Single-Server Multi-Message Individually-Private Information Retrieval with Side Information

Anoosheh Heidarzadeh, Swanand Kadhe, Salim El Rouayheb, and Alex Sprintson

***Abstract*— We consider a multi-user variant of the private information retrieval problem described as follows. Suppose there are D users, each of which wants to privately retrieve a distinct message from a server with the help of a *trusted agent*. We assume that the agent has a subset of M messages whose indices are unknown to the server. The goal of the agent is to collectively retrieve the users' requests from the server. For this problem, we introduce the notion of *individual-privacy* – the agent is required to protect the privacy only for each individual user (but may leak some correlations among user requests). We refer to this problem as *Individually-Private Information Retrieval with Side Information (IPIR-SI)*.**

We first establish a lower bound on the capacity, which is defined as the maximum achievable download rate, of the IPIR-SI problem by presenting a novel achievability protocol. Next, we characterize the capacity of IPIR-SI problem for $M = 1$ and $D = 2$. In the process of characterizing the capacity for arbitrary M and D we present a novel combinatorial conjecture, that may be of independent interest.

I. INTRODUCTION

In the conventional Private Information Retrieval (PIR) problem, a user wants to privately download a message belonging to a database with copies stored on a single or multiple remote servers (see [1]). The multiple-server PIR problem has been predominantly studied in the PIR literature, with breakthrough results for the information-theoretic privacy model in the past few years (see e.g., [2]–[5], and references therein). The multi-message extension of the PIR problem enables a user to privately download multiple messages from the server(s) [6], [7]. There have been a number of recent works on the PIR problem when some side information is present at the user [7]–[11].

Recently, in [12], [13], the authors considered the single-server PIR with Side Information (PIR-SI) problem, wherein the user knows a random subset of messages whose indices are unknown to the server. It was shown that the side information enables the user to substantially reduce the download cost and still achieve information-theoretic privacy for the requested message. The multi-message version of PIR-SI is considered in [14], [15], and the case of coded side information is considered in [16]. Single-server multi-user PIR-SI problem wherein all users have the same demand but different side-information sets was considered in [17].

In this work, we consider the following scenario. Suppose there are D users, each of which wants to privately retrieve a distinct message from a server. The users send their demands to a *trusted agent*, who is an entity that makes a profit by offering privacy to users. The agent has a subset of M messages whose indices are unknown to the server. This side information could have been obtained in several ways, e.g., from the current users and/or the users in the past, or from previous interactions with multiple (yet not-presently-available) servers storing identical copies of the database. Followed by aggregating the users' requests, the agent then collectively retrieves information from the server.

One natural solution for the agent to achieve privacy during the retrieval is to successively use the PIR-SI protocol in [12] for each request. However, the agent can achieve much higher download rate while preserving the privacy collectively for all the users by using the multi-message PIR protocol in [14], [15]. In this work, we introduce the notion of *individual-privacy* where the agent is required to protect the privacy only for each individual user, and we refer to this problem as *Individually-Private Information Retrieval with Side Information (IPIR-SI)*. We seek to answer the following questions: is it possible to further increase the download rate when individual-privacy is required? Moreover, what are the fundamental limits on the download rate for the IPIR-SI problem? We answer the first question affirmatively and take the first steps towards answering the second question.

A. Main Contributions

We first establish a lower bound on the capacity (defined as the supremum of all achievable download rates) of the IPIR-SI problem by presenting a new protocol which builds up on the Generalized Partition and Code (GPC) protocol of [14]. Next, we characterize the capacity of IPIR-SI problem for $M = 1$ and $D = 2$. In the process of characterizing the capacity for arbitrary M and D we present a novel combinatorial conjecture, that may be of independent interest.

For $M = 1$ and arbitrary D, our conjecture, rephrased in the language of graph theory, relates the size of an external mother vertex-set of any directed graph G whose nodes have certain in-degree and out-degree, to the size of an internal mother vertex-set of the transpose of G, where the

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

S. Kadhe is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA (E-mail: swanand.kadhe@berkeley.edu).

S. El Rouayheb is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854 USA (E-mail: sye8@soe.rutgers.edu).

The work of A. Heidarzadeh and A. Sprintson was supported in part by NSF Grants No. 1718658 and 1642983. The work of S. Kadhe was supported in part by NSF Grants CCF-1748585 and CNS-1748692. The work of S. El Rouayheb was supported in part by NSF Grant CCF-1817635.

notions of external and internal mother vertex-sets are novel generalizations of the notion of mother vertex of a graph.

II. Problem Formulation

Throughout, we denote random variables and their realizations by bold-face letters and regular letters, respectively.

For a prime q, let $\mathbb{F}_q$ be a finite field of size q, and let $\mathbb{F}_{q^l}$ be an extension field of $\mathbb{F}_q$ for some integer $l \geq 1$. Let $L \triangleq l \log_2 q$, and let $\mathbb{F}_q^\times \triangleq \mathbb{F}_q \setminus \{0\}$. For a positive integer i, let $[i] \triangleq \{1, \ldots, i\}$. Let $K \geq 1$, $M \geq 1$, and $D \geq 1$ be arbitrary integers such that $D + M \leq K$.

Suppose there is a server storing a set of K messages, denoted by $X \triangleq \{X_i\}_{i\in[K]}$, with each message X_i being independently and uniformly distributed over $\mathbb{F}_{q^l}$. That is, $H(\mathbf{X}_i) = L$ for $i \in [K]$ and $H(\mathbf{X}) = KL$, where $\mathbf{X} \triangleq \{\mathbf{X}_i\}_{i\in[K]}$. Also, suppose there are D users, each of which demands one distinct message X_i. Let W be the index set of the demanded messages. The users send the indices of their demanded messages to a trusted agent, called *aggregator*, who knows M messages $X_S \triangleq \{X_i\}_{i\in S}$ for some $S \subset [K]$, $|S| = M$, $S \cap W = \emptyset$. Then, the aggregator retrieves the D messages $X_W \triangleq \{X_i\}_{i\in W}$ from the server. We refer to W as the *demand index set*, X_W as the *demand*, D as the *demand size*, S as the *side information index set*, X_S as the *side information*, and M as the *side information size*.

Let $\mathcal{W}$ and $\mathcal{S}$ be the set of all D-subsets and all M-subsets of $[K]$, respectively. We assume that $\mathbf{S}$ is distributed uniformly, i.e., $\Pr(\mathbf{S} = S) = \binom{K}{M}^{-1}$ for all $S \in \mathcal{S}$; and $\mathbf{W}$, conditional on $\mathbf{S} = S$, is uniformly distributed, i.e., $\Pr(\mathbf{W} = W|\mathbf{S} = S) = \binom{K-M}{D}^{-1}$ for all $W \in \mathcal{W}$ such that $W \cap S = \emptyset$. Note that $\Pr(i \in \mathbf{W}) = D/K$ for all $i \in [K]$.

We assume that the server *a priori* knows the demand size D, the side information size M, the distribution of $\mathbf{S}$ and the conditional distribution of $\mathbf{W}$ given $\mathbf{S}$. In contrast, the realizations S and W are unknown to the server *a priori*.

For any S, W, for retrieving X_W the aggregator sends to the server a query $Q^{[W,S]}$. The aggregator's query is a (potentially stochastic) function of W, S, and X_S. The query $Q^{[W,S]}$ is required to protect the privacy of the demand index of every user individually from the server, i.e.,

$$\Pr(i \in \mathbf{W}|\mathbf{Q}^{[\mathbf{W},\mathbf{S}]} = Q^{[W,S]}, \mathbf{X} = X) = \Pr(i \in \mathbf{W})$$

for all $i \in [K]$. This condition is referred to as the *individual-privacy condition*. It should be noted that the individual-privacy condition is weaker than the *joint-privacy condition*, a.k.a. the *W-privacy condition*, being studied in [14], where the privacy of all indices in the demand index set must be protected jointly. The notions of individual privacy and joint privacy coincide for $D = 1$, which was previously settled in [12], and hence, in this work, we focus on $D \geq 2$.

Upon receiving $Q^{[W,S]}$, the server sends to the aggregator an answer $A^{[W,S]}$. The server's answer is a deterministic function of the query $Q^{[W,S]}$ and the messages in X. In other words, $(\mathbf{W}, \mathbf{S}) \to (\mathbf{Q}^{[\mathbf{W},\mathbf{S}]}, \mathbf{X}) \to \mathbf{A}^{[\mathbf{W},\mathbf{S}]}$ forms a Markov chain, and $H(\mathbf{A}^{[\mathbf{W},\mathbf{S}]}|\mathbf{Q}^{[\mathbf{W},\mathbf{S}]}, \mathbf{X}) = 0$. In addition, the answer $A^{[W,S]}$ along with the side information X_S and the index sets W, S must enable the aggregator to retrieve the demand X_W, i.e.,

$$H(\mathbf{X}_\mathbf{W}|\mathbf{A}^{[\mathbf{W},\mathbf{S}]}, \mathbf{Q}^{[\mathbf{W},\mathbf{S}]}, \mathbf{X}_\mathbf{S}, \mathbf{W}, \mathbf{S}) = 0.$$

This condition is referred to as the *recoverability condition*.

The problem is to design a protocol that, for any given W, S, generates a query $Q^{[W,S]}$ and its corresponding answer $A^{[W,S]}$ (given $Q^{[W,S]}$ and X) which satisfy both the privacy and recoverability conditions. We refer to this problem as *single-server multi-message Individually-Private Information Retrieval with Side Information (IPIR-SI)*.

The *rate* of an IPIR-SI protocol is defined as the ratio of the entropy of the demand messages, i.e., DL, to the total entropy of the answer, i.e., $H(\mathbf{A}^{[\mathbf{W},\mathbf{S}]})$. The supremum of rates over all IPIR-SI protocols is defined as the *capacity* of the IPIR-SI problem. In this work, our goal is to characterize the capacity of the IPIR-SI problem, and to design an IPIR-SI protocol that achieves the capacity.

III. Main Results

In this section, we present our main results. Theorem 1 provides a lower bound on the capacity of IPIR-SI problem for $M \geq 1$ and $D \geq 2$, and Theorem 2 characterizes the capacity of IPIR-SI problem for the special case of $M = 1$ and $D = 2$. The proofs of Theorems 1 and 2 are given in Sections IV and V, respectively.

Theorem 1. *The capacity of IPIR-SI problem with K messages, side information size $M \geq 1$, and demand size $D \geq 2$ is lower bounded by $D(K - M\lfloor \frac{K}{M+D} \rfloor)^{-1}$ if $\frac{K-D}{M+D} \leq \lfloor \frac{K}{M+D} \rfloor$, and by $\lceil \frac{K}{M+D} \rceil^{-1}$ otherwise.*

The proof is based on constructing an IPIR-SI protocol that achieves the rate $D(K - M\lfloor K/(M+D) \rfloor)^{-1}$ or $\lceil K/(M+D) \rceil^{-1}$, depending on K, M, and D (see, for details, Section IV). This protocol, which is a variation of the Generalized Partition and Code (GPC) protocol previously proposed in [14] for single-server multi-message PIR-SI where joint-privacy is required, is referred to as *GPC for Individual Privacy*, or *GPC-IP* for short.

Remark 1. A lower bound on the capacity of single-server multi-message PIR with side information, when the privacy of the demand indices must be protected jointly, was previously presented in [14, Theorem 1]. Surprisingly, this lower bound reduces to the lower bound of Theorem 1 where M (in [14, Theorem 1]) is replaced by MD. This correspondence implies that each message in the side information, when achieving individual-privacy, can be as effective as D side information messages when joint-privacy is required. This also suggests that, as one would expect, relaxing the privacy condition (from joint to individual) can increase the capacity.

Theorem 2. *The capacity lower bound given in Theorem 1 is tight for $M = 1$ and $D = 2$.*

The proof of converse is based on new combinatorial and information-theoretic arguments, relying on two necessary conditions imposed by the individual-privacy and recoverability conditions (see Lemmas 2 and 3).

Remark 2. As we will show later, the tightness of the result of Theorem 1 for arbitrary M and D, which remains open in general, is subject to the correctness of a novel conjecture in combinatorics, formally stated in Section V, which may be of independent interest. Interestingly, for $M = 1$ and $D \geq 2$, our conjecture relates the size of an external mother vertex-set of any directed graph G, whose nodes have in-degree at least one and out-degree either zero or at least D, to the size of an internal mother vertex-set of the transpose of G. (The notions of external and internal mother vertex-sets, formally defined in Section V, are two generalizations of the notion of the mother vertex in graph theory.) In this work, we prove the simplest non-trivial case of this conjecture for $M = 1$ and $D = 2$, and leave the complete proof for the future work.

IV. Proof of Theorem 1

In this section, we propose an IPIR-SI protocol, referred to as *Generalized Partition and Code for Individual Privacy (GPC-IP)*, achieving the rate lower bound of Theorem 1.

For simplifying the notation, we define $\alpha \triangleq M + D$, $\beta \triangleq \lfloor K/\alpha \rfloor$, $\rho \triangleq K - \alpha\beta$, $\gamma \triangleq \min\{\rho, D\}$. (Note that $0 \leq \rho < \alpha$.) We also define

$$\theta_1 \triangleq \frac{\binom{\alpha-1}{M}}{\prod_{j=1}^{\beta-1}\binom{K-j\alpha}{\alpha}}, \quad \theta_2 \triangleq \frac{\binom{\alpha-1}{M+\rho}\binom{M+\rho}{M}(\frac{\alpha\beta}{D-\rho}-1)}{\binom{D}{\rho}\binom{K-\alpha}{\rho}\prod_{j=1}^{\beta-1}\binom{K-j\alpha-\rho}{\alpha}},$$

and

$$\theta_3 \triangleq \frac{\beta\binom{\rho}{D}\binom{K-\rho}{\alpha-\rho}}{\binom{M}{\rho-D}\prod_{j=0}^{\beta-1}\binom{K-j\alpha-\rho}{\alpha}}.$$

We assume that $q \geq \alpha$, and let $\omega_1, \ldots, \omega_\alpha$ be α distinct elements from $\mathbb{F}_q$. (For $q < \alpha$, the achievability of the rate lower bound of Theorem 1 remains open.)

GPC-IP Protocol: This protocol consists of four steps as follows:

Step 1: First, the aggregator constructs a set Q_0 of size ρ from the indices in $[K]$, and β disjoint sets $Q_1, \ldots, Q_\beta$ (also disjoint from Q_0), each of size α, from the indices in $[K]$, where the construction procedure is described below. There are two cases based on ρ: (i) $\rho < D$, and (ii) $\rho \geq D$.

Case (i): With probability $\frac{\theta_1}{\theta_1+\theta_2}$, the aggregator places ρ randomly chosen elements (demand indices) from W into Q_0 and the remaining elements of W along with all elements of S (side information indices) into Q_1. Then the aggregator randomly places all other elements of $[K]$ into $Q_2, \ldots, Q_\beta$ and the remaining positions in Q_1; otherwise, with probability $\frac{\theta_2}{\theta_1+\theta_2}$, the aggregator places all elements of $S \cup W$ into Q_1, and randomly places all other elements of $[K]$ into $Q_0, Q_2, \ldots, Q_\beta$.

Case (ii): With probability $\frac{\theta_1}{\theta_1+\theta_3}$, the aggregator places all elements of W along with $\rho - D$ randomly chosen elements from S into Q_0, and places the remaining elements of S together with all other elements of $[K]$ into $Q_1, \ldots, Q_\beta$ at random; otherwise, with probability $\frac{\theta_3}{\theta_1+\theta_3}$, the aggregator places all elements of $S \cup W$ into Q_1, and randomly places all other elements of $[K]$ into $Q_0, Q_2, \ldots, Q_\beta$.

Next, the aggregator creates a collection Q' of γ sequences $Q'_1, \ldots, Q'_\gamma$, each of length ρ, such that $Q'_j = \{\omega_1^{j-1}, \ldots, \omega_\rho^{j-1}\}$ for $j \in [\gamma]$, and a collection Q'' of D sequences $Q''_1, \ldots, Q''_D$, each of length α, such that $Q''_j = \{\omega_1^{j-1}, \ldots, \omega_\alpha^{j-1}\}$ for $j \in [D]$.

Step 2: The aggregator constructs $Q_0^* = (Q_0, Q')$ and $Q_j^* = (Q_j, Q'')$ for $j \in [\beta]$, and sends to the server the query $Q^{[W,S]} = \{Q_0^*, Q_{\sigma^{-1}(1)}^*, \ldots, Q_{\sigma^{-1}(\beta)}^*\}$ for a randomly chosen permutation $\sigma : [\beta] \to [\beta]$.

Step 3: By using $Q_0^* = (Q_0, Q')$ and $Q_j^* = (Q_j, Q'')$ for $j \in [\beta]$, the server computes $A_0 = \{A_0^1, \ldots, A_0^\gamma\}$, where $A_0^k = \sum_{l=1}^{\rho} \omega_l^{k-1} X_{i_l}$ for $k \in [\gamma]$ where $Q_0 = \{i_1, \ldots, i_\rho\}$, and computes $A_j = \{A_j^1, \ldots, A_j^D\}$ for $j \in [\beta]$, where $A_j^k = \sum_{l=1}^{\alpha} \omega_l^{k-1} X_{i_l}$ for $j \in [D]$ where $Q_j = \{i_1, \ldots, i_\alpha\}$. The server then sends to the aggregator the answer $A^{[W,S]} = \{A_0, A_{\sigma^{-1}(1)} \ldots, A_{\sigma^{-1}(\beta)}\}$.

Step 4: Upon receiving the answer from the server, the aggregator retrieves X_i for $i \in W \cap Q_0$ (or $i \in W \cap Q_j$ for some $j \in [\beta]$) by subtracting off the contribution of the side information messages X_S from the γ (or D) equations associated with A_0 (or A_j), and solving the resulting system of γ (or D) linear equations with γ (or D) unknowns.

Lemma 1. *The GPC-IP protocol is an IPIR-SI protocol, and achieves the rate $D(K - M\lfloor \frac{K}{M+D} \rfloor)^{-1}$ if $\frac{K-D}{M+D} \leq \lfloor \frac{K}{M+D} \rfloor$, and the rate $\lceil \frac{K}{M+D} \rceil^{-1}$ otherwise.*

Proof: If $\frac{K-D}{M+D} \leq \lfloor \frac{K}{M+D} \rfloor$, then $\rho < D$. Thus, $\gamma = \rho$. In this case, $H(\mathbf{A}_0) = \rho L$ and $H(\mathbf{A}_j) = DL$ for $j \in [\beta]$, where $L = H(\mathbf{X}_i)$ for all $i \in [K]$. Thus, for any $W \in \mathcal{W}, S \in \mathcal{S}$ such that $S \cap W = \emptyset$, we have $H(\mathbf{A}^{[W,S]}) = H(\mathbf{A}_0, \ldots, \mathbf{A}_\gamma) = \sum_{j=0}^{\beta} H(\mathbf{A}_j) = (\rho + \beta D)L$. By the uniformity of the joint distribution of $\mathbf{W}$ and $\mathbf{S}$, the rate in this case is then equal to $DL/H(\mathbf{A}^{[\mathbf{W},\mathbf{S}]}) = DL/H(\mathbf{A}^{[W,S]}) = D/(\rho + \beta D) = D(K - M\lfloor \frac{K}{M+D} \rfloor)^{-1}$. If $\frac{K-D}{M+D} > \lfloor \frac{K}{M+D} \rfloor$, then $\rho \geq D$. Thus, $\gamma = D$. In this case, $H(\mathbf{A}_0) = H(\mathbf{A}_j) = DL$ for all $j \in [\beta]$, and thus, $H(A^{[W,S]}) = (\beta + 1)DL$. Then, the rate is equal to $D/(\beta + 1)D = \lceil \frac{K}{M+D} \rceil^{-1}$.

The recoverability condition is obviously satisfied. For the proof of individual privacy, we need to show that for any $Q \triangleq \{Q_0, \ldots, Q_\beta\}$ constructed by the protocol, $\Pr(i \in \mathbf{W} | \mathbf{Q} = Q)$ is the same for all $i \in [K]$. Here we only give the proof for the case of $\rho < D$, and the proof for the case of $\rho \geq D$ can be found in [18].

For simplifying the notation, let $\mathbf{E}$ denote the event $\{\mathbf{Q} = Q\}$. For the case of $\rho < D$, $\Pr(i \in \mathbf{W} | \mathbf{E})$ is given by

$$\sum_{j=1}^{\beta} \sum_{\substack{W \subset Q_j: \\ |W| = D-\rho}} \sum_{\substack{S \subset Q_j \setminus W: \\ |S| = M}} \Pr(\mathbf{W} = Q_0 \cup W, \mathbf{S} = S | \mathbf{E}) \quad (1)$$

for all $i \in Q_0$, and

$$\sum_{\substack{W \subset Q_j: \\ |W| = D, i \in W}} \Pr(\mathbf{W} = W, \mathbf{S} = Q_j \setminus W | \mathbf{E}) \quad (2)$$

$$+ \sum_{\substack{W \subset Q_j: \\ |W| = D-\rho, i \in W}} \sum_{\substack{S \subset Q_j \setminus W: \\ |S| = M}} \Pr(\mathbf{W} = Q_0 \cup W, \mathbf{S} = S | \mathbf{E})$$

for all $i \in Q_j$, $j \in [\beta]$. From (1) and (2), one can see that $\Pr(i \in \mathbf{W} | \mathbf{E})$ is the same for all $i \in Q_0$, say equal to p_0,

and is the same for all $i \in Q_j$ and all $j \in [\beta]$, say equal to p_1. We need to show that p_0 and p_1 are equal. Note that p_0 and p_1 are equal if the following two summations are equal:

$$\sum_{j=1}^{\beta} \sum_{\substack{W\subset Q_j:\\ |W|=D-\rho}} \sum_{\substack{S\subset Q_j\setminus W:\\ |S|=M}} \Pr(\mathbf{E}|\mathbf{W}=Q_0\cup W,\mathbf{S}=S) \quad (3)$$

$$\sum_{\substack{W\subset Q_j:\\ |W|=D,i\in W}} \Pr(\mathbf{E}|\mathbf{W}=W,\mathbf{S}=Q_j\setminus W) \quad (4)$$

$$+ \sum_{\substack{W\subset Q_j:\\ |W|=D-\rho,i\in W}} \sum_{\substack{S\subset Q_j\setminus W:\\ |S|=M}} \Pr(\mathbf{E}|\mathbf{W}=Q_0\cup W,\mathbf{S}=S)$$

Fix $j \in [\beta]$. For any $W \subset Q_j$, $|W|= D-\rho$, and any $S \subset Q_j \setminus W$, $|S|= M$, a simple counting yields

$$\Pr(\mathbf{E}|\mathbf{W}=Q_0\cup W,\mathbf{S}=S) = \left(\frac{\theta_1}{\theta_1+\theta_2}\right) \times (\beta-1)!\left(\binom{D}{\rho}\binom{K-\alpha}{\rho}\prod_{j=1}^{\beta-1}\binom{K-j\alpha-\rho}{\alpha}\right)^{-1}.$$

and accordingly, (3) is equal to

$$\left(\frac{\theta_1}{\theta_1+\theta_2}\right)\binom{\alpha}{M+\rho}\binom{M+\rho}{M} \times \beta!\left(\binom{D}{\rho}\binom{K-\alpha}{\rho}\prod_{j=1}^{\beta-1}\binom{K-j\alpha-\rho}{\alpha}\right)^{-1}.$$

For any $W \subset Q_j$, $|W|= D$ such that $i \in W$, we have

$$\Pr(\mathbf{E}|\mathbf{W}=W,\mathbf{S}=Q_j\setminus W) = \left(\frac{\theta_2}{\theta_1+\theta_2}\right) \times (\beta-1)!\left(\prod_{j=1}^{\beta-1}\binom{K-j\alpha}{\alpha}\right)^{-1},$$

and for any $W \subset Q_j$, $|W|= D-\rho$ such that $i \in W$, and any $S \subset Q_j \setminus W$, $|S|= M$, we have

$$\Pr(\mathbf{E}|\mathbf{W}=Q_0\cup W,\mathbf{S}=S) = \left(\frac{\theta_1}{\theta_1+\theta_2}\right) \times (\beta-1)!\left(\binom{D}{\rho}\binom{K-\alpha}{\rho}\prod_{j=1}^{\beta-1}\binom{K-j\alpha-\rho}{\alpha}\right)^{-1}.$$

Accordingly, (4) is equal to

$$\binom{\alpha-1}{M}(\beta-1)!\left(\left(\frac{\theta_2}{\theta_1+\theta_2}\right)\left(\prod_{j=1}^{\beta-1}\binom{K-j\alpha}{\alpha}\right)^{-1} + \left(\frac{\theta_1}{\theta_1+\theta_2}\right)\left(\frac{D-\rho}{D}\right) \times \left(\binom{K-\alpha}{\rho}\prod_{j=1}^{\beta-1}\binom{K-j\alpha-\rho}{\alpha}\right)^{-1}\right).$$

It is easy to verify that (3) and (4) are equal for the choice of θ_1 and θ_2 defined as in the protocol. □

V. Proof of Theorem 2

In this section, we first present a new combinatorial conjecture which, if holds, proves the tightness of the result of Theorem 1. Next, we prove the simplest non-trivial case of this conjecture, yielding the tightness of the capacity lower bound in Theorem 1 for $M = 1$ and $D = 2$.

Before stating the conjecture, we give two necessary conditions, due to individual-privacy and recoverability, which are essential to relate the IPIR-SI problem to our conjecture.

Lemma 2. *For any $j \in [K]$, there must exist $W \in \mathcal{W}, j \in W, S \in \mathcal{S}$ where $S \cap W = \emptyset$, such that*

$$H(\mathbf{X}_W|\mathbf{A}^{[\mathbf{W},\mathbf{S}]},\mathbf{Q}^{[\mathbf{W},\mathbf{S}]},\mathbf{X}_S) = 0.$$

Proof: The proof is by the way of contradiction, and is omitted for brevity. □

Lemma 3. *For any $W \in \mathcal{W}, S \in \mathcal{S}, W \cap S = \emptyset, J \subseteq [K]$, if $\Pr(\cup_{j\in J}\{j \in \mathbf{W}\}|\mathbf{Q}^{[\mathbf{W},\mathbf{S}]} = Q^{[W,S]}) = 1$, then $|J|\geq K/D$.*

Proof: Let $\mathbf{E}$ denote the event $\{\mathbf{Q}^{[\mathbf{W},\mathbf{S}]} = Q^{[W,S]}\}$. Take an arbitrary $J \subseteq [K]$ such that $\Pr(\cup_{j\in J}\{j \in \mathbf{W}\}|\mathbf{E}) = 1$. By the individual-privacy condition, $\Pr(j \in \mathbf{W}|\mathbf{E}) = D/K$ for all $j \in [K]$. By the union bound, $\Pr(\cup_{j\in J}\{j \in \mathbf{W}\}|\mathbf{E})$ is bounded from above by $\sum_{j\in J}\Pr(j \in \mathbf{W}|\mathbf{E}) = |J|D/K$. Thus, $|J|D/K \geq 1$, or equivalently, $|J|\geq K/D$. □

We would like to show that $H(\mathbf{A}^{[\mathbf{W},\mathbf{S}]})$, or particularly $H(\mathbf{A}^{[\mathbf{W},\mathbf{S}]}|\mathbf{Q}^{[\mathbf{W},\mathbf{S}]})$, is bounded from below by $\min\{K - M\lfloor\frac{K}{M+D}\rfloor, D\lceil\frac{K}{M+D}\rceil\}$, for any protocol that generates a query-answer pair $(Q^{[W,S]}, A^{[W,S]})$ (for any W, S) satisfying the conditions in Lemmas 2 and 3. Any such protocol can be represented by a mapping as follows.

Let K, M, D be arbitrary positive integers such that $K \geq D + M$. Let $\mathcal{I}$ and $\mathcal{J}$ be the set of all subsets I and J of $[K]$ such that $0 \leq |I|\leq M$ and $|J|\geq D$, respectively. Let $f : \mathcal{I} \to \mathcal{J}$ be an arbitrary mapping. A relation f is called *good* if the following conditions hold: (i) $I \subseteq f(I)$ for any $I \in \mathcal{I}$; (ii) for any $I_1, I_2 \in \mathcal{I}$, if $I_2 \subseteq f(I_1)$, then $f(I_2) \subseteq f(I_1)$; (iii) for any $j \in [K]$, there exist $I \in \mathcal{I}$ and $J \in \mathcal{J}$, $|J|= D$, $j \in J$ where $I \cap J = \emptyset$ such that $J \subseteq f(I)$; and (iv) for any $J \subseteq [K]$, $|J|< \lceil K/D\rceil$, there exists $I \in \mathcal{I}$, $I \neq \emptyset$ such that $f(I) \cap J = \emptyset$.

Thinking of the M-subsets in $\mathcal{I}$ as the potential side information index sets S, and the D-subsets in $\mathcal{J}$ as the possible demand index sets W, one can observe that a good mapping f, satisfying the conditions (i)-(iv), represents an arbitrary protocol that satisfies the conditions in Lemmas 2 and 3. Then, it holds that for any IPIR-SI protocol, $H(\mathbf{A}^{[\mathbf{W},\mathbf{S}]}|\mathbf{Q}^{[\mathbf{W},\mathbf{S}]}) \geq K - \theta$ (for any integer $\theta \geq 0$) so long as for any good mapping f (defined earlier) there exists a subset $I^* \subseteq [K], |I^*|\leq \theta$ such that the union of $f(I)$ for all $I \subseteq I^*$ is equal to $[K]$. This is because, thinking of f (or in turn, the protocol) as an oracle, given the messages $\{X_j\}_{j\in I^*}$, all other messages $\{X_j\}_{j\in[K]\setminus I^*}$ are recoverable from $A^{[W,S]}$ and $Q^{[W,S]}$ (for any W, S); and hence, $H(\mathbf{A}^{[\mathbf{W},\mathbf{S}]}|\mathbf{Q}^{[\mathbf{W},\mathbf{S}]}) \geq K - |I^*|\geq K - \theta$, as desired.

Conjecture 1. *For any good mapping f, there exists $I^* \subset [K]$, $|I^*| \leq \max\{K - D\lceil \frac{K}{M+D} \rceil, M\lfloor \frac{K}{M+D} \rfloor\}$ such that $\cup_{I \subseteq I^*} f(I) = [K]$.*

For $M = 1$ and $D \geq 2$, the statement of Conjecture 1 can be rephrased in the language of graph theory as follows. Let $G = (V, E)$ be an arbitrary directed graph (without parallel edges), where V and E are the set of nodes and edges of G, respectively. Denote by $d_{\text{in}}(v)$ and $d_{\text{out}}(v)$ the in-degree and out-degree of node $v \in V$, respectively, over G. We define an *external* (or respectively, *internal*) *mother vertex-set* of G as a minimal subset I^* of nodes in V from which all other nodes u in $V \setminus I^*$ such that $d_{\text{out}}(u) \neq 0$ (or respectively, $d_{\text{in}}(u) \neq 0$) can be reached (i.e., for any $u \in V \setminus I^*$, $d_{\text{out}}(u) \neq 0$ (or respectively, $d_{\text{in}}(u) \neq 0$), there exists $v \in I^*$ such that there is a directed path from v to u in G), and denote the size of an external (or respectively, internal) mother vertex-set I^* of G by $\mu_{\text{ext}}(G)$ (or respectively, $\mu_{\text{int}}(G)$). Also, let G^{T} be the transpose of G, which is formed by reversing the direction of all edges in G (i.e., $G^{\mathsf{T}} = (V, E^{\mathsf{T}})$, where $E^{\mathsf{T}} = \{(u, v) : (v, u) \in E\}$). We call G a *D-graph* if the following conditions hold: (i) for any $v \in V$, $d_{\text{in}}(v) \geq 1$, and $d_{\text{out}}(v) = 0$ or $d_{\text{out}}(v) \geq D$; and (ii) $\mu_{\text{int}}(G^{\mathsf{T}}) \geq \lceil \frac{K}{D} \rceil$.

Conjecture 2. *For any D-graph G on K nodes, $\mu_{\text{ext}}(G) \leq \lfloor \frac{K}{D+1} \rfloor$.*

Note that the upper bound on $|I^*|$ in Conjecture 1 reduces to $\lfloor \frac{K}{D+1} \rfloor$ for $M = 1$. This is because $K - D\lceil \frac{K}{D+1} \rceil \leq \lfloor \frac{K}{D+1} \rfloor$ for any $D \leq K - 1$. Moreover, for any D-graph $G = (V, E)$ on K nodes, we can define $f(v)$ for any $v \in V$ as the set of all nodes (including v) that can be reached from node v (via a directed path in G). Then, it is easy to verify that f satisfies the conditions (i)-(iv) for a good mapping. Note also that $\mu_{\text{ext}}(G)$ represents the size of a (minimal) subset $I^* \subseteq V$ such that $\cup_{v \in I^*} f(v) = V$. This shows the equivalence between Conjectures 1 and 2 for $M = 1$.

In the following, we prove Conjecture 2 for $M = 1$ and $D = 2$, and hence the proof of Theorem 2.

Lemma 4. *For any 2-graph G on K nodes, $\mu_{\text{ext}}(G) \leq \lfloor \frac{K}{3} \rfloor$.*

Proof: Let G be an arbitrary 2-graph on K nodes. Suppose that $\mu_{\text{ext}}(G) > \lfloor \frac{K}{3} \rfloor$. We need to show a contradiction. Let $n \triangleq \mu_{\text{ext}}(G)$. Consider an arbitrary partition of the nodes in G into n parts, $V_1, \ldots, V_n$, such that each part V_j contains a node v_j from which all other nodes in V_j can be reached. (Note that a node in a part can potentially reach some other nodes in other parts.) Obviously, $I^* \triangleq \{v_1, \ldots, v_n\}$ is an external mother vertex-set of G.

By the minimality of I^*, it follows that no node v_j can be reached from any node out of the part V_j. (Otherwise, from the nodes in $I^* \setminus \{v_j\}$ all other nodes can be reached, and this contradicts the minimality of I^*.) Since $d_{\text{in}}(v_j) \geq 1$ (by definition), then there must exist another node u_j in V_j that reaches v_j. Also, no part V_j can contain only a single node v_j, simply because $d_{\text{in}}(v_j) \geq 1$, and the node v_j can be reached from some other node(s) in some other part(s), which again contradicts the minimality of I^*.

Take an arbitrary part $V_j = \{v_j, u_j\}$ of size 2 (if exists). Since v_j reaches u_j (over G), then $d_{\text{out}}(v_j) \geq 1$, and particularly, $d_{\text{out}}(v_j) \geq 2$. Thus, the node v_j reaches some other node(s), say w, in some other part(s) over G. Equivalently, the node w reaches both nodes v_j and u_j over G^{T}. For any other part V_j of size $i \geq 3$, the nodes v_j and u_j can be reached from each node in $V_j \setminus \{v_j, u_j\}$ over G^{T}.

By these arguments, each node in $\{v_j, u_j\}_{j \in [n]}$ can be reached from some node(s) in $J^* \triangleq V \setminus \{v_j, u_j\}_{j \in [n]}$ via a directed path in G^{T}. Then, $\mu_{\text{int}}(G^{\mathsf{T}}) \leq |J^*| = K - 2n$. By assumption, $\mu_{\text{ext}}(G) = n > \lfloor \frac{K}{3} \rfloor$. Thus, $|J^*| < K - 2\lfloor \frac{K}{3} \rfloor$, and $\mu_{\text{int}}(G^{\mathsf{T}}) < K - 2\lfloor \frac{K}{3} \rfloor$. Since $K - 2\lfloor \frac{K}{3} \rfloor \leq \lceil \frac{K}{2} \rceil$, then $\mu_{\text{int}}(G^{\mathsf{T}}) < \lceil \frac{K}{2} \rceil$. This is a contradiction because $\mu_{\text{int}}(G^{\mathsf{T}}) \geq \lceil \frac{K}{2} \rceil$ for any 2-graph G on K nodes. □

REFERENCES

[1] S. Yekhanin, "Private information retrieval," *Communications of the ACM*, vol. 53, no. 4, pp. 68–73, 2010.

[2] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. on Info. Theory*, vol. 63, no. 7, pp. 4075–4088, July 2017.

[3] ——, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. on Info. Theory*, vol. 64, no. 4, pp. 2361–2370, April 2018.

[4] R. Tajeddine and S. E. Rouayheb, "Robust private information retrieval on coded data," in *IEEE Int. Sympo. on Info. Theory (ISIT'17)*, June 2017, pp. 1903–1907.

[5] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. on Info. Theory*, vol. 64, no. 3, pp. 1945–1956, March 2018.

[6] ——, "Multi-message private information retrieval: Capacity results and near-optimal schemes," Feb 2017. [Online]. Available: arXiv:1702.01739

[7] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," May 2018. [Online]. Available: arXiv:1805.11892

[8] 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.

[9] 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.

[10] ——, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Trans. on Info. Theory*, pp. 1–1, 2018.

[11] Z. Chen, Z. Wang, and S. Jafar, "The capacity of private information retrieval with private side information," Sept 2017. [Online]. Available: arXiv:1709.03022

[12] S. Kadhe, B. Garcia, A. Heidarzadeh, S. E. Rouayheb, and A. Sprintson, "Private information retrieval with side information: The single server case," in *2017 55th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2017, pp. 1099–1106.

[13] ——, "Private information retrieval with side information," Sept 2017. [Online]. Available: arXiv:1709.00112

[14] A. Heidarzadeh, B. Garcia, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," in *2018 56th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2018.

[15] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," in *2018 56th Annual Allerton Conf. on Commun., Control, and Computing*, Oct 2018.

[16] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "Capacity of single-server single-message private information retrieval with coded side information," June 2018. [Online]. Available: arXiv:1806.00661

[17] S. Li and M. Gastpar, "Single-server multi-user private information retrieval with side information," in *IEEE Int. Sympo. on Info. Theory (ISIT'18)*, June 2018, pp. 1954–1958.

[18] A. Heidarzadeh, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "Single-server multi-message individually-private information retrieval with side information," Feb 2019. [Online]. Available: arXiv:1901.07509