server 0 is expressed as

$$\mathsf{X}_0 = \left\{\mathsf{W}_{k_1}(\sigma_{k_1}^{j_1}) + \cdots + \mathsf{W}_{k_t}(\sigma_{k_t}^{j_t}) \;\middle|\; \{k_1, \cdots, k_t\} \subset [K]\right\},$$

where $\sigma_k = (\sigma_k^1, \cdots, \sigma_k^L)$, $k \in [K]$ is a uniform random permutation of $(1, 2, \cdots, L)$. $j_1, \cdots, j_t \in [L]$ are chosen such that $\sigma_k^j, j \in [L]$ appears at most once in X_0. It follows that $\Delta_0 = \binom{K}{t}$ since there are $\binom{K}{t}$ terms in X_0. The downloads from the dedicated servers follow the same rule as in the above examples. The download cost from the servers that do not store the desired message is $\binom{K-2}{t-2}$ since it is the number of interference terms corresponding to one server. The download cost from the server that stores W_θ is $L - \binom{K-1}{t-1} = \binom{K-2}{t-2}$ since it is the number of bits in W_θ that does not appear in X_0. Thus, the costs from the dedicated servers are balanced and $\Delta = \binom{K-2}{t-2}$.

For the K-star-graph PIR, the above generalized scheme provides K achievable tuples of $(\overline{\Delta}_0, \overline{\Delta})$. Note that $(0, 1)$ is trivial and thus we have $K+1$ achievable tuples. Time-sharing between adjacent tuples results in the achievable region

$$\mathcal{D}_K^{\text{achi}} = \left\{ (\overline{\Delta}_0, \overline{\Delta}) : \begin{array}{c} \binom{t+1}{2}\overline{\Delta}_0 + \left(\binom{K}{2} + Kt\right)\overline{\Delta} \geq Kt, \\ \text{for } t \in \{1, 2, \cdots, K-1\}; \\ \overline{\Delta}_0 + \overline{\Delta} \geq 1. \end{array} \right\}.$$

From previous works and Theorem 1 we have $\mathcal{D}_K^* = \mathcal{D}_K^{\text{achi}}$ for K up to 4. It remains open whether $\mathcal{D}_K^* = \mathcal{D}_K^{\text{achi}}$ for $K > 4$.

IV. Proof of the bound: $3\overline{\Delta}_0 + 14\overline{\Delta} \geq 8$

Due to the privacy constraint regarding Server 0, there must be such $z_1, z_2, z_3, z_4 \in \mathcal{Z}$ and a resulting query $\overline{Q}_0$ that

$$\mu_0(1, z_1) = \mu_0(2, z_2) = \mu_0(3, z_3) = \mu_0(4, z_4) = \overline{Q}_0, \quad (26)$$

because otherwise some information about θ is revealed by Q_0, thus violating the privacy constraint (3). For $k \in [4], s \in [4]$, let $\overline{Q}_k^s \triangleq \mu_k(s, z_s)$. Then let

$$\overline{\mathsf{X}}_0 \triangleq \phi_0(\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \mathsf{W}_4, \ \overline{Q}_0), \quad (27)$$

$$\overline{\mathsf{X}}_k^s \triangleq \phi_k\big(\mathsf{W}_k, \overline{Q}_k^s\big), \quad \forall k \in [4], \ s \in [4]. \quad (28)$$

This means that for $s \in [4]$, when $\mathsf{Z} = z_s$ and $\theta = s$, the answer from Server 0 is $\overline{\mathsf{X}}_0$, and the answer from Server $k \in [K]$ is $\overline{\mathsf{X}}_k^s$. (28) implies that

$$H(\overline{\mathsf{X}}_k^s | \mathsf{W}_k) = 0, \ \forall k \in [4], \ s \in [4]. \quad (29)$$

Note that (7) implies

$$\mathsf{W}_s = \psi(\overline{\mathsf{X}}_0, \overline{\mathsf{X}}_1^s, \overline{\mathsf{X}}_2^s, \overline{\mathsf{X}}_3^s, \overline{\mathsf{X}}_4^s, s, z_s). \quad (30)$$

Since z_1, z_2, z_3, z_4 are not random, (30) implies that W_s is a function of $(\overline{\mathsf{X}}_0, \overline{\mathsf{X}}_1^s, \overline{\mathsf{X}}_2^s, \overline{\mathsf{X}}_3^s, \overline{\mathsf{X}}_4^s)$, i.e.,

$$H(\mathsf{W}_s | \overline{\mathsf{X}}_0, \overline{\mathsf{X}}_1^s, \overline{\mathsf{X}}_2^s, \overline{\mathsf{X}}_3^s, \overline{\mathsf{X}}_4^s) = 0, \ \forall s \in [4]. \quad (31)$$

The proof of (24) is then as follows. First,

$$\begin{aligned} &H(\mathsf{W}_1) + H(\overline{\mathsf{X}}_2^2) + H(\overline{\mathsf{X}}_3^2) + H(\overline{\mathsf{X}}_4^2) + H(\overline{\mathsf{X}}_0) \\ &\geq H(\mathsf{W}_1, \overline{\mathsf{X}}_2^2, \overline{\mathsf{X}}_3^2, \overline{\mathsf{X}}_4^2, \overline{\mathsf{X}}_0) \\ &\overset{(29)(31)}{\geq} \underbrace{H(\mathsf{W}_1, \mathsf{W}_2, \overline{\mathsf{X}}_3^2, \overline{\mathsf{X}}_4^2, \overline{\mathsf{X}}_0)}_{T_1}. \end{aligned} \quad (32)$$

Due to symmetry, we have

$$\begin{aligned} &H(\mathsf{W}_3) + H(\overline{\mathsf{X}}_1^1) + H(\overline{\mathsf{X}}_2^1) + H(\overline{\mathsf{X}}_4^1) + H(\overline{\mathsf{X}}_0) \\ &\geq \underbrace{H(\mathsf{W}_3, \mathsf{W}_1, \overline{\mathsf{X}}_2^1, \overline{\mathsf{X}}_4^1, \overline{\mathsf{X}}_0)}_{T_2}, \end{aligned} \quad (33)$$

$$\begin{aligned} &H(\mathsf{W}_4) + H(\overline{\mathsf{X}}_1^1) + H(\overline{\mathsf{X}}_2^1) + H(\overline{\mathsf{X}}_3^1) + H(\overline{\mathsf{X}}_0) \\ &\geq \underbrace{H(\mathsf{W}_4, \mathsf{W}_1, \overline{\mathsf{X}}_2^1, \overline{\mathsf{X}}_3^1, \overline{\mathsf{X}}_0)}_{T_3}. \end{aligned} \quad (34)$$

Then, by submodularity and (31),

$$\begin{aligned} &T_1 + T_2 \\ &\geq \underbrace{H(\mathsf{W}_1, \overline{\mathsf{X}}_3^2, \overline{\mathsf{X}}_2^1, \overline{\mathsf{X}}_0)}_{T_4} + \underbrace{H(\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \overline{\mathsf{X}}_4^1, \overline{\mathsf{X}}_4^2, \overline{\mathsf{X}}_0)}_{T_5}. \end{aligned} \quad (35)$$

Again by submodularity,

$$\begin{aligned} &T_3 + T_4 \\ &\geq \underbrace{H(\mathsf{W}_1, \overline{\mathsf{X}}_2^1, \overline{\mathsf{X}}_0)}_{T_6} + \underbrace{H(\mathsf{W}_1, \mathsf{W}_4, \overline{\mathsf{X}}_2^1, \overline{\mathsf{X}}_3^1, \overline{\mathsf{X}}_3^2, \overline{\mathsf{X}}_0)}_{T_7}. \end{aligned} \quad (36)$$

Then

$$T_5 + H(\overline{\mathsf{X}}_4^4) \overset{(29)(31)}{\geq} H(\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \mathsf{W}_4), \quad (37)$$

and by submodularity

$$\begin{aligned} &T_6 + H(\mathsf{W}_2) + H(\overline{\mathsf{X}}_3^3) + H(\overline{\mathsf{X}}_4^3) + H(\overline{\mathsf{X}}_4^4) \\ &\geq T_6 + H(\mathsf{W}_2, \overline{\mathsf{X}}_3^3, \overline{\mathsf{X}}_4^3, \overline{\mathsf{X}}_4^4, \overline{\mathsf{X}}_0) \\ &\overset{(31)}{\geq} H(\overline{\mathsf{X}}_2^1) + H(\mathsf{W}_1, \mathsf{W}_2, \overline{\mathsf{X}}_3^3, \overline{\mathsf{X}}_4^3, \overline{\mathsf{X}}_4^4, \overline{\mathsf{X}}_0) \\ &\overset{(29)(31)}{\geq} H(\overline{\mathsf{X}}_2^1) + H(\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \mathsf{W}_4), \end{aligned} \quad (38)$$

and finally

$$\begin{aligned} &T_7 + H(\overline{\mathsf{X}}_2^2) + H(\overline{\mathsf{X}}_3^3) \\ &\geq H(\mathsf{W}_1, \mathsf{W}_4, \overline{\mathsf{X}}_2^1, \overline{\mathsf{X}}_3^1, \overline{\mathsf{X}}_3^2, \overline{\mathsf{X}}_2^2, \overline{\mathsf{X}}_3^3, \overline{\mathsf{X}}_0) \\ &\overset{(29)(31)}{\geq} H(\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \mathsf{W}_4). \end{aligned} \quad (39)$$

Adding (32), (33), (34), (35), (36), (37), (38), (39) (two terms of $H(\overline{\mathsf{X}}_2^1)$ are canceled from both sides), we have

$$\begin{aligned} &2H(\overline{\mathsf{X}}_1^1) + 2H(\overline{\mathsf{X}}_2^2) + 2H(\overline{\mathsf{X}}_3^3) + 2H(\overline{\mathsf{X}}_4^4) \\ &+ H(\overline{\mathsf{X}}_2^1) + H(\overline{\mathsf{X}}_3^1) + H(\overline{\mathsf{X}}_4^1) + H(\overline{\mathsf{X}}_3^2) + H(\overline{\mathsf{X}}_4^2) + H(\overline{\mathsf{X}}_4^3) \\ &+ 3H(\overline{\mathsf{X}}_0) + H(\mathsf{W}_1) + H(\mathsf{W}_2) + H(\mathsf{W}_3) + H(\mathsf{W}_4) \\ &\geq 3H(\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \mathsf{W}_4). \end{aligned} \quad (40)$$

Note that $\overline{\mathsf{X}}_0 \in \mathcal{X}_0$, $\overline{\mathsf{X}}_k^s \in \mathcal{X}_k, \forall k \in [4], s \in [4]$. With (8) and (12), we have that $\Delta_0 \geq H(\overline{\mathsf{X}}_0)$, $\Delta \geq H(\overline{\mathsf{X}}_k^s)$, $\forall k \in [4], s \in [4]$, and by the definition of the messages, $H(\mathsf{W}_1) = H(\mathsf{W}_2) = H(\mathsf{W}_3) = H(\mathsf{W}_4) = L$, $H(\mathsf{W}_1, \mathsf{W}_2, \mathsf{W}_3, \mathsf{W}_4) = 4L$. Combining this with (40), we conclude that

$$14\Delta + 3\Delta_0 + 4L \geq 12L \implies (24). \quad (41)$$

V. Conclusion

The capacity as well as the normalized feasible cost tradeoff between the dedicated servers and the universal server are found for $K = 4$. The achievable scheme, generalized to the setting with $K > 4$ messages, offers clues to the general capacity result for $K > 4$ for future work. Generalizing the insights from star-graph PIR to other communication/computation problems is especially of interest. For example, retrospective interference alignment schemes for K-user interference channel [10] are similar to star-graph PIR, and proofs of optimality are not yet available for such schemes. In these schemes there are two phases. The first phase typically yields random linear combinations of messages at the receivers, whereas the second phase involves transmissions from individual transmitters, one at a time. Conceptually, the first phase is quite similar to the downloads from the universal server in K-star-graph PIR, while the second phase is analogous to the downloads from each dedicated server. Remarkably, the approximate rate of K-star-graph PIR characterized by Sadeh et al. in [3] as $\Theta(1/\sqrt{K})$, matches the approximate degrees of freedom per user achieved by the retrospective interference alignment scheme in [10], indicative of a deeper connection between these problems.

Acknowledgment

This work was supported in part by research grants NSF CCF-1907053, CCF-2221379, and ONR N00014-21-1-2386.

REFERENCES

[1] S. Ulukus, S. Avestimehr, M. Gastpar, S. A. Jafar, R. Tandon, and C. Tian, "Private Retrieval, Computing, and Learning: Recent Progress and Future Challenges," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 3, pp. 729–748, 2022.

[2] M. Effros, S. El Rouayheb, and M. Langberg, "An Equivalence Between Network Coding and Index Coding," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2478–2487, 2015.

[3] B. Sadeh, Y. Gu, and I. Tamo, "Bounds on the Capacity of PIR over Graphs," 2021. [Online]. Available: https://arxiv.org/abs/2105.07704

[4] N. Raviv, I. Tamo, and E. Yaakobi, "Private Information Retrieval in Graph-Based Replication Systems," *IEEE Transactions on Information Theory*, vol. 66, no. 6, pp. 3590–3602, 2020.

[5] K. Banawan and S. Ulukus, "Private information retrieval from non-replicated databases," *International Symposium on Information Theory (ISIT)*, July 2019.

[6] Z. Jia and S. A. Jafar, "On the Asymptotic Capacity of X-Secure T-Private Information Retrieval With Graph-Based Replicated Storage," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6280–6296, 2020.

[7] H. Sun and S. A. Jafar, "Optimal Download Cost of Private Information Retrieval for Arbitrary Message Length," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2920–2932, 2017.

[8] R. Tandon, "The capacity of cache aided private information retrieval," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2017, pp. 1078–1082.

[9] C. Tian, J. S. Plank, B. Hurst, and R. Zhou, "Computational Techniques for Investigating Information Theoretic Limits of Information Systems," *Information*, vol. 12, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2078-2489/12/2/82

[10] D. Castanheira, A. Silva, and A. Gameiro, "Retrospective Interference Alignment for the K -User $M \times N$ MIMO Interference Channel," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8368–8379, 2016.