The Capacity of Private Information Retrieval Under Arbitrary Collusion Patterns

Xinyu Yao∗, Nan Liu†, Wei Kang∗

∗School of Information Science and Engineering, Southeast University, Nanjing, China, {xyyao,wkang}@seu.edu.cn
†National Mobile Communications Research Laboratory, Southeast University, China, nanliu@seu.edu.cn

Abstract—We study the private information retrieval (PIR) problem under arbitrary collusion patterns for replicated databases. We find its capacity, which is the same as the capacity of the original PIR problem with the number of databases $N$ replaced by a number $S^*$. The number $S^*$ is the optimal solution to a linear programming problem that is a function of the collusion pattern. Hence, the collusion pattern affects the capacity of the PIR problem only through the number $S^*$.

I. INTRODUCTION

The problem of private information retrieval (PIR) was first proposed in [1], where the user wants to retrieve a certain bit out of $K$ bits from $N$ replicated databases without revealing which bit is of interest to any single database. The design objective in [1] is to minimize the upload cost and the download cost between the user and the databases. The PIR problem was reformulated in [2] from an information-theoretic perspective, where the user wants to retrieve a sufficiently large message from the databases so that the download cost is minimized. This problem was fully solved by Sun and Jafar [2], where the capacity of the PIR problem was shown to be

$$C_{\text{PIR}} = \left( 1 + \frac{T}{N} + \frac{T^2}{N^2} + \cdots + \frac{T^{K-1}}{N^{K-1}} \right)^{-1},$$

which is defined as the ratio of the size of the desired message to the total number of downloaded symbols from the databases. The capacity increases with the number of databases $N$, since with the help of more databases, the privacy of the user can be hidden better from any single database. Many interesting extensions and variations for the PIR problem have since then been studied in over a hundred papers, which due to limited space, we can not list here.

One of the first variations studied was that of the colluding databases [3], where some subsets of databases may communicate and collude to learn about the message index that is of interest to the user. To preserve privacy under possible collusion among databases, the number of downloaded symbols needs to be increased. The first study on database collusion focused on the case where we have replicated databases, i.e., each database stores a replica of the entirety of the $K$ files, and $T$-colluding databases, where it is assumed that up to $T$ number of databases may collude. Sun and Jafar [3] proved that the capacity of the $T$-colluding PIR problem for replicated databases, is

$$C_{\text{PIR}} = \left( 1 + \frac{T}{N} + \frac{T^2}{N} + \cdots + \frac{T^{K-1}}{N^{K-1}} \right)^{-1}. \quad (2)$$

Comparing (2) with (1), we see that when any $T$ databases may collude, the number of effective databases has decreased from $N$ to $N^T$, where $N^T$ does not need to be an integer.

Following [3], many extensions of $T$-colluding PIR have been studied [4]–[20], among which MDS-coded databases with $T$-colluding generated a lot of research interest [4]–[10]. The MDS-coded databases scenario is the case where the messages are encoded using an $[N, J]$ MDS code, and the coded bits are stored in the $N$ databases. Unlike the replicated-databases scenario, where each database has the ability to reconstruct all $K$ messages, here, any $J$ databases together can reconstruct the $K$ messages. Thus, the replicated-databases scenario is a special case of the MDS-coded-databases scenario when $J = 1$. Finding the capacity of the $T$-colluding PIR problem with MDS-coded databases is difficult, and remains open in general [6], [8].

While most works focused on the $T$-colluding structure of the databases, where any up to $T$ databases may collude, it is of interest to study more general collusion patterns due to the possible heterogeneity of the databases. An arbitrary collusion pattern may be represented by its maximal colluding sets [4], [10] as $P = \{T_1, T_2, \ldots, T_M\}$, where the databases in set $T_m$, $m \in [1 : M]$ may collude, and there are $M$ such colluding sets.

Tajeddine et. al [4] proposed the PIR problem under arbitrary collusion patterns and studied it for MDS-coded databases. Several other works followed, including [13] for replicated databases, [10, Section VII] for MDS-coded databases, and some discussions in [6, Appendix D], for both the replicated and MDS-coded databases scenarios.

In this paper, we focus on the PIR problem under arbitrary collusion patterns for the replicated databases scenario. The known results for this problem thus far is 1) the capacity for the special case of disjoint colluding sets [13]; 2) the capacity for the special case of cyclically contiguous databases [6, Appendix D]; 3) a rate of (2) is achievable for $T \triangleq \max_{T \in P} |T|$, i.e., we may consider the more strict collusion pattern where any up to the maximum number of colluding
databases in $P$ may collude. This is also the result we obtain when specializing [10] to the replicated databases scenario; 4) a rate indicated by Theorem 2 in [4], specialized to the replicated databases scenario by setting $k = 1$, is achievable. As can be seen, the understanding of the PIR problem under arbitrary collusion patterns for replicated databases is still rather limited.

In this paper, we find the PIR capacity under arbitrary collusion patterns for the replicated databases scenario. Though collusion patterns are diverse, and at first glance, the problem requires a case-by-case analysis due to the property of each specific collusion pattern [10], we provide a general formula for the PIR capacity that holds true for any collusion pattern $P$. The capacity formula is shown to be

$$ C_P = \left( 1 + \frac{1}{S^*} + \left( \frac{1}{S^*} \right)^2 + \cdots + \left( \frac{1}{S^*} \right)^{K-1} \right)^{-1}, $$

where $S^*$ is the optimal value of the following linear programming problem

$$ \max \quad 1^T y \\
\text{subject to} \quad B_P^T y \leq 1_M \\
y \geq 0_N, $$

where $B_P$ is the incidence matrix, of size $N \times M$, of the collusion pattern $P$, i.e., if $DB$ is in the $m$-th colluding set $T_m$ in $P$, we let the $(n,m)$-th element of $B_P$ be 1, otherwise, it is zero. $1_k$ ($0_k$) is the column vector of size $k$ whose elements are all one (zero). Comparing (3) with (1) and (2), we find that the number of effective databases under arbitrary collusion pattern $P$ is $S^*$ which is related to the collusion pattern $P$ through a linear programming solution.

The difficulty of finding the capacity of the PIR problem under arbitrary collusion patterns for replicated databases comes from finding a common proof and capacity expressions that work for any collusion pattern. Towards this end, the tools and ideas that we use in proving the capacity result include 1) using the sub-modular property of the entropy function [21] to prove a general inequality, which is used in place of Han’s inequality for $T$-colluding [3], in the induction argument of the converse; 2) linking the achievable PIR rate and its converse to the optimal solution of two linear programming problems; 3) using the duality of linear programming problems to show that the achievability and converse results meet, yielding the capacity.

II. SYSTEM MODEL

Consider the problem where $K$ messages are stored on $N$ replicated databases. The $K$ messages, denoted as $W_1, \ldots, W_K$, are independent and each message consists of $L$ symbols, which are independently and uniformly distributed over a finite field $F_q$, where $q$ is the size of the field, i.e.,

$$ H(W_k) = L, \quad k = 1, \ldots, K, $$

$$ H(W_1, \ldots, W_K) = H(W_1) + H(W_2) + \cdots + H(W_K). $$

A user wants to retrieve message $W_\theta$, $\theta \in [1 : K]$, by sending designed queries to the databases, where the query sent to the $n$-th database is denoted as $Q_n[\theta]$. Since the queries are designed by the user, who do not know the content of the messages, we have

$$ I(W_{1:K}; Q_{1:N}^\theta) = 0, \quad \forall \theta \in [1 : K]. $$

Upon receiving the query $Q_n[\theta]$, Database $n$ calculates the answer, denoted as $A_n[\theta]$, based on the query received $Q_n[\theta]$ and the messages $W_{1:K}$, i.e.,

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

The queries need to be designed such that the user is able to reconstruct the desired message $W_\theta$ from all the answers received from the databases, i.e.,

$$ H(W_\theta | A_{1:N}^\theta, Q_{1:N}^\theta) = 0, \quad \forall \theta \in [1 : K]. $$

The queries also need to be designed such that the privacy of the user is preserved. In this paper, we consider colluding databases, and furthermore, the collusion pattern can be arbitrary. We represent the collusion pattern as $P = \{T_1, T_2, \ldots, T_M\}$, where $M$ is the number of colluding sets and $T_m \subseteq [1 : N], \forall m \in [1 : M]$ is the $m$-th colluding set in $P$. The representation $P$ means that the databases in set $T_m$ may collude, and there are $M$ such colluding sets. As an example, for $N = 4$ databases, the 2-colluding case considered by [3] is denoted as $P = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}\}$, and the disjoint collusion pattern considered in [13] would cover cases such as $P = \{\{1\}, \{2, 3\}, \{3, 4\}, \{2, 4\}\}, P = \{\{1, 2\}, \{3, 4\}\}$. Note that the defined collusion pattern $P$ satisfy the following two constraints: 1) we only include the maximal collusion set as elements of $P$. For example, if $\{1, 2, 3\} \in P$, then by definition, $\{1, 2\}$ is a colluding set too. But we do not include $\{1, 2\}$ in $P$ for ease of representation; 2) all databases must appear in at least one element of $P$, because at the very least, the privacy of the user must be preserved at each single database, which is the requirement of the original PIR problem [2].

To protect the privacy of the user, we require that databases that are in a colluding set can not learn anything about the desired message index $\theta$, i.e.,

$$ (Q_T^{[1]}, A_T^{[1]}, W_{1:K}) \sim (Q_T^{[\theta]}, A_T^{[\theta]}, W_{1:K}), \quad \forall \theta \in [1 : K], \forall T \in P. $$

The rate of the PIR problem with collusion pattern $P$, denoted as $R_P$, is defined as the ratio between the message size $L$ and the total number of downloaded information from the databases, i.e.,

$$ R_P = \frac{L}{\sum_{n=1}^N H(A_n^{[\theta]}|Q_n[\theta])}, $$

which is not a function of $\theta$ due to the privacy constraint in (8). The capacity of the PIR problem with collusion pattern $P$ is $C_P = \sup R_P$, where the supremum is over all possible
retrieval schemes.

We define an incidence matrix $B_P$, of size $N \times M$, to describe the collusion pattern $P$, where if DB $n$ is in the $m$-th colluding set in $P$, we let the $(n,m)$-th element of $B_P$ be 1, otherwise, it is zero. For example, $P = \{\{1,2\}, \{2,3\}, \{2,4\}, \{1,3,4\}\}$ would correspond to an incidence matrix of

$$B_P = \begin{bmatrix}
1 & 0 & 0 & 1 \\
1 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1
\end{bmatrix}.$$  

Throughout the paper, we will denote the $k \times 1$ column vector of all ones as $\mathbf{1}_k$, and the $k \times 1$ column vector of all zeros as $\mathbf{0}_k$. $\mathbf{I}_k$ is the size $k \times k$ identity matrix and when the size is evident, we write it as $\mathbf{I}$. Similarly, $\mathbf{0}_{k \times i}$ is the size $k \times i$ matrix of all zeros, and when the size is evident, we write it as $\mathbf{0}$.

III. MAIN RESULTS

The main result of the paper is the PIR capacity under arbitrary collusion patterns for replicated databases, as shown in the next theorem.

**Theorem 1** The capacity of the PIR problem under collusion pattern $P$ for replicated databases is

$$C_P = \left(1 + \frac{1}{S^*} + \left(\frac{1}{S^*}\right)^2 + \cdots + \left(\frac{1}{S^*}\right)^{K-1}\right)^{-1},$$  

where $S^*$ is the optimal value of the following linear programming problem, which we will call (LP1),

$$\begin{align*}
\text{(LP1)} & \quad \max_y \quad \mathbf{1}_N^T y \\
& \quad \text{subject to} \quad B_P^T y \leq \mathbf{1}_M \\
& \quad \quad y \geq \mathbf{0}_N,
\end{align*}$$

where $B_P$ is the incidence matrix, of size $N \times M$, of the collusion pattern $P$.

Theorem 1 will be proved in the following section. We will first show that (10) is achievable when the amount of data queried to each database is proportional to the optimal solution $y^*$ of (LP1). Next, we present a converse theorem where the upper bound on capacity has the same form as (10) with $S^*$ replaced by $S_2$, and $S_2$ is the optimal value of another linear programming problem (LP2). Finally, we show that (LP1) and (LP2) are dual problems, which means $S^* = S_2$. This concludes the proof that (10) is the capacity of the PIR problem under arbitrary collusion patterns for replicated databases.

We make a few remarks here regarding the main result.

**Remark 1** Theorem 1 shows that the arbitrary collusion pattern $P$ affects the capacity of the PIR problem only through the linear programming problem (LP1). More specifically, the capacity formula under arbitrary collusion patterns take on the same form as that of the original PIR problem of (1), with $N$ replaced by the optimal solution of (LP1).

**Remark 2** Our results coincide with known capacity results of PIR colluding for replicated databases:

1) In the case of non-colluding databases [2], the collusion pattern is $P = \{\{1\}, \{2\}, \ldots, \{N\}\}$, whose incidence matrix is $B_P = \mathbf{I}_N$. It is straightforward to see that the optimal solution to (LP1) is $y^* = \mathbf{1}_N$, and the corresponding optimal value $S^* = N$. Hence, the capacity formula in (10) becomes (1), consistent with [2].

2) In the case of $T$-colluding databases [3], the collusion pattern $P$ consists of all size $T$ subsets of $[1:N]$, and there are a total of \(\binom{N}{T}\) many colluding sets, i.e., $M = \binom{N}{T}$. The corresponding incidence matrix of size $N \times M$ consists of $\binom{N}{T}$ columns, each with $T$ number of 1s and $N - T$ number of 0s. It is straightforward to see that the optimal solution to (LP1) is $y^* = \frac{1}{T} \mathbf{1}_N$, and the corresponding optimal value $S^* = \frac{N}{T}$. Hence, the capacity formula in (10) becomes (2), consistent with [3].

3) In the case $T$-colluding cyclically contiguous databases [6, Appendix D], the collusion pattern $P = \{\{1,2,\ldots,T\}, \{2,3,\ldots,T+1\}, \ldots, \{N,1,2,\ldots,T-1\}\}$, where $M = N$. The transpose of the corresponding incidence matrix, i.e., $B_P^T$, is a circulant matrix, where the first row consists of $T$ number of 1s followed by $N - T$ number of 0s. It is straightforward to see that though the incidence matrix is different from that of the $T$-colluding case, the optimal solution $y^*$, and hence the optimal value $S^*$, is the same. Thus, the capacity formula in (10) becomes (2), consistent with [6, Appendix D].

4) In the case of disjoint colluding sets [13], the $N$ servers are split into $J$ disjoint sets, where Set $j$ consists of $N_j$ databases, $j \in [1:J]$. Within Set $j$, up to $T_j$ databases may collude, where $T_j \leq N_j$. The corresponding incidence matrix to this collusion pattern is

$$B_P = \begin{bmatrix}
B_1 & 0 & \cdots & 0 \\
0 & B_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & B_J
\end{bmatrix},$$

where $B_j$ is an $N_j \times \binom{N_j}{T_j}$ matrix, with each column consisting of $T_j$ 1s and $N_j - T_j$ 0s, $j \in [1:J]$. It is straightforward to see that the optimal solution to (LP1) is $y^* = \left[\begin{smallmatrix}1 \\
\frac{T_1}{T_j} \\
\frac{T_2}{T_j} \\
\vdots \\
\frac{T_J}{T_j}\end{smallmatrix}\right]^T$. The corresponding optimal value $S^* = \sum_{j=1}^J \frac{N_j}{T_j}$. Hence, the capacity formula in (10) becomes $1 + \left(\sum_{j=1}^J \frac{N_j}{T_j}\right)^{-1} + \left(\sum_{j=1}^J \frac{N_j}{T_j}\right)^{-2} + \cdots + \left(\sum_{j=1}^J \frac{N_j}{T_j}\right)^{-(K-1)}$, consistent with [13, Theorem 2].
IV. PROOFS

A. Achievability

Recall that for each collusion pattern $\mathcal{P}$, there is a corresponding incidence matrix $B_{\mathcal{P}}$, as defined in Section II. Let $y = [y_1, y_2, \ldots, y_N]^T$ be a feasible and rational solution of (LP1), i.e., $y$ consists of rational elements, and it satisfies the constraints (11) and (12). Let the value of the objective function in (LP1) corresponding to $y$ be $S$, i.e., $S = \sum_{n=1}^{N} y_n$. Then, we have the following achievability theorem.

**Theorem 2** Consider the PIR problem with collusion pattern $\mathcal{P}$, whose incidence matrix is $B_{\mathcal{P}}$. Suppose $y$ is a rational and feasible solution of (LP1) and $S = 1/y$. Then the following rate is achievable, i.e.,

$$C_{\mathcal{P}} \geq \left(1 + \frac{1}{S} + \left(\frac{1}{S}\right)^2 + \cdots + \left(\frac{1}{S}\right)^{K-1}\right)^{-1}.$$  

(13)

**Proof:** The details of the proof of Theorem 2, along with an illustrative example, can be found in the full version of this paper. The proof is similar to [3, Section IV.D], and we note the differences here: 1) In place of $N^K$ in [3, Section IV.D], we have $L$, which is the message length. $L$ will be chosen such that the number of $k$-sum symbols downloaded from each of the databases is an integer, $k \in [1 : K]$. Such an $L$ can be found since $y$ is rational. 2) In place of $\mathcal{S}$ in [3, Section IV.D], we have $S$. 3) Rather than distributing the queries evenly among all databases, we distribute the queries proportionally according to $(y, S)$, more specifically, the number of queries to Database $n$ is based on the proportion $\frac{y_n}{S}$, $n \in [1 : N]$.

**Remark 3** The main novelty in our achievable scheme is, rather than distributing the queries evenly among all databases, we propose distributing the queries proportionally according to $(y, S)$, i.e., the number of queries to Database $n$ is based on the proportion $\frac{y_n}{S}$, $n \in [1 : N]$). First of all, this is possible because $y$ satisfies the constraint in (12), which means $y_n \geq 0$, $n \in [1 : N]$. Secondly, $y$ that satisfies the constraint (11) will guarantee the user’s privacy. This can be intuitively explained as follows: the databases in each colluding $\mathcal{T}_m \in \mathcal{P}$ can not see too many symbols being queried, i.e., the $m$-th element of $B_{\mathcal{P}}^T y$ is no greater than 1, otherwise, the dependency of the undesired symbol will be revealed to the colluding databases in $\mathcal{T}_m$, violating the privacy of the user.

**Remark 4** It is easy to see that $y = \frac{1}{S} 1_N$ is a feasible and rational solution. The corresponding $S = 1/y = 1$. This is the suboptimal retrieval scheme of downloading all $K$ messages, evenly from all the databases.

Note that the right-hand side of (13) is an increasing function of $S$. Based on the result of Theorem 2, to find the largest possible achievable rate, we should find the maximum $S = \sum_{n=1}^{N} y_n$, achievable over all $y$ satisfying (11) and (12). Applying Theorem 2 for the optimal solution of (LP1), i.e., $(y^*, S^*)$, and noting that $y^*$ is rational due to the fact that the objective function and the linear constraints in (LP1) are both with integer coefficients, the rate of Theorem 1 is achievable.

B. Converse

Recall that for each collusion pattern $\mathcal{P}$, there is a corresponding incidence matrix $B_{\mathcal{P}}$, as defined in Section II. Consider the following linear programming problem, which will be called (LP2),

$$\begin{align*}
\min_{x} & \quad x^T M x \\
\text{subject to} & \quad B_{\mathcal{P}} x \geq 1_N \quad \text{(14)} \\
& \quad x \geq 0_M. \quad \text{(15)}
\end{align*}$$

Let $x = [x_1, x_2, \ldots, x_M]^T$ be a feasible and rational solution of (LP2), i.e., $x$ consists of rational elements, and it satisfies the constraints (14) and (15). Let the value of the objective function in (LP2) corresponding to $x$ be $S_2$, i.e., $S_2 = \sum_{m=1}^{M} x_m$. We have the following converse theorem.

**Theorem 3** Consider the PIR problem with collusion pattern $\mathcal{P}$, whose incidence matrix is $B_{\mathcal{P}}$. Suppose $x$ is a rational and feasible solution of (LP2) and $S_2 = 1_M^T x$. Then, the capacity of the PIR problem is upper bounded by

$$C_{\mathcal{P}} \leq \left(1 + \frac{1}{S_2} + \left(\frac{1}{S_2}\right)^2 + \cdots + \left(\frac{1}{S_2}\right)^{K-1}\right)^{-1}.$$  

(16)

**Proof:** The details of the proof can be found in the full version of this paper. We comment on the main idea here. Using standard PIR converse techniques such as those in [2], we can obtain for $k = 2, 3, \ldots, K$,

$$H(A^{[k-1]}_{1:N} | W_{1:k-1}, Q^{[k-1]}_{1:N}) \geq H(A^{[k]}_{\mathcal{T}_m} | W_{1:k-1}, Q^{[k]}_{1:N}),$$

$m = 1, 2, \ldots, M$.  

(17)

For each $m \in [1 : M]$, multiply both sides of (17) by $x_m$, which is the $m$-th element of $x$. Note that $x$ satisfies (15), which means that we are multiplying non-negative numbers and the sign of the inequality does not need to be changed. Then, adding all these $M$ inequalities together, we obtain

$$S_2 H(A^{[k-1]}_{1:N} | W_{1:k-1}, Q^{[k-1]}_{1:N}) \geq \sum_{m=1}^{M} x_m H(A^{[k]}_{\mathcal{T}_m} | W_{1:k-1}, Q^{[k]}_{1:N}),$$

(18)

where we have used the definition of $S_2$, i.e., $S_2 = \sum_{m=1}^{M} x_m$. The fact that $x$ is rational and non-negative means that there exist non-negative integers $G_x^{1}, G_x^{2}, \ldots, G_x^{M}, G_x$, such that each $x_m$ can be expressed as $x_m = \frac{G_x}{G_x^m}$, $m \in [1 : M]$. Thus, we have

$$G_x \sum_{m=1}^{M} x_m H(A^{[k]}_{\mathcal{T}_m} | W_{1:k-1}, Q^{[k]}_{1:N}) = \sum_{m=1}^{M} G_x H(A^{[k]}_{\mathcal{T}_m} | W_{1:k-1}, Q^{[k]}_{1:N}).$$

(19)
Since $G_m^m, m \in [1 : M]$ are integers, the right-hand side of (19) can be written as a summation of the form
\[ \sum_{v=1}^{V} H(A_{\bar{r}_v}^{[k]} | W_{1,k-1}, Q_{1:N}^{[k]}), \] (20)
where $V$ is a positive integer, and $\bar{r}_v \subseteq [1 : N], v \in [1 : V]$.

We have the following results for a summation of the form (20): we say that the summation in (20) satisfies the even property with the number $G$, if the number of times $n$ appears in $A \triangleq \{ \bar{r}_1, \bar{r}_2, \ldots, \bar{r}_V \}$ is equal to $G$ for each $n \in [1 : N]$. For a summation that satisfies the even property, we have
\[ \sum_{v=1}^{V} G_A^{[k]} H(A_{\bar{r}_v}^{[k]} | W_{1,k-1}, Q_{1:N}^{[k]}), \]
which follows by applying the sub-modular property of the entropy function multiple times.

Note that $x$ satisfies constraint (14). In the case of $B_P \cdot x = 1_N$, the sum on the right-hand side of (19) satisfies the even property with $G = G_x$. In the case of $B_P \cdot x > 1_N$, after writing the right-hand side of (19) in the form of (20), we may delete some indices of $n$ in sets $\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_V$, until each $n$ appears only $G_x$ number of times. This gives us a lower bound to the right-hand side of (19), and this lower bound is a summation that satisfies the even property with the number $G_x$. Hence, for all cases of $B_P \cdot x \geq 1_N$, we have
\[ \sum_{m=1}^{M} G_x^{[k]} H(A_{\bar{r}_v}^{[k]} | W_{1,k-1}, Q_{1:N}^{[k]}), \]
which follows by applying the sub-modular property of the entropy function multiple times.

Note that $x$ satisfies constraint (14). In the case of $B_P \cdot x = 1_N$, the sum on the right-hand side of (19) satisfies the even property with $G = G_x$. In the case of $B_P \cdot x > 1_N$, after writing the right-hand side of (19) in the form of (20), we may delete some indices of $n$ in sets $\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_V$, until each $n$ appears only $G_x$ number of times. This gives us a lower bound to the right-hand side of (19), and this lower bound is a summation that satisfies the even property with the number $G_x$. Hence, for all cases of $B_P \cdot x \geq 1_N$, we have
\[ \sum_{m=1}^{M} G_x^{[k]} H(A_{\bar{r}_v}^{[k]} | W_{1,k-1}, Q_{1:N}^{[k]}), \]
which follows by applying the sub-modular property of the entropy function multiple times.

Utilizing (18), (19) and (22), we may obtain the induction argument
\[ S_2 H(A_{1:N}^{[k-1]} | W_{1,k-1}, Q_{1:N}^{[k-1]}), \]
for $k \in [2 : K]$, from which the result of Theorem 3 follows from standard PIR converse techniques such as those in [2].

\begin{remark}
The main novelty of our converse proof is proving (22), which is a general version of Han’s inequality. We use (22) in place of Han’s inequality, which was used for $T$-colluding [3], for the induction argument of the converse. We show that when $x$ satisfies constraints (14) and (15), the sum corresponding to $x$ either satisfies the even property or a lower bound of it satisfies the even property, resulting in (22).
\end{remark}

The reason why $x$ has to satisfy (14), (15) and is rational is stated in the proof main idea above. Note that the right-hand side of (16) is an increasing function of $S_2$. Based on the result of Theorem 3, to find the tightest possible upper bound, we should find the minimum $S_2 = \sum_{m=1}^{M} x_m a_m$ achievable over all $x$ satisfying (14) and (15). Applying Theorem 3 for the optimal solution of (LP2), i.e., $(x^*, S_2^*)$, and noting that $x^*$ is rational, the right-hand side of (19) can be written as a summation of the form
\[ \sum_{v=1}^{V} H(A_{\bar{r}_v}^{[k]} | W_{1,k-1}, Q_{1:N}^{[k]}), \]
where $V$ is a positive integer, and $\bar{r}_v \subseteq [1 : N], v \in [1 : V]$.

We have the following results for a summation of the form (20): we say that the summation in (20) satisfies the even property with the number $G$, if the number of times $n$ appears in $A \triangleq \{ \bar{r}_1, \bar{r}_2, \ldots, \bar{r}_V \}$ is equal to $G$ for each $n \in [1 : N]$. For a summation that satisfies the even property, we have
\[ \sum_{v=1}^{V} G_A^{[k]} H(A_{\bar{r}_v}^{[k]} | W_{1,k-1}, Q_{1:N}^{[k]}), \]
which follows by applying the sub-modular property of the entropy function multiple times.

Note that $x$ satisfies constraint (14). In the case of $B_P \cdot x = 1_N$, the sum on the right-hand side of (19) satisfies the even property with $G = G_x$. In the case of $B_P \cdot x > 1_N$, after writing the right-hand side of (19) in the form of (20), we may delete some indices of $n$ in sets $\bar{r}_1, \bar{r}_2, \ldots, \bar{r}_V$, until each $n$ appears only $G_x$ number of times. This gives us a lower bound to the right-hand side of (19), and this lower bound is a summation that satisfies the even property with the number $G_x$. Hence, for all cases of $B_P \cdot x \geq 1_N$, we have
\[ \sum_{m=1}^{M} G_x^{[k]} H(A_{\bar{r}_v}^{[k]} | W_{1,k-1}, Q_{1:N}^{[k]}), \]
which follows by applying the sub-modular property of the entropy function multiple times.

Utilizing (18), (19) and (22), we may obtain the induction argument
\[ S_2 H(A_{1:N}^{[k-1]} | W_{1,k-1}, Q_{1:N}^{[k-1]}), \]
for $k \in [2 : K]$, from which the result of Theorem 3 follows from standard PIR converse techniques such as those in [2].

\section{Conclusions}

We have found the capacity of the PIR problem under arbitrary collusion patterns for replicated databases. We first link the achievable PIR rate and its converse to the solutions of two linear programming problems (LP1) and (LP2), i.e., we have
\[ \left( 1 + \frac{1}{S^*} + \left( \frac{1}{S^2} \right)^2 + \cdots + \left( \frac{1}{S^2} \right)^{K-1} \right)^{-1} \leq C_P \]
where $S^*$ and $S_2^*$ are the optimal solutions to (LP1) and (LP2), respectively. It is easy to see that (LP1) and (LP2) are actually dual problems of each other, which means $S^* = S_2^*$. Hence, we have found the capacity of the PIR problem under arbitrary collusion pattern $P$ for replicated databases, as described in Theorem 1.

To aid in a better understanding of the PIR problem under arbitrary collusion patterns for replicated databases and its proofs, we provide several examples in the full version of the paper [22].

\begin{thebibliography}{99}
\item [1] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In Proceedings of IEEE 36th Annual Foundations of Computer Science, pages 41–50, Oct. 1995.
\item [2] Hua Sun and Syed Ali Jafar. The capacity of private information retrieval. IEEE Transactions on Information Theory, 63(7):4075–4088, Jul. 2017.
\item [3] Hua Sun and Syed Ali Jafar. The capacity of robust private information retrieval with colluding databases. IEEE Transactions on Information Theory, 64(4):2361–2370, Apr. 2018.
\item [4] Razane Tajeddine, Oliver W Gnilke, David Karpuk, Ragnar Freij-Hollanti, Camilla Hollanti, and Salim El Rouayheb. Private information retrieval schemes for coded data with arbitrary collusion patterns. In 2017 IEEE International Symposium on Information Theory (ISIT), pages 1908–1912, Jun. 2017.
\end{thebibliography}
[5] Razane Tajeddine, Oliver W Gnilke, David Karpuk, Ragnar Freij-Hollanti, and Camilla Hollanti. Robust private information retrieval from coded systems with byzantine and colluding servers. In 2018 IEEE International Symposium on Information Theory (ISIT), pages 2451–2455, Jun. 2018.

[6] Hua Sun and Syed Ali Jafar. Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al. IEEE Transactions on Information Theory, 64(2):1000–1022, Feb. 2018.

[7] Razane Tajeddine, Oliver W. Gnilke, David Karpuk, Ragnar Freij-Hollanti, and Camilla Hollanti. Private information retrieval from coded storage systems with colluding, byzantine, and unresponsive servers. IEEE Transactions on Information Theory, 65(6):3898–3906, Jun. 2019.

[8] Lukas Holzbaur, Ragnar Freij-Hollanti, Jie Li, and Camilla Hollanti. Capacity of linear private information retrieval from coded, colluding, and adversarial servers. arXiv preprint https://arxiv.org/pdf/1903.12552.pdf, Mar. 2019.

[9] Yiwei Zhang and Gennian Ge. A general private information retrieval scheme for MDS coded databases with colluding servers. Designs, Codes and Cryptography, May. 2019.

[10] Yiwei Zhang and Gennian Ge. Private information retrieval from MDS coded databases with colluding servers under several variant models. arXiv preprint arXiv:1705.03186, Oct. 2017.

[11] Qiwen Wang and Mikael Skoglund. Secure symmetric private information retrieval from colluding databases with adversaries. In 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 1083–1090, Oct. 2017.

[12] Qiwen Wang and Mikael Skoglund. Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers. In 2017 IEEE Information Theory Workshop (ITW), pages 71–75, Nov. 2017.

[13] Zhuqing Jia, Hua Sun, and Syed A. Jafar. The capacity of private information retrieval with disjoint colluding sets. In 2017 IEEE Global Communications Conference, Dec. 2017.

[14] Qiwen Wang and Mikael Skoglund. Secure private information retrieval from colluding databases with eavesdroppers. In 2018 IEEE International Symposium on Information Theory (ISIT), pages 2456–2460, Jun. 2018.

[15] Karim Banawan and Sennur Ulukus. The capacity of private information retrieval from byzantine and colluding databases. IEEE Transactions on Information Theory, 65(2):1206–1219, Feb. 2019.

[16] Qiwen Wang and Mikael Skoglund. Symmetric private information retrieval from MDS coded distributed storage with non-colluding and colluding servers. IEEE Transactions on Information Theory, 65(8):5160–5175, Aug. 2019.

[17] Qiwen Wang and Mikael Skoglund. On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers. IEEE Transactions on Information Theory, 65(5):3183–3197, May 2019.

[18] Zhifang Zhang and Jingke Xu. The optimal sub-packetization of linear capacity-achieving PIR schemes with colluding servers. IEEE Transactions on Information Theory, 65(5):2723–2735, May 2019.

[19] Umberto Martinez-Penas. Private information retrieval from locally repairable databases with colluding servers. arXiv preprint https://arxiv.org/pdf/1901.02938.pdf, Jan. 2019.

[20] Ragnar Freij-Hollanti, Oliver W. Gnilke, Camilla Hollanti, Anna-Lena Horlemann-Trautmann, David Karpuk, and Ivo Kubjas. T-private information retrieval schemes using transitive codes. IEEE Transactions on Information Theory, 65(4):2107–2118, Apr. 2019.

[21] Alexander Schrijver. Combinatorial optimization: polyhedra and efficiency, volume 24. Springer Science & Business Media, 2003.

[22] Xinyu Yao, Nan Liu, and Wei Kang. The capacity of private information retrieval schemes using transitive codes. IEEE Transactions on Information Theory, 65(4):2107–2118, Apr. 2019.

[23] Qiwen Wang, Hua Sun, and Mikael Skoglund. The capacity of private information retrieval with eavesdroppers. IEEE Transactions on Information Theory, 65(1):322–329, Jan. 2019.

[24] Xinyu Yao, Nan Liu, and Wei Kang. Capacity of symmetric private information retrieval under arbitrary collision patterns. In preparation, Jan. 2020.

[25] Qiwen Wang, Hua Sun, and Mikael Skoglund. The capacity of private information retrieval under arbitrary collision patterns. In preparation, Jan. 2020.

[26] Yang Li, Nan Liu, and Wei Kang. Capacity of symmetric private information retrieval under arbitrary collision patterns. In preparation, Jan. 2020.