# Multi-Message Private Information Retrieval: Capacity Results and Near-Optimal Schemes 

Karim Banawan, Student Member, IEEE, and Sennur Ulukus ${ }^{\oplus}$, Fellow, IEEE


#### Abstract

We consider the problem of multi-message private information retrieval (MPIR) from $N$ non-communicating replicated databases. In MPIR, the user is interested in retrieving $P$ messages out of $M$ stored messages without leaking the identity of the retrieved messages. The information-theoretic sum capacity of MPIR $C_{s}^{P}$ is the maximum number of desired message symbols that can be retrieved privately per downloaded symbol, where the symbols are defined over the same field. For the case $P \geq M / 2$, we determine the exact sum capacity of MPIR as $C_{s}^{P^{-}}=1 /(1+(M-P) /(P N))$. The achievable scheme in this case is based on downloading MDS-coded mixtures of all messages. For $P \leq M / 2$, we develop lower and upper bounds for all $M, P, N$. These bounds match if the total number of messages $M$ is an integer multiple of the number of desired messages $P$, i.e., $M / P \in \mathbb{N}$. In this case, $C_{s}^{P}=\left(1+1 / N+\cdots+1 / N^{M / P-1}\right)^{-1}$, i.e., $C_{s}^{P}=(1-1 / N) /\left(1-1 / N^{M / P}\right)$ for $N>1$, and $C_{s}^{P}=P / M$ for $N=1$. The achievable scheme in this case generalizes the singlemessage capacity achieving scheme to have unbalanced number of stages per round of download. For all the remaining cases, the difference between the lower and upper bound is at most 0.0082 , which occurs for $M=5, P=2, N=2$. Our results indicate that joint retrieval of desired messages is more efficient than successive use of single-message retrieval schemes even after considering the free savings that result from downloading undesired symbols in each single-message retrieval round.


Index Terms-Private information retrieval, multi-message, MDS code, capacity, alignment, IIR filters.

## I. Introduction

THE privacy of the contents of the downloaded information from curious public databases has attracted considerable research within the computer science community [1]-[4]. The problem is motivated by practical examples such as: ensuring privacy of investors as they download records in a stock market, since revealing the interest in a certain record may influence its value; ensuring the privacy of an inventor as they look up existing patents in a database, since revealing what they are looking at leaks some information about the current invention they are

[^0]working on; and protecting activists in oppressive regimes as they browse restricted content on the internet [2]. In the classical private information retrieval (PIR) problem, a user wishes to download a certain message (or file) from $N$ noncommunicating databases without leaking any information about the identity of the downloaded message. The contents of the databases are identical. The user performs this operation by preparing and submitting queries to all databases. The databases respond truthfully with answer strings which are functions of the queries and the messages. The user needs to reconstruct the desired message from these answer strings. A trivial solution for this seemingly difficult problem is for the user to download the contents of any individual database. This solution however is extremely inefficient. The efficiency is measured by the retrieval rate which is the ratio of the number of retrieved desired message symbols to the number of total downloaded symbols, where the symbols are defined over the same field. The capacity of PIR is the maximum retrieval rate over all possible PIR schemes.
The computer science formulation of this problem assumes that the messages are of length one bit. The metrics in this case are the download cost, i.e., the sum of lengths of the answer strings, and the upload cost, i.e., the size of the queries. Most of this work is computational PIR as it ensures only that a server cannot get any information about user intent unless it solves a certain computationally hard problem [2], [5]. The information-theoretic re-formulation of the problem considers arbitrarily large message sizes, and ignores the upload cost. This formulation provides an absolute, i.e., informationtheoretic, guarantee that no server participating in the protocol gets any information about the user intent. Towards that end, recently, [6] has drawn a connection between the PIR problem and the blind interference alignment scheme proposed in [7]. Then, [8] has determined the exact capacity of the classical PIR problem. The retrieval scheme in [8] is based on three principles: message symmetry, symmetry across databases, and exploiting side information from the undesired messages through alignment.

The basic PIR setting has been extended in several interesting directions. The first extension is the coded PIR (CPIR) problem [9]-[11]. The contents of the databases in this problem are coded by an $(N, K)$ storage code instead of being replicated. This is a natural extension since most storage systems nowadays are in fact coded to achieve reliability against node failures and erasures with manageable storage cost. In [12], the exact capacity of the MDS-coded PIR is
determined. Another interesting extension is PIR with colluding databases (TPIR). In this setting, $T$ databases can communicate and exchange the queries to identify the desired message. The exact capacity of colluded PIR is determined in [13]. The case of coded PIR with colluding databases is investigated in [14]. The robust PIR problem (RPIR) extension considers the case when some databases are not responsive [13]. Lastly, in the symmetric PIR problem (SPIR) the privacy of the remaining records should be maintained against the user in addition to the usual privacy constraint on the databases, i.e., the user should not learn any other messages other than the one it wished to retrieve. The exact capacity of symmetric PIR is determined in [15]; and the exact capacity of symmetric PIR from coded databases is determined in [16].

In some applications, the user may be interested in retrieving multiple messages from the databases without revealing the identities of these messages. Returning to the examples presented earlier: the investor may be interested in comparing the values of multiple records at the same time, and the inventor may be looking up several patents that are closely related to their work. One possible solution to this problem is to use single-message retrieval scheme in [8] successively. We show in this work that multiple messages can be retrieved more efficiently than retrieving them one-by-one in a sequence. This resembles superiority of joint decoding in multiple access channels over multiple simultaneous singleuser transmissions [17]. To motivate the multi-message private information retrieval problem (MPIR), consider the example in [8, Sec. 4.3] where the number of messages is $M=3$, number of databases is $N=2$, and the user is interested in retrieving only $P=1$ message. Here the optimal scheme retrieves 8 desired bits in 14 downloads, hence with a rate $4 / 7$. When the user wishes to retrieve $P=2$ messages, if we use the scheme in [8] twice in a row, we retrieve 16 bits in 28 downloads, hence again a sum rate of $4 / 7$. Even considering the fact that the scheme in [8] retrieves 2 bits of the second message for free in downloading the first message, i.e., it actually retrieves 10 bits in 14 downloads, hence a sum rate of $5 / 7$, we show in this paper that a better sum rate of $4 / 5$ can be achieved by joint retrieval of the messages.

Although there is a vast literature on classical PIR in the computer science literature, only a few works exist in MPIR: Reference [18] proposes a multi-block (multi-message) scheme and observes that if the user requests multiple blocks (messages), it is possible to reuse randomly mixed data blocks (answer strings) across multiple requests (queries). Reference [19] develops a multi-block scheme which further reduces the communication overhead. An achievable scheme for the multi-block PIR by designing $k$-safe binary matrices that uses XOR operations is developed in [20]. Reference [20] extends the scheme in [1] to multiple blocks. Reference [21] designs an efficient non-trivial multiquery computational PIR protocol and gives a lower bound on the communication of any multi-query information retrieval protocol. Reference [22] suggests using batch codes to allow a single client to retrieve multiple records simultaneously while allowing the server computation to scale sublinearly with the number of records fetched. This idea is extended further in [23]
to design a PIR server algorithm that achieves sublinear scaling in the number of records fetched, even when they are requested by distinct, non-collaborating clients. These works do not consider determining the information-theoretic capacity.

In this paper, we formulate the MPIR problem with noncolluding repeated databases from an information-theoretic perspective. Our goal is to characterize the sum capacity of the MPIR problem $C_{s}^{P}$, which is defined as the maximum ratio of the number of retrieved symbols from the $P$ desired messages to the number of total downloaded symbols. When the number of desired messages $P$ is at least half of the total number of messages $M$, i.e., $P \geq \frac{M}{2}$, we determine the exact sum capacity of MPIR as $C_{s}^{P}=\frac{1}{1+\frac{M-P}{P N}}$. We use a novel achievable scheme which downloads MDS-coded mixtures of all messages. We show that joint retrieving of the desired messages strictly outperforms successive use of single-message retrieval for $P$ times. Additionally, we present an achievable rate region to characterize the trade-off between the retrieval rates of the desired $P$ messages.

For the case of $P \leq \frac{M}{2}$, we derive lower and upper bounds that match if the total number of messages $M$ is an integer multiple of the number of desired messages $P$, i.e., $\frac{M}{P} \in \mathbb{N}$. In this case, the sum capacity is $C_{s}^{P}=\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{M / P}}$. The result resembles the single-message capacity with the number of messages equal to $\frac{M}{P}$. In other cases, although the exact capacity is still an open problem, we show numerically that the gap between the lower and upper bounds is monotonically decreasing in $N$ and is upper bounded by 0.0082 . The achievable scheme when $P \leq \frac{M}{2}$ is inspired by the greedy algorithm in [8], which retrieves all possible combinations of messages. The main difference of our scheme from the scheme in [8] is the number of stages required in each download round. For example, round $M-P+1$ to round $M-1$, which correspond to retrieving the sum of $M-P+1$ to sum of $M-1$ messages, respectively, are suppressed in our scheme. This is because, they do not generate any useful side information for our purposes here, in contrast to [8]. Interestingly, the number of stages for each round is related to the output of a $P$-order IIR filter [24]. This intriguing connection to IIR filtering is a result of constructing the greedy achievable scheme in [8] backwards and observing the required side information needed in previous rounds. Our converse proof generalizes the proof in [8] for $P \geq 1$. The essence of the proof is captured in two lemmas: the first lemma lower bounds the uncertainty of the interference for the case $P \geq \frac{M}{2}$, and the second lemma upper bounds the remaining uncertainty after conditioning on $P$ interfering messages.

## II. Problem Formulation

Consider a classical PIR setting storing $M$ messages (or files). Each message is a vector $W_{i} \in \mathbb{F}_{q}^{L}, i \in\{1, \cdots, M\}$, whose elements are picked uniformly and independently from sufficiently large field ${ }^{1} \mathbb{F}_{q}$. Denote the contents of

[^1]message $W_{m}$ by the vector $\left[w_{m}(1), w_{m}(2), \cdots, w_{m}(L)\right]^{T}$. The messages are independent and identically distributed, and thus,
\[

$$
\begin{align*}
H\left(W_{i}\right) & =L, \quad i \in\{1, \cdots, M\}  \tag{1}\\
H\left(W_{1: M}\right) & =M L \tag{2}
\end{align*}
$$
\]

where $L$ is measured in $q$-ary bits, $W_{1: M}=\left(W_{1}, W_{2}, \cdots\right.$, $W_{M}$ ). The messages are stored in $N$ non-colluding (noncommunicating) databases. Each database stores an identical copy of all $M$ messages, i.e., the databases encode the messages via $(N, 1)$ repetition storage code [12].

In the MPIR problem, the user aims to retrieve a subset of messages indexed by the index set $\mathcal{P}=\left\{i_{1}, \cdots, i_{P}\right\} \subseteq$ $\{1, \cdots, M\}$ out of the available messages, where $|\mathcal{P}|=P$, without leaking the identity of the subset $\mathcal{P}$. We assume that the cardinality of the potential message set, $P$, is known to all databases. To retrieve $W_{\mathcal{P}}=\left(W_{i_{1}}, W_{i_{2}}, \cdots, W_{i_{P}}\right)$, the user generates a query $Q_{n}^{[\mathcal{P}]}$ and sends it to the $n$th database. The user does not have any knowledge about the messages in advance, hence the messages and the queries are statistically independent,

$$
\begin{equation*}
I\left(W_{1}, \cdots, W_{M} ; Q_{1}^{[\mathcal{P}]}, \cdots, Q_{N}^{[\mathcal{P}]}\right)=I\left(W_{1: M} ; Q_{1: N}^{[\mathcal{P}]}\right)=0 \tag{3}
\end{equation*}
$$

The privacy is satisfied by ensuring statistical independence between the queries and the message index set $\mathcal{P}=$ $\left\{i_{1}, \cdots, i_{P}\right\}$, i.e., the privacy constraint is given by,

$$
\begin{equation*}
I\left(Q_{n}^{\left[i_{1}, \cdots, i_{P}\right]} ; i_{1}, \cdots, i_{P}\right)=I\left(Q_{n}^{[\mathcal{P}]} ; \mathcal{P}\right)=0 \tag{4}
\end{equation*}
$$

for all $n \in\{1, \cdots, N\}$.
The $n$th database responds with an answer string $A_{n}^{[\mathcal{P}]}$, which is a deterministic function of the queries and the messages, hence

$$
\begin{equation*}
H\left(A_{n}^{[\mathcal{P}]} \mid Q_{n}^{[\mathcal{P}]}, W_{1: M}\right)=0 \tag{5}
\end{equation*}
$$

We further note that by the data processing inequality and (4),

$$
\begin{equation*}
I\left(A_{n}^{[\mathcal{P}]} ; \mathcal{P}\right)=0, \quad n \in\{1, \cdots, N\} \tag{6}
\end{equation*}
$$

In addition, the user should be able to reconstruct the messages $W_{\mathcal{P}}$ reliably from the collected answers from all databases given the knowledge of the queries. Thus, we write the reliability constraint as,

$$
\begin{gather*}
H\left(W_{i_{1}}, \cdots, W_{i_{P}} \mid A_{1}^{[\mathcal{P}]}, \cdots, A_{N}^{[\mathcal{P}]}, Q_{1}^{[\mathcal{P}]}, \cdots, Q_{N}^{[\mathcal{P}]}\right) \\
=H\left(W_{\mathcal{P}} \mid A_{1: N}^{[\mathcal{P}]}, Q_{1: N}^{[\mathcal{P}]}\right)=0 \tag{7}
\end{gather*}
$$

We denote the retrieval rate of the $i$ th message by $R_{i}$, where $i \in \mathcal{P}$. The retrieval rate of the $i$ th message is the ratio between the length of message $i$ and the total download cost of the message set $\mathcal{P}$ that includes $W_{i}$. Hence,

$$
\begin{equation*}
R_{i}=\frac{H\left(W_{i}\right)}{\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}]}\right)} \tag{8}
\end{equation*}
$$

The sum retrieval rate of $W_{\mathcal{P}}$ is given by,

$$
\begin{equation*}
\sum_{i=1}^{P} R_{i}=\frac{H\left(W_{\mathcal{P}}\right)}{\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}]}\right)}=\frac{P L}{\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}]}\right)} \tag{9}
\end{equation*}
$$

The sum capacity of the MPIR problem is given by

$$
\begin{equation*}
C_{s}^{P}=\sup \sum_{i=1}^{P} R_{i} \tag{10}
\end{equation*}
$$

where the sup is over all private retrieval schemes.
In this paper, we follow the information-theoretic assumptions of large enough message size, large enough field size, and ignore the upload cost as in [8] and [11]-[13]. A formal treatment of the capacity under message and field size constraints for $P=1$ can be found in [25]. We note that the MPIR problem described here reduces to the classical PIR problem when $P=1$, whose capacity is characterized in [8].

## III. Main Results and Discussions

Our first result is the exact characterization of the sum capacity for the case $P \geq \frac{M}{2}$, i.e., when the user wishes to privately retrieve at least half of the messages stored in the databases.

Theorem 1: For the MPIR problem with non-colluding and replicated databases, if the number of desired messages $P$ is at least half of the number of overall stored messages $M$, i.e., if $P \geq \frac{M}{2}$, then the sum capacity is given by,

$$
\begin{equation*}
C_{s}^{P}=\frac{1}{1+\frac{M-P}{P N}} \tag{11}
\end{equation*}
$$

The achievability proof for Theorem 1 is given in Section IV, and the converse proof is given in Section VI-A. We note that when $P=1$, the constraint of Theorem 1 is equivalent to $M=2$, and the result in (11) reduces to the known result of [8] for $P=1, M=2$, which is $\frac{1}{1+\frac{1}{N}}$. We observe that the sum capacity in (11) is a strictly increasing function of $N$, and $C_{s}^{P} \rightarrow 1$ as $N \rightarrow \infty$. We also observe that the sum capacity in this regime is a strictly increasing function of $P$, and approaches $1 \mathrm{as}^{2} P \rightarrow M$.

The following corollary compares our result and the rate corresponding to the repeated use of single-message retrieval scheme [8].

Corollary 1: For the MPIR problem with $P \geq \frac{M}{2}$, the repetition of the single-message retrieval scheme of [8] $P$ times in a row, which achieves a sum rate of,

$$
\begin{equation*}
R_{s}^{r e p}=\frac{(N-1)\left(N^{M-1}+P-1\right)}{N^{M}-1} \tag{12}
\end{equation*}
$$

is strictly sub-optimal with respect to the exact capacity in (11).

Corollary 1 implies that applying Sun-Jafar scheme [8] $P$ times is suboptimal, even if the user uses the undesired symbols, which are downloaded as a byproduct of Sun-Jafar scheme, as a head start in downloading the remaining messages because in this case the user would achieve $R_{s}^{r e p}<C_{s}^{P}$.

Proof: In order to use the single-message capacity achieving PIR scheme as an MPIR scheme, the user repeats the

[^2]single-message achievable scheme for each individual message that belongs to $\mathcal{P}$. We note that at each repetition, the scheme downloads extra decodable symbols from other messages. By this argument, the following rate $R_{s}^{r e p}$ is achievable using a repetition of the single-message scheme,
\[

$$
\begin{equation*}
R_{s}^{r e p}=C+\Delta(M, P, N) \tag{13}
\end{equation*}
$$

\]

where $C$ is the single-message capacity which is given by $C=$ $\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{M}}$ [8], and $\Delta(M, P, N)$ is the rate of the extra decodable symbols that belong to $\mathcal{P}$. To calculate $\Delta(M, P, N)$, we note that the total download cost $D$ is given by $D=\frac{L}{C}$ by definition. Since $L=N^{M}$ in the single-message scheme, $D=\frac{N^{M}\left(1-\left(\frac{1}{N}\right)^{M}\right)}{1-\frac{1}{N}}=\frac{N^{M+1}-N}{N-1}$. The single-message scheme downloads one symbol from every message from every database, i.e., the scheme downloads extra $(P-1) N$ symbols from the remaining desired messages that belong to $\mathcal{P}$, thus,

$$
\begin{equation*}
\Delta(M, P, N)=\frac{(P-1) N(N-1)}{N^{M+1}-N}=\frac{(P-1)(N-1)}{N^{M}-1} \tag{14}
\end{equation*}
$$

Using this in (13) gives the $R_{s}^{r e p}$ expression in (12).
Now, the difference between the capacity in (11) and achievable rate in (12) is,

$$
\begin{align*}
C_{s}^{P}-R_{s}^{r e p} & =\frac{P N}{P(N-1)+M}-\frac{(N-1)\left(N^{M-1}+P-1\right)}{N^{M}-1} \\
& =\frac{\eta(P, M, N)}{\left(N^{M}-1\right)(P(N-1)+M)} \tag{15}
\end{align*}
$$

It suffices to prove that $\eta(P, M, N) \geq 0$ for all $P, M, N$ when $P \geq \frac{M}{2}$ and $N \geq 2$. Note,

$$
\begin{align*}
& \eta(P, M, N) \\
& \quad=(2 P-M) N^{M}+(M-P) N^{M-1}-P(P-1) N^{2} \\
& \quad+((P-1)(2 P-M)-P) N+(M-P)(P-1) \tag{17}
\end{align*}
$$

In the regime $P \geq \frac{M}{2}$, coefficients of $N^{M}, N^{M-1}, N^{0}$ are nonnegative. Denote the negative terms in $\eta(\cdot)$ by $v(P, N)$ which is $v(P, N)=P(P-1) N^{2}+P N$. We note $v(P, N)<P^{2} N^{2}$ when $N>1$, which is the case here. Thus,

$$
\begin{align*}
\geq & (P, M, N) \\
\geq & (2 P-M) N^{M}+(M-P) N^{M-1}+(P-1)(2 P-M) N \\
& +(M-P)(P-1)-P^{2} N^{2}  \tag{18}\\
> & (2 P-M) N^{M}+(M-P) N^{M-1}-P^{2} N^{2}  \tag{19}\\
= & N^{2}\left((2 P-M) N^{M-2}+(M-P) N^{M-3}-P^{2}\right)  \tag{20}\\
\geq & N^{2}\left((2 P-M) 2^{M-2}+(M-P) 2^{M-3}-P^{2}\right)  \tag{21}\\
= & N^{2}\left(2^{M-3}(3 P-M)-P^{2}\right)  \tag{22}\\
\geq & N^{2}\left(2^{M-3} \cdot \frac{M}{2}-M^{2}\right)  \tag{23}\\
= & M N^{2}\left(2^{M-4}-M\right) \tag{24}
\end{align*}
$$

where (21) follows from the fact that $(2 P-M) N^{M-2}+(M-$ $P) N^{M-3}-P^{2}$ is monotone increasing in $N \geq 2$ for $M \geq 3$, and (23) follows from $\frac{M}{2} \leq P \leq M$. From (24), we conclude
that $\eta(M, P, N)>0$ for all $M \geq 7, P \geq \frac{M}{2}$ and $N \geq 2$. Examining the expression in (17) for the remaining cases manually, i.e., when $M \leq 6$, we note that $\eta(M, P, N)>0$ in these cases as well. Therefore, $\eta(M, P, N)>0$ for all possible cases, and the MPIR capacity is strictly larger than the rate achieved by repeating the optimum single-message PIR scheme.

For the example in the introduction, where $M=3, P=2$, $N=2$, our MPIR scheme achieves a sum capacity of $\frac{4}{5}$ in (11), which is strictly larger than the repeating-based achievable sum rate of $\frac{5}{7}$ in (12).

The following corollary gives an achievable rate region for the MPIR problem.

Corollary 2: For the MPIR problem, for the case $P \geq \frac{M}{2}$, the following rate region is achievable,
$\mathcal{C}=\operatorname{conv}\{(C, \delta, \cdots, \delta),(\delta, C, \cdots, \delta), \cdots,(\delta, \cdots, \delta, C)$,

$$
\begin{align*}
& (C, 0,0, \cdots, 0),(0, C, 0, \cdots, 0), \cdots,(0,0, \cdots, C) \\
& \left.(0,0, \cdots, 0),\left(C^{P}, C^{P}, \cdots, C^{P}\right)\right\} \tag{25}
\end{align*}
$$

where

$$
\begin{align*}
C & =\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{M}} \\
C^{P} & =\frac{C_{s}^{P}}{P}=\frac{N}{P N+(M-P)} \\
\delta & =\frac{\Delta(M, P, N)}{P-1}=\frac{N-1}{N^{M}-1} \tag{26}
\end{align*}
$$

and where conv denotes the convex hull, and all corner points lie in the $P$-dimensional space.

Proof: This is a direct consequence of Theorem 1 and Corollary 1. The corner point $\left(C, \frac{\Delta(M, P, N)}{P-1}\right.$, $\left.\frac{\Delta(M, P, N)}{P-1}, \cdots, \frac{\Delta(M, P, N)}{P-1}\right)=\left(\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{M}}, \frac{N-1}{N^{M}-1}, \frac{N-1}{N^{M}-1}, \cdots\right.$, $\left.\frac{N-1}{N^{M}-1}\right)$ is achievable from the single-message achievable scheme. Due to the symmetry of the problem any other permutation for the coordinates of this corner point is also achievable by changing the roles of the desired messages. Theorem 1 gives the symmetric sum capacity corner point for the case of $P \geq \frac{M}{2}$, namely $\left(\frac{C_{s}^{P}}{P}, \frac{C_{s}^{P}}{P}, \cdots, \frac{C_{s}^{P}}{P}\right)=$ $\left(\frac{N}{P N+(M-P)}, \frac{N}{P N+(M-P)}, \cdots, \frac{N}{P N+(M-P)}\right)$. By time sharing of these corner points along with the origin, the region in (25) is achievable.
As an example for this achievable region, consider again the example in the introduction, where $M=3, P=2$, $N=2$. In this case, we have a two-dimensional rate region with three corner points: $\left(\frac{4}{7}, \frac{1}{7}\right)$, which corresponds to the single-message capacity achieving point that aims at retrieving $W_{1} ;\left(\frac{1}{7}, \frac{4}{7}\right)$, which corresponds to single-message capacity achieving point that aims at retrieving $W_{2}$; and $\left(\frac{2}{5}, \frac{2}{5}\right)$, which corresponds to the symmetric sum capacity point. The convex hull of these corner points together with the points on the axes gives the achievable region in Fig. 1. We note that in general, the rate region in Corollary 2 is merely an achievable region.


Fig. 1. The achievable rate region of $M=3, P=2, N=2$.

The capacity region that characterizes the exact tradeoff between the retrieval rates for the $P$ messages remains an open problem despite the optimality of the corner points. A converse argument is needed to show the optimality of time-sharing (if the rate region is indeed the capacity region).

For the case $P \leq \frac{M}{2}$, we have the following result, where the lower and upper bound match if $\frac{M}{P} \in \mathbb{N}$.

Theorem 2: For the MPIR problem with non-colluding and replicated databases, when $P \leq \frac{M}{2}$, the sum capacity is lower and upper bounded as,

$$
\begin{equation*}
\underline{R}_{s} \leq C_{s}^{P} \leq \bar{R}_{s} \tag{27}
\end{equation*}
$$

where the upper bound $\bar{R}_{S}$ is given by,

$$
\begin{align*}
\bar{R}_{s} & =\frac{1}{1+\frac{1}{N}+\cdots+\frac{1}{N^{\left\lfloor\frac{M}{P}\right\rfloor-1}}+\left(\frac{M}{P}-\left\lfloor\frac{M}{P}\right\rfloor\right) \frac{1}{N^{\left\lfloor\frac{M}{P}\right\rfloor}}}  \tag{28}\\
& =\frac{1}{\frac{1-\left(\frac{1}{N} \complement^{\left\lfloor\frac{M}{P}\right\rfloor}\right.}{1-\frac{1}{N}}+\left(\frac{M}{P}-\left\lfloor\frac{M}{P}\right\rfloor\right) \frac{1}{{ }^{\left\lfloor\left\lfloor\frac{M}{P}\right\rfloor\right.}}} \tag{29}
\end{align*}
$$

For the lower bound, define $r_{i}$ as,

$$
\begin{equation*}
r_{i}=\frac{e^{j 2 \pi(i-1) / P}}{N^{1 / P}-e^{j 2 \pi(i-1) / P}}, \quad i=1, \cdots, P \tag{30}
\end{equation*}
$$

where $j=\sqrt{-1}$, and denote $\gamma_{i}, i=1, \cdots, P$, to be the solutions of the linear equations $\sum_{i=1}^{P} \gamma_{i} r_{i}^{-P}=(N-1)^{M-P}$, and $\sum_{i=1}^{P} \gamma_{i} r_{i}^{-k}=0, k=1, \cdots, P-1$, then $\underline{R}_{S}$ is given by,

$$
\begin{equation*}
\underline{R}_{s}=\frac{\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[\left(1+\frac{1}{r_{i}}\right)^{M}-\left(1+\frac{1}{r_{i}}\right)^{M-P}\right]}{\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[\left(1+\frac{1}{r_{i}}\right)^{M}-1\right]} \tag{31}
\end{equation*}
$$

The achievability lower bound in Theorem 2 is shown in Section V and the upper bound is derived in Section VI-B. The following corollary states that the bounds in Theorem 2


Fig. 2. Summary of the state of the results.
match if the total number of messages is an integer multiple of the number of desired messages.

Corollary 3: For the MPIR problem with non-colluding and replicated databases, if $\frac{M}{P}$ is an integer, then the bounds in (27) match, and hence, for ${ }^{3} N>1$

$$
\begin{equation*}
C_{s}^{P}=\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{\frac{M}{P}}}, \quad \frac{M}{P} \in \mathbb{N} \tag{32}
\end{equation*}
$$

Proof: For the upper bound, observe that if $\frac{M}{P} \in \mathbb{N}$, then $\frac{M}{P}=\left\lfloor\frac{M}{P}\right\rfloor$. Hence, (28) becomes

$$
\begin{equation*}
\bar{R}_{s}=\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{\frac{M}{P}}} \tag{33}
\end{equation*}
$$

For the lower bound, consider the case $\frac{M}{P} \in \mathbb{N}$. From (30),

$$
\begin{equation*}
\left(1+\frac{1}{r_{i}}\right)^{M}=\left(\frac{N^{1 / P}}{e^{j 2 \pi(i-1) / P}}\right)^{M}=N^{\frac{M}{P}} \tag{34}
\end{equation*}
$$

since $e^{j 2 \pi(i-1) M / P}=1$ for $\frac{M}{P} \in \mathbb{N}$. Similarly, $\left(1+\frac{1}{r_{i}}\right)^{M-P}=N^{\frac{M}{P}-1}$. Hence, if $\frac{M}{P} \in \mathbb{N}$,

$$
\begin{align*}
\underline{R}_{S} & =\frac{\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[N^{\frac{M}{P}}-N^{\frac{M}{P}-1}\right]}{\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[N^{\frac{M}{P}}-1\right]}  \tag{35}\\
& =\frac{N^{\frac{M}{P}}-N^{\frac{M}{P}-1}}{N^{\frac{M}{P}}-1}  \tag{36}\\
& =\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{\frac{M}{P}}} \tag{37}
\end{align*}
$$

Thus, $\underline{R}_{s}=C_{s}^{P}=\bar{R}_{s}$ if $\frac{M}{P} \in \mathbb{N}$, and we have an exact capacity result in this case.
${ }^{3}$ If $N=1$, the optimal retrieval scheme is to download the contents of the database, hence $C_{S}^{P}=\frac{P}{M}$.


Fig. 3. Deviation of the achievable sum rate from the upper bound. (a) $N=2$. (b) $N=5$. (c) $N=10$. (d) $N=20$.

Examining the result, we observe that when the total number of messages is an integer multiple of the number of desired messages, the sum capacity of the MPIR is the same as the capacity of the single-message PIR with the number of messages equal to $\frac{M}{P}$. Note that, although at first the result may seem as if every $P$ messages can be lumped together as a single message, and the achievable scheme in [8] can be used, this is not the case. The reason for this is that, we need to ensure the privacy constraint for every subset of messages of size ${ }^{4} P$. That is why, in this paper, we develop a new achievable scheme.

The state of the results is summarized in Fig. 2: Consider the $(M, P)$ plane, where naturally $M \geq P$. The valid part of the

[^3]plane is divided into two regions. The first region is confined between the lines $P=\frac{M}{2}$ and $P=M$; the sum capacity in this region is exactly characterized (Theorem 1). The second region is confined between the lines $P=1$ and $P=\frac{M}{2}$; the sum capacity in this region is characterized only for the cases when $\frac{M}{P} \in \mathbb{N}$ (Corollary 3). The line $P=1$ corresponds to the previously known result for the single-message PIR [8]. The exact capacity for the rest of the cases is still an open problem; however, the achievable scheme in Theorem 2 yields nearoptimal sum rates for all the remaining cases with the largest difference of 0.0082 from the upper bound, as discussed next.

Fig. 3 shows the difference of the achievable rate $\underline{R}_{s}$ and the upper bound $\bar{R}_{s}$ in Theorem 2, i.e., $\bar{R}_{s}-\underline{R}_{s}$. The figure shows that the difference decreases as $N$ increases. This difference in all cases is small and is upper bounded by 0.0082 , which occurs when $M=5, P=2, N=2$. In addition, the difference is zero for the cases $P \geq \frac{M}{2}$ (Theorem 1) or $\frac{M}{P} \in \mathbb{N}$ (Corollary 3).


Fig. 4. Effect of changing $M$ for fixed $P=5,6,10$ and fixed $N=2$.


Fig. 5. Effect of changing $N$ for fixed $(M, P)=(5,2),(10,5),(20,3)$.

Fig. 4 shows the effect of changing $M$ for fixed $(P, N)$. We observe that as $M$ increases, the sum rate monotonically decreases and has a limit of ${ }^{5} 1-\frac{1}{N}$. In addition, Fig. 5 shows the effect of changing $N$ for fixed $(P, M)$. We observe that as $N$ increases, the sum rate increases and approaches 1 , as expected.

## IV. Achievability Proof for the Case $P \geq \frac{M}{2}$

In this section, we present the general achievable scheme that attains the upper bound for the case $P \geq \frac{M}{2}$. The scheme applies the concepts of message symmetry, database symmetry, and exploiting side information as in [8]. However, our scheme requires the extra ingredient of MDS coding of the desired

[^4]symbols and the side information in its second stage. We note also that, here, by message symmetry, we mean symmetry across group of messages of size $P$, which is realized by MDS coding.

## A. Motivating Example: $M=3, P=2$ Messages, $N=2$ Databases

We start with a simple motivating example in this subsection. The scheme operates over message size $N^{2}=4$. For sake of clarity, we assume that the three messages after interleaving their indices are $W_{1}=\left(a_{1}, \cdots, a_{4}\right)^{T}, W_{2}=$ $\left(b_{1}, \cdots, b_{4}\right)^{T}$, and $W_{3}=\left(c_{1}, \cdots, c_{4}\right)^{T}$. We use $\mathbf{G}_{2 \times 3}$ ReedSolomon generator matrix over $\mathbb{F}_{3}$ as

$$
\mathbf{G}_{2 \times 3}=\left[\begin{array}{lll}
1 & 1 & 1  \tag{38}\\
1 & 2 & 0
\end{array}\right]
$$

The user picks a random permutation for the columns of $\mathbf{G}_{2 \times 3}$ from the 6 possible permutations, e.g., in this example we use the permutation $2,1,3$. In the first round, the user starts by downloading one symbol from each database and each message, i.e., the user downloads $\left(a_{1}, b_{1}, c_{1}\right)$ from the first database, and $\left(a_{2}, b_{2}, c_{2}\right)$ from the second database. In the second round, the user encodes the side information from database 2 which is $c_{2}$ with two new symbols from $W_{1}, W_{2}$ which are $\left(a_{3}, b_{3}\right)$ using the permuted generator matrix, i.e., the user downloads two equations from database 1 in the second round,

$$
\mathbf{G S}_{1}\left[\begin{array}{l}
a_{3}  \tag{39}\\
b_{3} \\
c_{2}
\end{array}\right]=\left[\begin{array}{lll}
1 & 1 & 1 \\
1 & 2 & 0
\end{array}\right]\left[\begin{array}{lll}
0 & 1 & 0 \\
1 & 0 & 0 \\
0 & 0 & 1
\end{array}\right]\left[\begin{array}{l}
a_{3} \\
b_{3} \\
c_{2}
\end{array}\right]=\left[\begin{array}{c}
a_{3}+b_{3}+c_{2} \\
2 a_{3}+b_{3}
\end{array}\right]
$$

The user repeats this operation for the second database with $\left(a_{4}, b_{4}\right)$ as desired symbols and $c_{1}$ as the side information from the first database.

For the decodability: The user subtracts out $c_{2}$ from round two in the first database, then the user can decode $\left(a_{3}, b_{3}\right)$ from $a_{3}+b_{3}$ and $2 a_{3}+b_{3}$. Similarly, by subtracting out $c_{1}$ from round two in the second database, the user can decode $\left(a_{4}, b_{4}\right)$ from $a_{4}+b_{4}$ and $2 a_{4}+b_{4}$.

For the privacy: Single bit retrievals of $\left(a_{1}, b_{1}, c_{1}\right)$ and $\left(a_{2}, b_{2}, c_{2}\right)$ from the two databases in the first round satisfy message symmetry and database symmetry, and do not leak any information. In addition, due to the private shuffling of bit indices, the different coefficients of 1,2 and 0 in front of the bits in the MDS-coded summations in the second round do not leak any information either; see a formal proof in Section IV-C. To see the privacy constraint intuitively from another angle, we note that the user can alter the queries for the second database when the queries for the first database are fixed, when the user wishes to retrieve another set of two messages. For instance, if the user wishes to retrieve $\left(W_{1}, W_{3}\right)$ instead of $\left(W_{1}, W_{2}\right)$, it can alter the queries for the second database by changing every $c_{2}$ in the queries of the second database with $c_{3}, c_{1}$ with $c_{4}, b_{2}$ with $b_{3}$, and $b_{4}$ with $b_{1}$.

The query table for this case is shown in Table I. The scheme retrieves $a_{1}, \cdots, a_{4}$ and $b_{1}, \cdots, b_{4}$, i.e., 8 bits in 10 downloads ( 5 from each database). Thus, the achievable

TABLE I
The Query Table for the Case $M=3, P=2, N=2$

| Database 1 | Database 2 |
| :---: | :---: |
| $a_{1}, b_{1}, c_{1}$ | $a_{2}, b_{2}, c_{2}$ |
| $a_{3}+b_{3}+c_{2}$ | $a_{4}+b_{4}+c_{1}$ |
| $2 a_{3}+b_{3}$ | $2 a_{4}+b_{4}$ |

sum rate for this scheme is $\frac{8}{10}=\frac{4}{5}=\frac{1}{1+\frac{M-P}{P N}}$. If we use the single-message optimal scheme in [8], which is given in [8, Example 4.3] for this specific case, twice in a row to retrieve two messages, we achieve a sum rate of $\frac{20}{28}=\frac{5}{7}<\frac{4}{5}$ as discussed in the introduction.

## B. General Achievable Scheme

The scheme requires $L=N^{2}$, and is completed in two rounds. The main ingredient of the scheme is the MDS coding of the desired symbols and side information in the second round. The details of the scheme are as follows.

1) Index preparation: The user interleaves the contents of each message randomly and independently from the remaining messages using a random interleaver $\pi_{m}($. which is known privately to the user only, i.e.,

$$
\begin{equation*}
x_{m}(i)=w_{m}\left(\pi_{m}(i)\right), \quad i \in\{1, \cdots, L\} \tag{40}
\end{equation*}
$$

where $X_{m}=\left[x_{m}(1), \cdots, x_{m}(L)\right]^{T}$ is the interleaved message. Thus, the downloaded symbol $x_{m}(i)$ at any database appears to be chosen at random and independent from the desired message subset $\mathcal{P}$.
2) Round one: As in [8], the user downloads one symbol from every message from every database, i.e., the user downloads $\left(x_{1}(n), x_{2}(n), \cdots, x_{M}(n)\right)$ from the $n$th database. This implements message symmetry, symmetry across databases, and satisfies the privacy constraint.
3) Round two: The user downloads a coded mixture of new symbols from the desired messages and the undesired symbols downloaded from the other databases. Specifically,
a) The user picks an MDS generator matrix $\mathbf{G} \in$ $\mathbb{F}_{q}^{P \times M}$, which has the property that every $P \times P$ submatrix is full-rank. This implies that if the user can cancel out any $M-P$ symbols from the mixture, the remaining symbols can be decoded. One explicit MDS generator matrix is the ReedSolomon generator matrix over $\mathbb{F}_{q}$, where $q>M$, [26], [27]. The matrix is constructed by choosing $M$ distinct elements of $\mathbb{F}_{q}$. Let us denote these elements by $\left\{\theta_{1}, \theta_{2}, \cdots, \theta_{M}\right\}$. Then,

$$
\mathbf{G}=\left[\begin{array}{ccccc}
1 & 1 & 1 & \cdots & 1  \tag{41}\\
\theta_{1} & \theta_{2} & \theta_{3} & \cdots & \theta_{M} \\
\theta_{1}^{2} & \theta_{2}^{2} & \theta_{3}^{2} & \cdots & \theta_{M}^{2} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\theta_{1}^{P-1} & \theta_{2}^{P-1} & \theta_{3}^{P-1} & \cdots & \theta_{M}^{P-1}
\end{array}\right]_{P \times M}
$$

b) The user picks uniformly and independently at random the permutation matrices $\mathbf{S}_{1}, \mathbf{S}_{2}, \cdots, \mathbf{S}_{N-1}$ of size $M \times M$. These matrices shuffle the order of the columns of $\mathbf{G}$ to be independent of $\mathcal{P}$.
c) At the first database, the user downloads an MDS-coded version of $P$ new symbols from the desired set $\mathcal{P}$ and $M-P$ undesired symbols that are already decoded from the second database in the first round, i.e., the user downloads $P$ equations of the form

$$
\begin{align*}
& \mathbf{G S}_{1}\left[x_{i_{1}}(N+1) \quad x_{i_{2}}(N+1) \cdots x_{i_{P}}(N+1)\right. \\
& \left.x_{j_{1}}(2) \quad x_{j_{2}}(2) \cdots x_{j_{M-P}}(2)\right]^{T} \tag{42}
\end{align*}
$$

where $\mathcal{P}=\left\{i_{1}, i_{2}, \cdots, i_{P}\right\}$ are the indices of the desired messages and $\overline{\mathcal{P}}=\left\{j_{1}, j_{2}, \cdots, j_{M-P}\right\}$ are the indices of the undesired messages. In this case, the user can cancel out the undesired messages and be left with a $P \times P$ invertible system of equations that it can solve to get $\left[x_{i_{1}}(N+1), x_{i_{2}}(N+\right.$ 1), $\left.\cdots, x_{i_{P}}(N+1)\right]$. This implements exploiting side information as in [8].
d) The user repeats the last step for each set of side information from database 3 to database $N$, each with different permutation matrix.
e) By database symmetry, the user repeats all steps of round two at all other databases.

## C. Decodability, Privacy, and Calculation of the Achievable Rate

Now, we verify that this achievable scheme satisfies the reliability and privacy constraints.

For the reliability: The user gets individual symbols from all databases in the first round, and hence they are all decodable by definition. In the second round, the user can subtract out all the undesired message symbols using the undesired symbols downloaded from all other databases during the first round. Consequently, the user is left with a $P \times P$ system of equations which is guaranteed to be invertible by the MDS property, hence all symbols that belong to $W_{\mathcal{P}}$ are decodable.

For the privacy: At each database, for every message subset $\mathcal{P}$ of size $P$, the achievable scheme retrieves randomly interleaved symbols which are encoded by the following matrix:

$$
\mathbf{H}_{\mathcal{P}}=\left[\begin{array}{ccccc}
\mathbf{I}_{P} & \mathbf{0}_{P} & \mathbf{0}_{P} & \cdots & \mathbf{0}_{P}  \tag{43}\\
\mathbf{0}_{P} & \mathbf{G}_{\mathcal{P}}^{1} & \mathbf{0}_{P} & \cdots & \mathbf{0}_{P} \\
\mathbf{0}_{P} & \mathbf{0}_{P} & \mathbf{G}_{\mathcal{P}}^{2} & \cdots & \mathbf{0}_{P} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0}_{P} & \mathbf{0}_{P} & \mathbf{0}_{P} & \cdots & \mathbf{G}_{\mathcal{P}}^{N-1}
\end{array}\right]
$$

where $\mathbf{G}_{\mathcal{P}}^{n}=\mathbf{G S}_{n}(:, \mathcal{P})$ are the columns of the encoding matrix that correspond to the message subset $\mathcal{P}$ after applying the random permutation $\mathbf{S}_{n}$. Since the permutation matrices are chosen uniformly and independently from each other, the probability distribution of $\mathbf{H}_{\mathcal{P}}$ is uniform irrespective to $\mathcal{P}$ (the probability of realizing such a matrix is $\left(\frac{(M-P)!}{M!}\right)^{N-1}$ ). Furthermore, the symbols are chosen randomly and uniformly
by applying the random interleaver. Hence, the retrieval scheme is private.

To calculate the achievable rate: We note that at each database, the user downloads $M$ individual symbols in the first round that includes $P$ desired symbols. The user exploits the side information from the remaining $(N-1)$ databases to generate $P$ equations for each side information set. Each set of $P$ equations in turn generates $P$ desired symbols. Hence, the achievable rate is calculated as,

$$
\begin{align*}
\sum_{i=1}^{P} R_{i} & =\frac{\text { total number of desired symbols }}{\text { total downloaded equations }}  \tag{44}\\
& =\frac{N(P+P(N-1))}{N(M+P(N-1))}  \tag{45}\\
& =\frac{P N}{(M-P)+P N}  \tag{46}\\
& =\frac{1}{1+\frac{M-P}{P N}} \tag{47}
\end{align*}
$$

## D. Further Examples for the Case $P \geq \frac{M}{2}$

In this section, we illustrate our achievable scheme with two more basic examples. In Section IV-A, we considered the case $M=3, P=2, N=2$. In the next two sub-sections, we will consider examples with larger $M, P$ (Section IV-D.1), and larger $N$ (Section IV-D.2).

1) $M=5$ Messages, $P=3$ Messages, $N=2$ Databases: Let $\mathcal{P}=\{1,2,3\}$, and $a$ to $e$ denote the contents of $W_{1}$ to $W_{5}$, respectively. The achievable scheme is similar to the example in Section IV-A. The main difference is that in this case, we use $5 \times 5$ permutation matrix for $\mathbf{S}_{1}$ and $\mathbf{G}_{3 \times 5}$ Reed-Solomon generator matrix over $\mathbb{F}_{5}$ as:

$$
\mathbf{G}_{3 \times 5}=\left[\begin{array}{lllll}
1 & 1 & 1 & 1 & 1  \tag{48}\\
1 & 2 & 3 & 4 & 0 \\
1 & 4 & 4 & 1 & 0
\end{array}\right]
$$

The query table is shown in Table II with the following random permutation for the columns: $2,5,1,3,4$. The reliability and privacy constraints are satisfied due to the MDS property that implies that any subset of 3 messages corresponds to a $3 \times 3$ invertible submatrix if the remaining symbols are decodable from the other database. This scheme retrieves $a_{1}, \cdots, a_{4}, b_{1}, \cdots, b_{4}$ and $c_{1}, \cdots, c_{4}$, hence 12 bits in 16 downloads ( 8 from each database). Thus, the achievable sum rate is $\frac{12}{16}=\frac{3}{4}$ which equals the sum capacity $\frac{1}{1+\frac{M-P}{P N}}$ in (11). This strictly outperforms the repetition-based achievable sum rate $\frac{18}{31}$ in (12).
2) $M=4$ Messages, $P=2$ Messages, $N=3$ Databases: Next, we give an example with a larger $N$. Here, the message size is $N^{2}=9$. With a generator matrix $\mathbf{G}_{2 \times 4}=\mathbf{G}_{3 \times 5}$ ([1:2], [1:4]) to be the upper left submatrix of the previous example and two set of random permutations (corresponding to $\mathbf{S}_{1}, \mathbf{S}_{2}$ ) as $1,3,2,4$, and $4,1,3,2$. The query table is shown in Table III. This scheme retrieves $a_{1}, \cdots, a_{9}$ and $b_{1}, \cdots, b_{9}$, hence 18 bits in 24 downloads ( 8 from each database). Thus, the achievable rate is $\frac{18}{24}=\frac{3}{4}=\frac{1}{1+\frac{M-P}{P N}}$. This strictly

TABLE II
The Query Table for $M=5, P=3, N=2$

| Database 1 | Database 2 |
| :---: | :---: |
| $a_{1}, b_{1}, c_{1}, d_{1}, e_{1}$ | $a_{2}, b_{2}, c_{2}, d_{2}, e_{2}$ |
| $a_{3}+b_{3}+c_{3}+d_{2}+e_{2}$ | $a_{4}+b_{4}+c_{4}+d_{1}+e_{1}$ |
| $2 a_{3}+c_{3}+3 d_{2}+4 e_{2}$ | $2 a_{4}+c_{4}+3 d_{1}+4 e_{1}$ |
| $4 a_{3}+c_{3}+4 d_{2}+e_{2}$ | $4 a_{4}+c_{4}+4 d_{1}+e_{1}$ |

outperforms the repetition-based achievable scheme sum rate $\frac{7}{10}$ in (12).

## V. Achievability Proof for the Case $P \leq \frac{M}{2}$

In this section, we describe an achievable scheme for the case $P \leq \frac{M}{2}$. We show that this scheme is optimal when the total number of messages $M$ is an integer multiple of the number of desired messages $P$. The scheme incurs a small loss from the upper bound for all other cases. The scheme generalizes the ideas in [8]. Different than [8], our scheme uses unequal number of stages for each round of download. Interestingly, the number of stages at each round can be thought of as the output of an all-poles IIR filter. Our scheme reduces to [8] if we let $P=1$. In the sequel, we define the $i$ th round as the download queries that retrieve sum of $i$ different symbols. We define the stage as a block of queries that exhausts all $\binom{M}{i}$ combinations of the sum of $i$ symbols in the $i$ th round.

## A. Motivating Example: $M=5, P=2$ Messages, $N=2$ Databases

To motivate our achievable scheme, consider the case of retrieving two messages denoted by letters $(a, b)$ from five stored messages denoted by letters ( $a, b, c, d, e$ ). Instead of designing the queries beginning from the top as usual, i.e., beginning by downloading individual symbols, we design the scheme backwards starting from the last round that corresponds to downloading sums of all five messages and trace back to identify the side information needed at each round from the other database. Our steps described below can be followed through in the query table in Table IV.

Now, let us fix the number of stages in the 5th round to be 1 as in [8] since $N=2$. Round 5 corresponds to downloading the sum of all five messages and contains one combination of symbols $a+b+c+d+e$; please see the last line in Table IV. Since we wish to retrieve $(a, b)$, we need one side information equation in the form of $c+d+e$ from earlier rounds. The combination $c+d+e$ can be created directly from round 3 without using round 4 . Hence, we suppress round 4, as it does not create any useful side information in our case, and download one stage from round 3 to generate one side information equation $c+d+e$.

In round 3, we download sums of 3 messages. Each stage of round 3 consists of $\binom{5}{3}=10$ equations. One of those 10 equations is in the desired $c+d+e$ form, and the remaining 9 of them have either $a$ or $b$ or both $a, b$ in them. In tabulating all these 9 combinations, we recognize two categories of side information equations needed from earlier rounds. The first

TABLE III
The Query Table for the Case $M=4, P=2, N=3$

| Database 1 | Database 2 | Database 3 |
| :---: | :---: | :---: |
| $a_{1}, b_{1}, c_{1}, d_{1}$ | $a_{2}, b_{2}, c_{2}, d_{2}$ | $a_{3}, b_{3}, c_{3}, d_{3}$ |
| $a_{4}+b_{4}+c_{2}+d_{2}$ | $a_{6}+b_{6}+c_{1}+d_{1}$ | $a_{8}+b_{8}+c_{1}+d_{1}$ |
| $a_{4}+3 b_{4}+2 c_{2}+4 d_{2}$ | $a_{6}+3 b_{6}+2 c_{1}+4 d_{1}$ | $a_{8}+3 b_{8}+2 c_{1}+4 d_{1}$ |
| $a_{5}+b_{5}+c_{3}+d_{3}$ | $a_{7}+b_{7}+c_{3}+d_{3}$ | $a_{9}+b_{9}+c_{2}+d_{2}$ |
| $4 a_{5}+b_{5}+3 c_{3}+2 d_{3}$ | $4 a_{7}+b_{7}+3 c_{3}+2 d_{3}$ | $4 a_{9}+b_{9}+3 c_{2}+2 d_{2}$ |

category corresponds to equations of the form $a+b+(c, d, e)$, where ( $c, d, e$ ) means possible choices for the rest of the equation, i.e., these equations have both $a$ and $b$ in them and plus one more symbol in the form of $c$ or $d$ or $e$. This category requires downloading one stage of individual symbols (i.e., an individual $c$ or $d$ or $e$ ), that is, one stage of round 1. We note also that one of the symbols $(a, b)$ should be known as a side information from the second database in order to solve for the remaining new symbol. The second category corresponds to equations of the form $a+(c+d, c+e, d+e)$ and $b+(c+d, c+e, d+e)$, i.e., these equations have only one of $a$ or $b$ but not both. This category requires two stages of round 2, as we need different side information equations that contain sum of twos, e.g., $c+d, c+e, d+e$.

In round 2, we download sums of 2 messages. Each stage of the second round contains $\binom{5}{2}=10$ equations. In each stage, we need one category of side information equations, which is $a+(c, d, e)$ and $b+(c, d, e)$. This necessitates two different stages of individual symbols, i.e., two stages of round 1 for each stage of round 2.
Denoting $\alpha_{i}$ to be the number of stages needed for the $i$ th round, we sum all the required stages for round 1 to be $\alpha_{1}=2 \cdot 2+1=5$ stages. Hence, the user identifies the number stages as $\alpha_{1}=5, \alpha_{2}=2, \alpha_{3}=1, \alpha_{4}=0, \alpha_{5}=1$. These can be observed in the query table in Table IV. Note that, we have $\alpha_{1}=5$ stages in round 1 where we download individual bits; then we have $\alpha_{2}=2$ stages in round 2 where we download sums of two symbols; then we have $\alpha_{3}=1$ stage in round 3 where we download sums of three symbols; we skip round 4 as $\alpha_{4}=0$; and we have $\alpha_{5}=1$ stage of round 5 where we download sum of all five symbols.

Now, after designing the structure of the queries and the number of stages needed for each round, we apply the rest of the scheme described in [8]. The user randomly interleaves the messages as usual. In the first round, the user downloads one symbol from each message at each database. This is repeated $\alpha_{1}=5$ times for each database. Hence, the user downloads $a_{1: 10}, b_{1: 10}, c_{1: 10}, d_{1: 10}, e_{1: 10}$ from the two databases. In the second round, the user downloads sums of two messages. Each stage contains $\binom{5}{2}=10$ equations. This is repeated $\alpha_{2}=2$ times. For example, in database 1 , user exploits $c_{6}, d_{6}, e_{6}$ to get $a_{12}, a_{13}, a_{14}$ and $c_{7}, d_{7}, e_{7}$ to obtain $b_{11}, b_{12}, b_{13}$. These are from round 1. Round 2 generates $c_{11}+d_{11}, c_{12}+e_{11}, d_{12}+e_{12}$ from stage 1 , and $c_{13}+d_{13}$, $c_{14}+e_{13}, d_{14}+e_{14}$ from stage 2 as side information for round 3 . In round 3, the user downloads sum of three symbols. There are $\binom{5}{3}=10$ of them. Symbols $c_{10}, d_{10}, e_{10}$ downloaded from
round 1 in database 2 are used to be summed with mixtures of $a+b$. The two sets of side information generated in the second round are exploited in the equations that have one $a$ or $b$. Note that for each such equation, one of $a$ or $b$ is new and the other one is decoded from database 2. Round 3 generates one side information as $c_{19}+d_{19}+e_{19}$ that is used in round 5 . This last round includes the sum of all five messages.

Therefore, as seen in Table IV, we have retrieved $a_{1}, \cdots, a_{34}$ and $b_{1} \cdots, b_{34}$, i.e., 68 bits in a total of 112 downloads ( 56 from each database). Thus, the achievable sum rate is $\frac{68}{112}=\frac{17}{28}$. This is $\underline{R}_{s}$ in Theorem 2, whereas the upper bound $\bar{R}_{S}$ in Theorem 2 is $\frac{1}{1+\frac{1}{N}+\frac{1}{2 N^{2}}}=\frac{8}{13}$. The gap between $\underline{R}_{s}$ and $\bar{R}_{s}$ is equal to $\frac{3}{364} \simeq 0.0082$, which also is the largest possible gap between $\underline{R}_{s}$ and $\bar{R}_{S}$ over all possible values of $M$, $P$ and $N$.

## B. Calculation of the Number of Stages

The main new ingredient of our scheme in comparison to the scheme in [8] is the unequal number of stages in each round. In [8], the scheme is completed in $M$ rounds, and each round contains only 1 stage only when $N=2$. To generalize the ideas in Section V-A and calculate the number of stages needed per round, we use Vandermonde's identity

$$
\begin{equation*}
\binom{M}{i}=\sum_{k=0}^{P}\binom{P}{k}\binom{M-P}{i-k} \tag{49}
\end{equation*}
$$

The relation in (49) states that any combination of $i$ objects from a group of $M$ objects must have $k$ objects from a group of size $P$ and $i-k$ objects from a group of size $M-P$. In our context, the first group is the subset of the desired messages and the second group is the subset of the undesired messages. Then, the relation can be interpreted in our setting as follows: In the $i$ th round, the $\binom{M}{i}$ combinations of all possible sums of $i$ terms can be sorted into $P+1$ categories: The first category (i.e., $k=0$ ), contains no terms from the desired messages, the second category contains 1 term from the desired messages and $i-1$ terms from the undesired messages, etc. The relation gives also the number of query subgroups of each category $\binom{P}{k}$ and the number of queries in each subgroup $\binom{M-P}{i-k}$.
Let us consider the following concrete example for clarification: Consider that we have 6 messages denoted by ( $a, b, c, d, e, f$ ), and the desired group to be retrieved is $(a, b)$. Consider round 4 that consists of all combinations of sums of 4 symbols. From Vandermonde's identity, we know that $\binom{6}{4}=\binom{2}{0}\binom{4}{4}+\binom{2}{1}\binom{4}{3}+\binom{2}{2}\binom{4}{2}$. Which means that there are

TABLE IV
The Query Table for the Case $M=5, P=2, N=2$

three categories of sums: First category is with only undesired messages; we have $\binom{2}{0}=1$ query subgroup of the form $c+d+e+f$. The second category is to have 1 term from the desired group and the remaining are undesired; we have $\binom{2}{1}=2$ query subgroups, one corresponds to $a$ with combinations of 3 terms from $c, d, e, f$, and the other to $b$ with combinations of 3 terms from $c, d, e, f$. Each query subgroup contains $\binom{4}{3}$ queries, i.e., the first query subgroup is of the form $a+(c+d+e, c+d+f, c+e+f, d+e+f)$ and the second query subgroup is of the form $b+(c+d+e, c+$ $d+f, c+e+f, d+e+f)$. Third category is to have 2 terms from the desired group and 2 terms from the undesired group; we have $\binom{2}{2}=1$ query subgroup of this category that takes the form $a+b+(c+d, c+e, \cdots)$. The number of queries of this
group is $\binom{4}{2}$ corresponding to all combinations of 2 undesired symbols.
Back to the calculation of the number of stages: To be able to cancel the undesired symbols from an $i$-term sum, the user needs to download these undesired symbols as side information in the previous rounds. Hence, round $i$ requires downloading $\binom{P}{1}$ stages in round $(i-1),\binom{P}{2}$ stages in round ( $i-2$ ), etc. Note that these stages need to be downloaded from the remaining $(N-1)$ databases. Then, each database needs to download $\frac{1}{N-1}\binom{P}{1}$ stages in round $(i-1), \frac{1}{N-1}\binom{P}{2}$ stages in round ( $i-2$ ), etc.

From this observation, we can trace back the number of stages needed at each round. Denote $\alpha_{i}$ to be the number of stages in round $i$. Fix the number of stages in the last
round (round $M$ ) to be $\alpha_{M}=(N-1)^{M-P}$ stages. This choice ensures that the number of stages in any round is an integer. Note that in round $M$, the user downloads a sum of all $M$ messages, this requires side information in the form of the sum of the undesired $M-P$ messages. Hence, we suppress the rounds $M-P+1$ through $M-1$ since they do not generate any useful side information. Note that the side information equations in round $M$ at each database are collected from the remaining $(N-1)$ databases. Then, the number of stages in round $(M-P)$ should be $(N-1)^{M-P-1}$. Therefore, we write

$$
\begin{align*}
\alpha_{M} & =(N-1)^{M-P}  \tag{50}\\
\alpha_{M-1} & =\cdots=\alpha_{M-P+1}=0  \tag{51}\\
\alpha_{M-P} & =\frac{1}{N-1} \alpha_{M}=\frac{1}{N-1} \sum_{i=1}^{P}\binom{P}{i} \alpha_{M-P+i} \tag{52}
\end{align*}
$$

Now, in round $(M-P)$, each stage requires $\binom{P}{1}$ stages from round $(M-P-1),\binom{P}{2}$ stages from round $(M-P-2)$, and so on so forth, and these stages are divided across $(N-1)$ databases. Continuing with the same argument, for each round, we write

$$
\begin{align*}
& \alpha_{M-P-1}=\frac{1}{N-1}\binom{P}{1} \alpha_{M-P}=\frac{1}{N-1} \sum_{i=1}^{P}\binom{P}{i} \alpha_{M-P-1+i} \\
& \alpha_{M-P-2}=\frac{1}{N-1}\binom{P}{1} \alpha_{M-P-1}+\frac{1}{N-1}\binom{P}{2} \alpha_{M-P}  \tag{53}\\
&=\frac{1}{N-1} \sum_{i=1}^{P}\binom{P}{i} \alpha_{M-P-2+i}  \tag{54}\\
& \vdots  \tag{55}\\
& \alpha_{k}=\frac{1}{N-1} \sum_{i=1}^{P}\binom{P}{i} \alpha_{k+i}
\end{align*}
$$

Interestingly, this pattern closely resembles the output of an IIR filter $y[n][24]$, with the difference equation,

$$
\begin{equation*}
y[n]=\frac{1}{N-1} \sum_{i=1}^{P}\binom{P}{i} y[n-i] \tag{56}
\end{equation*}
$$

and with the initial conditions $y[-P]=(N-1)^{M-P}, y[-P+$ $1]=\cdots=y[-1]=0$. Note that the only difference between the two seemingly different settings is the orientation of the time axis. The calculation of the number of stages is obtained backwards in contrast to the output of this IIR filter. Hence, we can systematically obtain the number of stages at each round by observing the output of the IIR filter characterized by (56), and mapping it to the number of stages via $\alpha_{k}=y[(M-P)-k]$.

We note that for the special case $P=1$, the number of stages can be obtained from the first order filter $y[n]=$ $\frac{1}{N-1} y[n-1]$. The output of this filter is $y[n]=(N-1)^{M-2-n}$. Then, the number of stages in round $k$ is $\alpha_{k}=y[M-1-k]=$ $(N-1)^{k-1}$, which is exactly the number of stages used in [8]; in particular if $N=2$, then $\alpha_{k}=1$ for all $k$.

## C. General Achievable Scheme

1) Index preparation: The user interleaves the contents of each message randomly and independently from the remaining messages using a random interleaver $\pi_{m}($. which is known privately to the user only, i.e.,

$$
\begin{equation*}
x_{m}(i)=w_{m}\left(\pi_{m}(i)\right), \quad i \in\{1, \cdots, L\} \tag{57}
\end{equation*}
$$

2) Number of stages: We calculate the number of stages needed in each round. This can be done systematically by finding the output of the IIR filter characterized by,

$$
\begin{equation*}
y[n]=\frac{1}{N-1} \sum_{i=1}^{P}\binom{P}{i} y[n-i] \tag{58}
\end{equation*}
$$

with the initial conditions $y[-P]=(N-$ $1)^{M-P}, y[-P+1]=\cdots=y[-1]=0$. The number of stages in round $i$ is $\alpha_{i}=y[(M-P)-i]$ as discussed in Section V-B.
3) Initialization: From the first database, the user downloads one symbol from each message that belongs to the desired message set $\mathcal{P}$. The user sets the round index to $i=1$.
4) Message symmetry: In round $i$, the user downloads sums of $i$ terms from different symbols from the first database. To satisfy the privacy constraint, the user should download an equal amount of symbols from all messages. Therefore, the user downloads the remaining $\binom{M-P}{i}$ combinations in round $i$ from the undesired symbol set $\overline{\mathcal{P}}$. For example: In round 1 , the user downloads one symbol from every undesired message with a total of $\binom{M-P}{1}=M-P$ such symbols.
5) Repetition of stages: In the first database, the user repeats the operation in round $i$ according to the number of calculated stages $\alpha_{i}$. This in total results in downloading $\alpha_{i}\binom{M-P}{i}$ undesired equations, and $\alpha_{i}\left(\binom{M}{i}-\binom{M-P}{i}\right)$ desired equations.
6) Symmetry across databases: The user implements symmetry across databases by downloading $\alpha_{i}\binom{M-P}{i}$ new undesired equations, and $\alpha_{i}\left(\binom{M}{i}-\binom{M-P}{i}\right)$ new desired equations from each database. These undesired equations will be used as side information in subsequent rounds. For example: In round 1, each database generates $\alpha_{1}(M-P)$ undesired equations in the form of individual symbols. Hence, each database can exploit up to $\alpha_{1}(N-1)(M-P)$ side information equations from other $(N-1)$ databases.
7) Exploiting side information: Until now, we did not specify how the desired equations are constructed. Since each stage in round $i$ can be categorized using Vandermonde's identity as in the previous section, we form the desired equations as a sum of the desired symbols and the undesired symbols that can be decoded from other databases in the former $(i-1)$ rounds. If the user sums two or more symbols from $\mathcal{P}$, the user downloads one new symbol from one message only and the remaining symbols from $\mathcal{P}$ should be derived from other databases. Thus, in round $(i+1)$, the user mixes one symbol of
$\mathcal{P}$ with the sum of $i$ undesired symbols from round $i$. This should be repeated for all $\binom{P}{1}$ desired symbols. Then, the user mixes each sum of 2 desired symbols with the sum of $(i-1)$ undesired symbols generated in the $(i-1)$ th round. This should be repeated for all the $\binom{P}{2}$ combinations of the desired symbols, and so on.
8) Repeating steps: Repeat steps 4, 5, 6, 7 by setting $i=i+1$ until $i=M-P-1$.
9) Last round: We note that rounds $M-P+1$ to $M-1$ do not generate useful side information. Hence, $\alpha_{M-P+1}=$ $\cdots=\alpha_{M-1}=0$. In round $M$, which corresponds to summing all $M$ messages, the user mixes $P$ symbols from $\mathcal{P}$ (only one of them is new and the remaining are previously decoded from the other $(N-1)$ databases) and $M-P$ undesired symbol mixture that was generated in round $(M-P)$.
10) Shuffling the order of queries: After preparing the query table, the order of the queries are shuffled uniformly, so that all possible orders of queries are equally likely regardless of $\mathcal{P}$.

## D. Decodability, Privacy, and Calculation of the Achievable Rate

Now, we verify that the proposed scheme satisfies the reliability and privacy constraints.

For the reliability: The scheme is designed to download the exact number of undesired equations that will be used as side information equation at subsequent rounds in other databases. ${ }^{6}$ Hence, each desired symbol at any round is mixed with a known mixture of symbols that can be decoded from other databases. Note that if the scheme encounters the case of having a mixture of desired symbols, one of them only is chosen to be new and the remaining symbols are downloaded previously from other databases. Thus, the reliability constraint is satisfied by canceling out the side information.

For the privacy: The randomized mapping of message bits and the randomization of the order of queries guarantees privacy as in [8]. It can be checked that when we fix the queries for one database, we can adjust the queries for the remaining databases such that the user can decode any $\mathcal{P}$ subset of messages. This is true since all combinations of messages are generated by our scheme.

To calculate the achievable rate: From Vandermonde's identity $\binom{M}{i}=\sum_{p=0}^{P}\binom{P}{p}\binom{M-P}{i-p}$, round $i$ requires downloading $\binom{P}{p}$ stages in round $(i-p)$. These stages should be downloaded from the remaining $(N-1)$ databases. Hence, as shown in the previous section, the number of stages at each round is calculated as the output of an IIR filter whose input-output relation is given in (56) with the initial conditions $y[-P]=$ $(N-1)^{M-P}, y[-P+1]=\cdots=y[-1]=0$, with the conversion of time index of the filter to the round index of the schemes as $\alpha_{i}=y[(M-P)-i]$. These initial conditions imply

[^5]that the user downloads $(N-1)^{M-P}$ stages in the last round that corresponds to downloading the sum of all messages. The $(P-1)$ rounds before the last round are suppressed because we only need to form sums of $(M-P)$ messages to be used in the last round.

Now, to calculate the number of stages for round $i$, we first solve for the roots of the characteristic equation of (56) [24],

$$
\begin{equation*}
r^{P}-\frac{1}{N-1} \sum_{i=1}^{P}\binom{P}{i} r^{P-i}=0 \tag{59}
\end{equation*}
$$

which is equivalent to

$$
\begin{equation*}
r^{P}-\frac{r^{P}}{N-1} \sum_{i=1}^{P}\binom{P}{i} r^{-i}=0 \tag{60}
\end{equation*}
$$

which further reduces to

$$
\begin{equation*}
r^{P}-\frac{r^{P}}{N-1}\left[\left(1+\frac{1}{r}\right)^{P}-1\right]=0 \tag{61}
\end{equation*}
$$

using the binomial theorem. Simplifying (61), we have

$$
\begin{equation*}
N r^{P}-(r+1)^{P}=0 \tag{62}
\end{equation*}
$$

By applying the bijective mapping $t=N^{1 / P} \cdot \frac{r}{r+1}$, (62) is equivalent to $t^{P}=1$. The roots for this equation are the normal roots of unity, i.e., $t_{k}=e^{j 2 \pi(k-1) / P}, k=1, \cdots, P$, where $j=\sqrt{-1}$. Hence, the roots of the characteristic equation are given by,

$$
\begin{equation*}
r_{k}=\frac{t_{k}}{N^{1 / P}-t_{k}}=\frac{e^{j 2 \pi(k-1) / P}}{N^{1 / P}-e^{j 2 \pi(k-1) / P}}, \quad k=1, \cdots, P \tag{63}
\end{equation*}
$$

Thus, the complete response of the IIR filter is given by $y[n]=$ $\sum_{i=1}^{P} \gamma_{i} r_{i}^{n}$, where $\gamma_{i}$ are constants that result from solving the initial conditions, i.e., $\boldsymbol{\gamma}=\left(\gamma_{1}, \cdots, \gamma_{P}\right)^{T}$ is the solution of the system of equations,

$$
\left[\begin{array}{cccc}
r_{1}^{-P} & r_{2}^{-P} & \cdots & r_{P}^{-P}  \tag{64}\\
r_{1}^{-P+1} & r_{2}^{-P+1} & \cdots & r_{P}^{-P+1} \\
\vdots & \vdots & \cdots & \vdots \\
r_{1}^{-1} & r_{2}^{-1} & \cdots & r_{P}^{-1}
\end{array}\right]\left[\begin{array}{c}
\gamma_{1} \\
\gamma_{2} \\
\vdots \\
\gamma_{P}
\end{array}\right]=\left[\begin{array}{c}
(N-1)^{M-P} \\
0 \\
\vdots \\
0
\end{array}\right]
$$

Now, we are ready to calculate the number of stages $\alpha_{k}$ in round $k$. Since $\alpha_{k}=y[(M-P)-k]$ by construction, then

$$
\begin{equation*}
\alpha_{k}=\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P-k} \tag{65}
\end{equation*}
$$

In round $k$, the user downloads sums of $k$ symbols. The user repeats this round for $\alpha_{k}$ stages. Each stage contains all the combinations of any $k$ symbols which there are $\binom{M}{k}$ of them. Hence, the total download cost $D$ is,

$$
\begin{align*}
D & =\sum_{k=1}^{M}\binom{M}{k} \alpha_{k}  \tag{66}\\
& =\sum_{k=1}^{M} \sum_{i=1}^{P}\binom{M}{k} \gamma_{i} r_{i}^{M-P-k} \tag{67}
\end{align*}
$$

$$
\begin{align*}
& =\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P} \sum_{k=1}^{M}\binom{M}{k} r_{i}^{-k}  \tag{68}\\
& =\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[\left(1+\frac{1}{r_{i}}\right)^{M}-1\right] \tag{69}
\end{align*}
$$

Considering the undesired equations: in round $k$, the user downloads all combinations of the $(M-P)$ undesired messages which there are $\binom{M-P}{k}$ of them. Therefore, similar to the above calculation, the total number of undesired equations $U$ is,

$$
\begin{equation*}
U=\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[\left(1+\frac{1}{r_{i}}\right)^{M-P}-1\right] \tag{70}
\end{equation*}
$$

Hence, the achievable rate $\underline{R}_{S}$ is

$$
\begin{align*}
\underline{R}_{s} & =\frac{D-U}{D}  \tag{71}\\
& =\frac{\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[\left(1+\frac{1}{r_{i}}\right)^{M}-\left(1+\frac{1}{r_{i}}\right)^{M-P}\right]}{\sum_{i=1}^{P} \gamma_{i} r_{i}^{M-P}\left[\left(1+\frac{1}{r_{i}}\right)^{M}-1\right]} \tag{72}
\end{align*}
$$

which is (31) in Theorem 2.

## E. Further Examples for the Case $P \leq \frac{M}{2}$

In this section, we illustrate our proposed scheme with a few additional basic examples. In Section V-A, we considered the case $M=5, P=2, N=2$. In the next three subsections, we consider three more examples. In the example in Section V-E.1, the ratio $\frac{M}{P}$ is exactly equal to 2 , thus, both the achievable scheme here and the achievable scheme in Section IV can be used; we comment about the differences and advantages of both schemes. In the example in Section V-E.2, we present the case of a larger $N$ for the example in Section V-A. In the example in Section V-E.3, we present a case with larger $M, P$ and $N$.

1) $M=4$ Messages, $P=2$ Messages, $N=2$ Databases: The first step of the achievable scheme is to identify the number of stages needed for each round of download. The IIR filter in (56) that determines the number of stages reduces in this case to

$$
\begin{equation*}
y[n]=2 y[n-1]+y[n-2] \tag{73}
\end{equation*}
$$

with the initial conditions $y[-2]=1, y[-1]=0$. The number of stages in round $k$ is $\alpha_{k}=y[2-k]$. Since $M$ is small, we can calculate the output iteratively without using the canonical filter output as,

$$
\begin{align*}
& \alpha_{4}=y[-2]=1  \tag{74}\\
& \alpha_{3}=y[-1]=0  \tag{75}\\
& \alpha_{2}=y[0]=2 y[-1]+y[-2]=1  \tag{76}\\
& \alpha_{1}=y[1]=2 y[0]+y[-1]=2 \tag{77}
\end{align*}
$$

Hence, we should download 2 stages of individual symbols (round 1), and 1 stage of sums of two symbols (round 2). We should suppress the round that retrieves sums of three

TABLE V
The Query Table for the Case $M=4, P=2, N=2$

|  |  | Database 1 | Database 2 |
| :---: | :---: | :---: | :---: |
| $\square$ | $\operatorname{stg} 1$ | $a_{1}, b_{1}, c_{1}, d_{1}$ | $a_{3}, b_{3}, c_{3}, d_{3}$ |
| - | $\operatorname{stg} 2$ | $a_{2}, b_{2}, c_{2}, d_{2}$ | $a_{4}, b_{4}, c_{4}, d_{4}$ |
| $\begin{aligned} & \text { N } \\ & \text { B } \\ & 0 \\ & 0 \end{aligned}$ |  | $a_{5}+b_{3}$ | $a_{1}+b_{7}$ |
|  |  | $a_{6}+c_{3}$ | $a_{8}+c_{1}$ |
|  |  | $a_{7}+d_{3}$ | $a_{9}+d_{1}$ |
|  |  | $b_{5}+c_{4}$ | $b_{8}+c_{2}$ |
|  |  | $b_{6}+d_{4}$ | $b_{9}+d_{2}$ |
|  |  | $c_{5}+d_{5}$ | $c_{6}+d_{6}$ |
| $\stackrel{+}{+}$ | $\operatorname{stg} 1$ | $a_{3}+b_{10}+c_{6}+d_{6}$ | $a_{10}+b_{1}+c_{5}+d_{5}$ |

symbols (round 3), and have 1 stage of sums of all four symbols (round 4).

The user initializes the scheme by randomly and independently interleaving the symbols of each message. The query table for this example is shown in Table V. In round 1, the user downloads individual symbols from all messages at each database. The user downloads $a_{1}, b_{1}, c_{1}, d_{1}$ and $a_{2}, b_{2}, c_{2}, d_{2}$ from database 1 , as $\alpha_{1}=2$. This is repeated for database 2 . In round 2, the user downloads sums of two symbols. There are $\binom{4}{2}=6$ such equations. At database 1 , the undesired symbols from database 2 in the first round are exploited in some of these sums. These equations are either in the form $a+(c, d)$ or in the form $b+(c, d)$. This necessitates two sets of different individual symbols to be downloaded from database 2 in the first round, or otherwise the symbols are repeated and privacy is compromised. Moreover, we note that the user downloads $a_{5}+b_{3}$ which uses $b_{3}$ as side information even though $W_{2}$ is desired; this is reversed in database 2 to download $a_{1}+b_{7}$ with $a_{1}$ as a side information to have a symmetric scheme. Round 2 concludes with downloading $c_{5}+d_{5}$ and $c_{6}+d_{6}$ at the two databases, which will be used as side information in the last round. Round 3 is skipped and the user proceeds to round 4 (last round) directly. In round 4, the user downloads sum of four symbols, and uses the side information downloaded in round 2 and any decoded symbols for the other desired message. For example, in database 1 , the user downloads $a_{3}+b_{10}+c_{6}+d_{6}$, hence, the side information $c_{6}+d_{6}$ is exploited in this round as well as $a_{3}$. The user finishes the scheme by shuffling the order of all queries randomly. The user retrieves $a_{1}, \cdots, a_{10}$ and $b_{1}, \cdots, b_{10}$ privately in 30 downloads ( 15 from each database) and achieves a sum rate of $\frac{20}{30}=\frac{2}{3}=\frac{1}{1+\frac{1}{N}}$, which matches the upper bound in Theorem 2. This sum rate outperforms the repetition-based achievable rate which is $\frac{3}{5}$ in (12).

We note that this case can be solved using the achievable scheme presented in Section IV as well since $\frac{M}{P}=2$ in this case. In fact, this is equivalent to the case considered in Section IV-D.2, if the number of databases is reduced from $N=3$ to $N=2$. Starting from Table III in Section IV-D. 2 and removing the downloads from database 3, we obtain the query table which uses MDS-coded queries shown in Table VI

TABLE VI
Alternative Query Table for the Case $M=4, P=2, N=2$

| Database 1 | Database 2 |
| :---: | :---: |
| $a_{1}, b_{1}, c_{1}, d_{1}$ | $a_{2}, b_{2}, c_{2}, d_{2}$ |
| $a_{3}+b_{3}+c_{2}+d_{2}$ | $a_{4}+b_{4}+c_{1}+d_{1}$ |
| $a_{3}+3 b_{3}+2 c_{2}+4 d_{2}$ | $a_{4}+3 b_{4}+2 c_{1}+4 d_{1}$ |

below. Via the scheme in Table VI below, the user retrieves $a_{1}, \cdots, a_{4}$ and $b_{1}, \cdots, b_{4}$ privately in 12 downloads ( 6 from each database), therefore achieving the same optimal sum rate of $\frac{8}{12}=\frac{2}{3}=\frac{1}{1+\frac{1}{N}}$.

We presented this case here even though it could be solved using the scheme in Section IV, in order to give an example where the second achievable scheme achieves the upper bound in Theorem 2 and yields a capacity result since $\frac{M}{P}$ is an integer. Interestingly, we observe that for all cases where $P=\frac{M}{2}$, the two achievable schemes are both optimal. The two schemes present an interesting trade-off between the field size and the upload cost: The first achievable scheme in Section IV requires using an MDS code with field size $q \geq M$ but the number of queries for each database is limited to $M+P$. On the other hand, the second achievable scheme here in Section V does not use any coding and can work with the storage field size, however, the number of queries increase exponentially since the number of stages for each round is related to an unstable IIR filter.
2) $M=5$ Messages, $P=2$ Messages, $N=3$ Databases: In this example, we show an explicit query structure for $N>2$. In this case the corresponding difference equation for the IIR filter is

$$
\begin{equation*}
y[n]=y[n-1]+\frac{1}{2} y[n-2] \tag{78}
\end{equation*}
$$

with the initial conditions $y[-1]=0, y[-2]=(N-1)^{M-P}=8$. Thus, the number of stages in each round are: $\alpha_{1}=6$, $\alpha_{2}=4, \alpha_{3}=4, \alpha_{4}=0, \alpha_{5}=8$. The query table is shown in Tables VII, VIII and IX. This scheme retrieves $a_{1}, \cdots, a_{126}$ and $b_{1}, \cdots, b_{126}$ privately in 354 downloads (177 from each database), therefore, achieving a sum rate of $\frac{252}{354}=\frac{42}{59}<\frac{1}{1+\frac{1}{N}+\frac{1}{2 N^{2}}}=\frac{18}{25}$. The gap is $\frac{12}{1475} \simeq 0.0081$.
3) $M=7$ Messages, $P=3$ Messages, $N=3$ Databases: Finally, in this section, we consider an example with $N=3$ databases and larger $M$ and $P$ than in previous examples, where we describe the structure and the calculation of the number of queries without specifying the explicit query table as it grows quite large. We first calculate the number of stages at each round. The corresponding IIR filter is

$$
\begin{equation*}
y[n]=\frac{1}{2}(3 y[n-1]+3 y[n-2]+y[n-3]) \tag{79}
\end{equation*}
$$

with the initial conditions $y[-3]=(N-1)^{M-P}=16$, $y[-2]=0, y[-1]=0$. Hence, the number of stages for each round $\alpha_{k}=y[4-k], k=1, \cdots, 7$, are calculated iteratively as $\alpha_{1}=67, \alpha_{2}=30, \alpha_{3}=12, \alpha_{4}=8, \alpha_{5}=0, \alpha_{6}=0$, $\alpha_{7}=16$.

In round 1, the user downloads 67 individual symbols from each message and from each database. Each database can use the side information generated by the other two databases. Hence, each database has $67 \cdot 2=134$ side information equations in the form of single symbols from round 1 to exploit. In round 2, the user downloads sums of two symbols. Each stage in round 2 requires 3 stages from round 1, since the user faces with $a+(d, e, f, g), b+(d, e, f, g)$ or $c+(d, e, f, g)$ cases. Then, round 2 requires $30 \cdot 3=90$ stages from the generated side information in round 1 , and we are left with $134-90=44$ more stages of round 1 . Each database can use the side information stages from the other two databases, i.e., each can use up to $2 \cdot 30=60$ stages of side information in the form of sums of two.

In round 3, the user downloads sums of three symbols, which can be either of $a+b+(d, e, f, g), a+c+(d, e, f, g)$, $b+c+(d, e, f, g), a+(d+e, d+f, \cdots)$, and similarly for $b, c$. Therefore, each stage in round 3 requires 3 stages from round 2, and 3 stages from round 1 . This in total requires $12 \cdot 3=36$ stages from round 1 and 36 stages from round 2, and we will be left with 8 stages from round 1 and 24 stages from round 2. Round 3 generates $2 \cdot 12=24$ stages of side information in the form of sums of threes. In round 4, the user downloads sums of 4 symbols, which can be either $a+b+(d+e, d+f, \cdots)$, and similarly for $b+c$ and $a+c, a+(d+e+f, d+e+g, \cdots)$ and similarly for $b$, $c$, or $a+b+c+(d, e, f, g)$. This means that for each stage of round 3 , the user needs 1 stage of round 1,3 stages of round 2 , and 3 stages of round 3 . This in total requires $8 \cdot 3=24$ stages from round 2 and 3 and $8 \cdot 1$ stages from round 1 and hence, we exhaust all the generated side information by round 4. Round 4 generates 8 stages of side information in the form of sums of fours. This will be used in the last round to get $8 \cdot 2$ new symbols from the desired messages.

The achievable sum rate in this case is $\frac{3933}{5445}=\frac{437}{605}<$ $\frac{1}{1+\frac{1}{N}+\frac{1}{3 N^{2}}}=\frac{27}{37}$. The gap is $\frac{166}{22385} \simeq 0.0074$.

## VI. Converse Proof

In this section, we derive an upper bound for the MPIR problem. ${ }^{7}$ The derived upper bound is tight when $P \geq \frac{M}{2}$ and when $\frac{M}{P} \in \mathbb{N}$. We follow the notations and simplifications in [8] and [12], and we define

$$
\begin{equation*}
\mathcal{Q} \triangleq\left\{Q_{n}^{[\mathcal{P}]}: \mathcal{P} \subseteq\{1, \cdots, M\},|\mathcal{P}|=P, n \in\{1, \cdots, N\}\right\} \tag{80}
\end{equation*}
$$

and

$$
\begin{equation*}
A_{n_{1}: n_{2}}^{[\mathcal{P}]} \triangleq\left\{A_{n_{1}}^{[\mathcal{P}]}, A_{n_{1}+1}^{[\mathcal{P}]}, \cdots, A_{n_{2}}^{[\mathcal{P}]}\right\} \tag{81}
\end{equation*}
$$

for $n_{1} \leq n_{2}, n_{1}, n_{2} \in\{1, \cdots, N\}$.
Without loss of generality, the following simplifications hold for the MPIR problem:

1) We can assume that the MPIR scheme is symmetric. Since for every asymmetric scheme, there exists an
[^6]TABLE VII
The Query Table for the Case $M=5, P=2, N=3$

|  |  | Database 1 | Database 2 | Database 3 |
| :---: | :---: | :---: | :---: | :---: |
| ت | stg 1 | $a_{1}, b_{1}, c_{1}, d_{1}, e_{1}$ | $a_{7}, b_{7}, c_{7}, d_{7}, e_{7}$ | $a_{13}, b_{13}, c_{13}, d_{13}, e_{13}$ |
|  | stg 2 | $a_{2}, b_{2}, c_{2}, d_{2}, e_{2}$ | $a_{8}, b_{8}, c_{8}, d_{8}, e_{8}$ | $a_{14}, b_{14}, c_{14}, d_{14}, e_{14}$ |
|  | stg 3 | $a_{3}, b_{3}, c_{3}, d_{3}, e_{3}$ | $a_{9}, b_{9}, c_{9}, d_{9}, e_{9}$ | $a_{15}, b_{15}, c_{15}, d_{15}, e_{15}$ |
|  | stg 4 | $a_{4}, b_{4}, c_{4}, d_{4}, e_{4}$ | $a_{10}, b_{10}, c_{10}, d_{10}, e_{10}$ | $a_{16}, b_{16}, c_{16}, d_{16}, e_{16}$ |
|  | $\operatorname{stg} 5$ | $a_{5}, b_{5}, c_{5}, d_{5}, e_{5}$ | $a_{11}, b_{11}, c_{11}, d_{11}, e_{11}$ | $a_{17}, b_{17}, c_{17}, d_{17}, e_{17}$ |
|  | stg 6 | $a_{6}, b_{6}, c_{6}, d_{6}, e_{6}$ | $a_{12}, b_{12}, c_{12}, d_{12}, e_{12}$ | $a_{18}, b_{18}, c_{18}, d_{18}, e_{18}$ |
| $\begin{aligned} & \text { N } \\ & \stackrel{\rightharpoonup}{3} \\ & 0 \\ & 0 \end{aligned}$ | $\begin{aligned} & \underset{\sim}{0} \\ & \stackrel{0}{0} \\ & \underset{\sim}{0} \end{aligned}$ | $\begin{gathered} a_{19}+b_{7} \\ a_{20}+c_{7} \\ a_{21}+d_{7} \\ a_{22}+e_{7} \\ b_{19}+c_{8} \\ b_{20}+d_{8} \\ b_{21}+e_{8} \\ c_{19}+d_{19} \\ c_{20}+e_{19} \\ d_{20}+e_{20} \\ \hline \end{gathered}$ | $\begin{gathered} a_{33}+b_{1} \\ a_{34}+c_{1} \\ a_{35}+d_{1} \\ a_{36}+e_{1} \\ b_{33}+c_{2} \\ b_{34}+d_{2} \\ b_{35}+e_{2} \\ c_{27}+d_{27} \\ c_{28}+e_{27} \\ d_{28}+e_{28} \\ \hline \end{gathered}$ | $\begin{gathered} \hline a_{47}+b_{1} \\ a_{48}+c_{1} \\ a_{49}+d_{1} \\ a_{50}+e_{1} \\ b_{47}+c_{2} \\ b_{48}+d_{2} \\ b_{49}+e_{2} \\ c_{35}+d_{35} \\ c_{36}+e_{35} \\ d_{36}+e_{36} \\ \hline \end{gathered}$ |
|  | $\begin{aligned} & N \\ & 0 \\ & 0 \\ & \text { O } \\ & \text { © } \end{aligned}$ | $\begin{gathered} a_{7}+b_{22} \\ a_{23}+c_{9} \\ a_{24}+d_{9} \\ a_{25}+e_{9} \\ b_{23}+c_{10} \\ b_{24}+d_{10} \\ b_{25}+e_{10} \\ c_{21}+d_{21} \\ c_{22}+e_{21} \\ d_{22}+e_{22} \end{gathered}$ | $\begin{gathered} a_{1}+b_{36} \\ a_{37}+c_{3} \\ a_{38}+d_{3} \\ a_{39}+e_{3} \\ b_{37}+c_{4} \\ b_{38}+d_{4} \\ b_{39}+e_{4} \\ c_{29}+d_{29} \\ c_{30}+e_{29} \\ d_{30}+e_{30} \\ \hline \end{gathered}$ | $\begin{aligned} & a_{1}+b_{50} \\ & a_{51}+c_{3} \\ & a_{52}+d_{3} \\ & a_{53}+e_{3} \\ & b_{51}+c_{4} \\ & b_{52}+d_{4} \\ & b_{53}+e_{4} \\ & c_{37}+d_{37} \\ & c_{38}+e_{37} \\ & d_{38}+e_{38} \end{aligned}$ |
|  | $\begin{aligned} & \text { m } \\ & 0 \\ & \text { © } \\ & \text { 훙 } \end{aligned}$ | $\begin{aligned} & a_{26}+b_{13} \\ & a_{27}+c_{13} \\ & a_{28}+d_{13} \\ & a_{29}+e_{13} \\ & b_{26}+c_{14} \\ & b_{27}+d_{14} \\ & b_{28}+e_{14} \\ & c_{23}+d_{23} \\ & c_{24}+e_{23} \\ & d_{24}+e_{24} \end{aligned}$ | $\begin{aligned} & a_{40}+b_{13} \\ & a_{41}+c_{13} \\ & a_{42}+d_{13} \\ & a_{43}+e_{13} \\ & b_{40}+c_{14} \\ & b_{41}+d_{14} \\ & b_{42}+e_{14} \\ & c_{31}+d_{31} \\ & c_{32}+e_{31} \\ & d_{32}+e_{32} \end{aligned}$ | $\begin{gathered} a_{54}+b_{7} \\ a_{55}+c_{7} \\ a_{56}+d_{7} \\ a_{57}+e_{7} \\ b_{54}+c_{8} \\ b_{55}+d_{8} \\ b_{56}+e_{8} \\ c_{39}+d_{39} \\ c_{40}+e_{39} \\ d_{40}+e_{40} \end{gathered}$ |

equal rate symmetric scheme that can be constructed by replicating all permutations of databases and messages.
2) To invoke the privacy constraint, we fix the response of one database to be the same irrespective of the desired set of messages $\mathcal{P}$, i.e., $A_{n}^{\left[\mathcal{P}_{i}\right]}=A_{n}$, where $\left|\mathcal{P}_{i}\right|=P$ for every $i \in\{1,2, \cdots, \beta\}$ for some $n \in\{1, \cdots, N\}$, and $\beta=\binom{M}{P}$. No loss of generality is incurred due to the fact that the queries and the answers are statistically independent from $\mathcal{P}$. In the sequel, we fix the answer string of the first database, i.e.,

$$
\begin{equation*}
A_{1}^{[\mathcal{P}]}=A_{1}, \forall \mathcal{P} \tag{82}
\end{equation*}
$$

The following lemma is a consequence of the symmetry assumption; its proof can be found in [8].

Lemma 1 (Symmetry [8]): For any $W_{\mathcal{S}}=\left\{W_{i}: i \in \mathcal{S}\right\}$

$$
\begin{align*}
H\left(A_{n}^{[\mathcal{P}]} \mid W_{\mathcal{S}}, \mathcal{Q}\right) & =H\left(A_{1}^{[\mathcal{P}]} \mid W_{\mathcal{S}}, \mathcal{Q}\right), \quad n \in\{1, \cdots, N\}  \tag{83}\\
H\left(A_{1} \mid \mathcal{Q}\right) & =H\left(A_{n}^{[\mathcal{P}]} \mid \mathcal{Q}\right), \quad n \in\{1, \cdots, N\}, \forall \mathcal{P} \tag{84}
\end{align*}
$$

We construct the converse proof by induction over $\left\lfloor\frac{M}{P}\right\rfloor$ in a similar way to [8] and [12]. The base induction step is obtained for $1 \leq \frac{M}{P} \leq 2$ (this is the case $P \geq \frac{M}{2}$ as it was referred to so far, where the user wants to retrieve at least half of the messages). We obtain an inductive relation for the case $\frac{M}{P}>2$. The converse proof extends the proof in [8] for $P>1$.

## A. Converse Proof for the Case $1 \leq \frac{M}{P} \leq 2$

To prove the converse for the case $1 \leq \frac{M}{P} \leq 2$, we need the following lemma which gives a lower bound on the interference within an answer string.

TABLE VIII
The Query Table for the Case $M=5, P=2, N=3$ (Cont.)

|  |  | Database 1 | Database 2 | Database 3 |
| :---: | :---: | :---: | :---: | :---: |
| $\begin{aligned} & N \\ & 0 \\ & 0 \\ & 0 \end{aligned}$ |  | $\begin{aligned} & a_{13}+b_{29} \\ & a_{30}+c_{15} \\ & a_{31}+d_{15} \\ & a_{32}+e_{15} \\ & b_{30}+c_{16} \\ & b_{31}+d_{16} \\ & b_{32}+e_{16} \\ & c_{25}+d_{25} \\ & c_{26}+e_{25} \\ & d_{26}+e_{26} \end{aligned}$ | $\begin{aligned} & a_{13}+b_{43} \\ & a_{44}+c_{15} \\ & a_{45}+d_{15} \\ & a_{46}+e_{15} \\ & b_{44}+c_{16} \\ & b_{45}+d_{16} \\ & b_{46}+e_{16} \\ & c_{33}+d_{33} \\ & c_{34}+e_{33} \\ & d_{34}+e_{34} \end{aligned}$ | $\begin{gathered} a_{7}+b_{57} \\ a_{58}+c_{9} \\ a_{59}+d_{9} \\ a_{60}+e_{9} \\ b_{58}+c_{10} \\ b_{59}+d_{10} \\ b_{60}+e_{10} \\ c_{41}+d_{41} \\ c_{42}+e_{41} \\ d_{42}+e_{42} \end{gathered}$ |
| $\begin{aligned} & \text { m } \\ & 0 \\ & 0 \\ & 0 \end{aligned}$ |  | $\begin{gathered} a_{61}+b_{8}+c_{11} \\ a_{8}+b_{61}+d_{11} \\ a_{62}+b_{9}+e_{11} \\ a_{63}+c_{27}+d_{27} \\ a_{64}+c_{28}+e_{27} \\ a_{65}+d_{28}+e_{28} \\ b_{62}+c_{29}+d_{29} \\ b_{63}+c_{30}+e_{29} \\ b_{64}+d_{30}+e_{30} \\ c_{43}+d_{43}+e_{43} \end{gathered}$ | $\begin{gathered} a_{79}+b_{2}+c_{5} \\ a_{2}+b_{79}+d_{5} \\ a_{80}+b_{3}+e_{5} \\ a_{81}+c_{19}+d_{19} \\ a_{82}+c_{20}+e_{19} \\ a_{83}+d_{20}+e_{20} \\ b_{80}+c_{21}+d_{21} \\ b_{81}+c_{22}+e_{21} \\ b_{82}+d_{22}+e_{22} \\ c_{47}+d_{47}+e_{47} \end{gathered}$ | $\begin{gathered} \hline a_{97}+b_{2}+c_{5} \\ a_{2}+b_{97}+d_{5} \\ a_{98}+b_{3}+e_{5} \\ a_{99}+c_{19}+d_{19} \\ a_{100}+c_{20}+e_{19} \\ a_{101}+d_{20}+e_{20} \\ b_{98}+c_{21}+d_{21} \\ b_{99}+c_{22}+e_{21} \\ b_{100}+d_{22}+e_{22} \\ c_{51}+d_{51}+e_{51} \\ \hline \end{gathered}$ |
|  | $\begin{aligned} & N \\ & 0 \\ & \text { O. } \\ & \text { İ } \end{aligned}$ | $\begin{gathered} a_{9}+b_{65}+c_{12} \\ a_{66}+b_{10}+d_{12} \\ a_{10}+b_{66}+e_{12} \\ a_{67}+c_{31}+d_{31} \\ a_{68}+c_{32}+e_{31} \\ a_{69}+d_{32}+e_{32} \\ b_{67}+c_{33}+d_{33} \\ b_{68}+c_{34}+e_{33} \\ b_{69}+d_{34}+e_{34} \\ c_{44}+d_{44}+e_{44} \end{gathered}$ | $\begin{gathered} a_{3}+b_{83}+c_{6} \\ a_{84}+b_{4}+d_{6} \\ a_{4}+b_{84}+e_{6} \\ a_{85}+c_{23}+d_{23} \\ a_{86}+c_{24}+e_{23} \\ a_{87}+d_{24}+e_{24} \\ b_{85}+c_{25}+d_{25} \\ b_{86}+c_{26}+e_{25} \\ b_{87}+d_{26}+e_{26} \\ c_{48}+d_{48}+e_{48} \end{gathered}$ | $\begin{gathered} a_{3}+b_{101}+c_{6} \\ a_{102}+b_{4}+d_{6} \\ a_{4}+b_{102}+e_{6} \\ a_{103}+c_{23}+d_{23} \\ a_{104}+c_{24}+e_{23} \\ a_{105}+d_{24}+e_{24} \\ b_{103}+c_{25}+d_{25} \\ b_{104}+c_{26}+e_{25} \\ b_{105}+d_{26}+e_{26} \\ c_{52}+d_{52}+e_{52} \end{gathered}$ |
|  | $\begin{aligned} & m \\ & 0 \\ & \text { © } \\ & \text { İ } \end{aligned}$ | $\begin{aligned} & a_{70}+b_{14}+c_{17} \\ & a_{14}+b_{70}+d_{17} \\ & a_{71}+b_{15}+e_{17} \\ & a_{72}+c_{35}+d_{35} \\ & a_{73}+c_{36}+e_{35} \\ & a_{74}+d_{36}+e_{36} \\ & b_{71}+c_{37}+d_{37} \\ & b_{72}+c_{38}+e_{37} \\ & b_{73}+d_{38}+e_{38} \\ & c_{45}+d_{45}+e_{45} \end{aligned}$ | $\begin{aligned} & a_{88}+b_{14}+c_{17} \\ & a_{14}+b_{88}+d_{17} \\ & a_{89}+b_{15}+e_{17} \\ & a_{90}+c_{35}+d_{35} \\ & a_{91}+c_{36}+e_{35} \\ & a_{92}+d_{36}+e_{36} \\ & b_{89}+c_{37}+d_{37} \\ & b_{90}+c_{38}+e_{37} \\ & b_{91}+d_{38}+e_{38} \\ & c_{49}+d_{49}+e_{49} \end{aligned}$ | $\begin{gathered} a_{106}+b_{8}+c_{11} \\ a_{8}+b_{106}+d_{11} \\ a_{107}+b_{9}+e_{11} \\ a_{108}+c_{27}+d_{27} \\ a_{109}+c_{28}+e_{27} \\ a_{110}+d_{28}+e_{28} \\ b_{107}+c_{29}+d_{29} \\ b_{108}+c_{30}+e_{29} \\ b_{109}+d_{30}+e_{30} \\ c_{53}+d_{53}+e_{53} \\ \hline \end{gathered}$ |

Lemma 2 (Interference Lower Bound): For the MPIR problem with $P \geq \frac{M}{2}$, the uncertainty of the interfering messages $W_{P+1: M}$ within the answer string $A_{1}^{[1: P]}$ is lower bounded as,

$$
\begin{equation*}
H\left(A_{1}^{[1: P]} \mid W_{1: P}, \mathcal{Q}\right) \geq \frac{(M-P) L}{N} \tag{85}
\end{equation*}
$$

Furthermore, (85) is true for any set of desired messages $\mathcal{P}$ with $|\mathcal{P}|=P$, i.e.,

$$
\begin{equation*}
H\left(A_{1}^{[\mathcal{P}]} \mid W_{\mathcal{P}}, \mathcal{Q}\right) \geq \frac{(M-P) L}{N} \tag{86}
\end{equation*}
$$

Proof: For clarity of presentation, we assume that $\mathcal{P}=$ $\{1, \cdots, P\}$ without loss of generality. Hence,

$$
\begin{align*}
(M-P) L= & H\left(W_{P+1: M}\right)  \tag{87}\\
= & H\left(W_{P+1: M} \mid W_{1: P}, \mathcal{Q}\right)  \tag{88}\\
= & H\left(W_{P+1: M} \mid W_{1: P}, \mathcal{Q}\right) \\
& \quad-H\left(W_{P+1: M} \mid A_{1: N}^{[M-P+1: M]}, W_{1: P}, \mathcal{Q}\right)  \tag{89}\\
= & I\left(W_{P+1: M} ; A_{1: N}^{[M-P+1: M]} \mid W_{1: P}, \mathcal{Q}\right)  \tag{90}\\
= & H\left(A_{1: N}^{[M-P+1: M]} \mid W_{1: P}, \mathcal{Q}\right)  \tag{91}\\
\leq & \sum_{n=1}^{N} H\left(A_{n}^{[M-P+1: M]} \mid W_{1: P}, \mathcal{Q}\right)  \tag{92}\\
= & N H\left(A_{1} \mid W_{1: P}, \mathcal{Q}\right) \tag{93}
\end{align*}
$$

TABLE IX
The Query Table for the case $M=5, P=2, N=3$ (Cont.)

|  |  | Database 1 | Database 2 | Database 3 |
| :---: | :---: | :---: | :---: | :---: |
| $\begin{aligned} & \text { m } \\ & 0 \\ & 0 \\ & 0 \end{aligned}$ |  | $\begin{aligned} & a_{15}+b_{74}+c_{18} \\ & a_{75}+b_{16}+d_{18} \\ & a_{16}+b_{75}+e_{18} \\ & a_{76}+c_{39}+d_{39} \\ & a_{77}+c_{40}+e_{39} \\ & a_{78}+d_{40}+e_{40} \\ & b_{76}+c_{41}+d_{41} \\ & b_{77}+c_{42}+e_{41} \\ & b_{78}+d_{42}+e_{42} \\ & c_{46}+d_{46}+e_{46} \end{aligned}$ | $\begin{aligned} & a_{15}+b_{92}+c_{18} \\ & a_{93}+b_{16}+d_{18} \\ & a_{16}+b_{93}+e_{18} \\ & a_{94}+c_{39}+d_{39} \\ & a_{95}+c_{40}+e_{39} \\ & a_{96}+d_{40}+e_{40} \\ & b_{94}+c_{41}+d_{41} \\ & b_{95}+c_{42}+e_{41} \\ & b_{96}+d_{42}+e_{42} \\ & c_{50}+d_{50}+e_{50} \end{aligned}$ | $\begin{gathered} a_{9}+b_{110}+c_{12} \\ a_{111}+b_{10}+d_{12} \\ a_{10}+b_{111}+e_{12} \\ a_{112}+c_{31}+d_{31} \\ a_{113}+c_{32}+e_{31} \\ a_{114}+d_{32}+e_{32} \\ b_{112}+c_{33}+d_{33} \\ b_{113}+c_{34}+e_{33} \\ b_{114}+d_{34}+e_{34} \\ c_{54}+d_{54}+e_{54} \\ \hline \end{gathered}$ |
| $\begin{aligned} & n \\ & \text { n } \\ & 0 \\ & 0 \end{aligned}$ | stg 1 | $a_{115}+b_{11}+c_{47}+d_{47}+e_{47}$ | $a_{119}+b_{5}+c_{43}+d_{43}+e_{43}$ | $a_{123}+b_{5}+c_{43}+d_{43}+e_{43}$ |
|  | $\operatorname{stg} 2$ | $a_{11}+b_{115}+c_{48}+d_{48}+e_{48}$ | $a_{5}+b_{119}+c_{44}+d_{44}+e_{44}$ | $a_{5}+b_{123}+c_{44}+d_{44}+e_{44}$ |
|  | $\operatorname{stg} 3$ | $a_{116}+b_{12}+c_{49}+d_{49}+e_{49}$ | $a_{120}+b_{6}+c_{45}+d_{45}+e_{45}$ | $a_{124}+b_{6}+c_{45}+d_{45}+e_{45}$ |
|  | $\operatorname{stg} 4$ | $a_{12}+b_{116}+c_{50}+d_{50}+e_{50}$ | $a_{6}+b_{120}+c_{46}+d_{46}+e_{46}$ | $a_{6}+b_{124}+c_{46}+d_{46}+e_{46}$ |
|  | $\operatorname{stg} 5$ | $a_{117}+b_{17}+c_{51}+d_{51}+e_{51}$ | $a_{121}+b_{17}+c_{51}+d_{51}+e_{51}$ | $a_{125}+b_{11}+c_{47}+d_{47}+e_{47}$ |
|  | stg 6 | $a_{17}+b_{117}+c_{52}+d_{52}+e_{52}$ | $a_{17}+b_{121}+c_{52}+d_{52}+e_{52}$ | $a_{11}+b_{125}+c_{48}+d_{48}+e_{48}$ |
|  | $\operatorname{stg} 7$ | $a_{118}+b_{18}+c_{53}+d_{53}+e_{53}$ | $a_{122}+b_{18}+c_{53}+d_{53}+e_{53}$ | $a_{126}+b_{12}+c_{49}+d_{49}+e_{49}$ |
|  | $\operatorname{stg} 8$ | $a_{18}+b_{118}+c_{54}+d_{54}+e_{54}$ | $a_{18}+b_{122}+c_{54}+d_{54}+e_{54}$ | $a_{12}+b_{126}+c_{50}+d_{50}+e_{50}$ |

where (88) follows from the independence of the messages $W_{P+1: M}$ from the messages $W_{1: P}$ and the queries as in (2) and (3); (89) follows from the reliability constraint (7), since messages $W_{P+1: M}$ can be decoded correctly from the answer strings $A_{1: N}^{[M-P+1: M]}$ if $P \geq \frac{M}{2}$ as $\{P+1, \cdots, M\} \subseteq\{M-P+$ $1, \cdots, M\}$ in this regime; (91) follows from the fact that the answer strings are deterministic functions of all messages and queries $\left(\mathcal{Q}, W_{1: M}\right)$; and (93) follows from the independence bound and Lemma 1.

Consequently, $H\left(A_{1} \mid W_{1: P}, \mathcal{Q}\right) \geq \frac{(M-P) L}{N}$. The proof of the general statement can be done replacing $W_{1: P}$ by $W_{\mathcal{P}}, W_{P+1: M}$ by $W_{\overline{\mathcal{P}}}$ which corresponds to the complement set of messages of $W_{\mathcal{P}}$, and the answer strings $A_{1: N}^{[M-P+1: M]}$ by $A_{1: N}^{\left[\mathcal{P}^{*}\right]}$, where $\overline{\mathcal{P}} \subseteq \mathcal{P}^{*},\left|\mathcal{P}^{*}\right|=P$.

Now, we are ready to prove the converse of the case $P \geq \frac{M}{2}$. We use a similar converse technique to the case of $M=2, P=1$ in [8],
$M L$

$$
\begin{align*}
& =H\left(W_{1: M}\right)  \tag{94}\\
& =H\left(W_{1: M} \mid \mathcal{Q}\right)  \tag{95}\\
& =H\left(W_{1: M} \mid \mathcal{Q}\right)-H\left(W_{1: M} \mid A_{1: N}^{\left[\mathcal{P}_{1}\right]}, A_{1: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]}, \mathcal{Q}\right)  \tag{96}\\
& =I\left(W_{1: M} ; A_{1: N}^{\left[\mathcal{P}_{1}\right]}, A_{1: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]} \mid \mathcal{Q}\right)  \tag{97}\\
& =H\left(A_{1: N}^{\left[\mathcal{P}_{1}\right]}, A_{1: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]} \mid \mathcal{Q}\right)  \tag{98}\\
& =H\left(A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, A_{2: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid \mathcal{Q}\right)  \tag{99}\\
& =H\left(A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right)+H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, \mathcal{Q}\right)
\end{align*}
$$

$$
\begin{equation*}
=H\left(A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right) \tag{100}
\end{equation*}
$$

$$
\begin{equation*}
+H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, W_{\mathcal{P}_{1}}, \mathcal{Q}\right) \tag{101}
\end{equation*}
$$

$$
\begin{align*}
\leq & \sum_{n=1}^{N} H\left(A_{n}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right)+H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid A_{1}, W_{\mathcal{P}_{1}}, \mathcal{Q}\right)  \tag{102}\\
= & \sum_{n=1}^{N} H\left(A_{n}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right)+H\left(A_{1: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]} \mid W_{\mathcal{P}_{1}}, \mathcal{Q}\right) \\
& \quad-H\left(A_{1} \mid W_{\mathcal{P}_{1}}, \mathcal{Q}\right) \tag{103}
\end{align*}
$$

where $\beta=\binom{M}{P}$ represents the total number of message subsets of size $P$ that can be constructed from $M$ messages; (95) follows from the independence between the messages and the queries; (96) follows from the reliability constraint in (7) with noting that $A_{1: N}^{\left[\mathcal{P}_{1}\right]}, A_{1: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]}$ represent all answer strings from all databases to every possible subset of messages $\mathcal{P}_{i} \subseteq\{1, \cdots, M\}, i=1,2, \cdots, \beta$, hence all messages can be correctly decoded as all possible answer strings are known; (98) follows from the fact that answer strings are deterministic functions of the messages and the queries; (99) follows from simplification (82) without loss of generality; (101) follows from the fact that the messages $W_{\mathcal{P}}=\left(W_{i_{1}}, W_{i_{2}}, \cdots, W_{i_{P}}\right)$ can be reconstructed from $A_{1: N}^{[\mathcal{P}]}$; and (102) is a consequence of the fact that conditioning does not increase entropy and Lemma 1.

Now, every message appears in $\binom{M-1}{P-1}$ different message subsets of size $P$, therefore the answer strings $\left(A_{1: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]}\right)$ are sufficient to construct all messages $W_{1: M}$ irrespective of $\mathcal{P}_{1}$. Therefore,

$$
\begin{equation*}
H\left(A_{1: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]} \mid W_{\mathcal{P}_{1}}, \mathcal{Q}\right)=(M-P) L \tag{104}
\end{equation*}
$$

Using this and Lemma 2 in (103) yields

$$
\begin{equation*}
M L \leq \sum_{n=1}^{N} H\left(A_{n}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right)+(M-P) L-\frac{(M-P) L}{N} \tag{105}
\end{equation*}
$$

which can be written as,

$$
\begin{equation*}
P L+\frac{(M-P) L}{N} \leq \sum_{n=1}^{N} H\left(A_{n}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right) \tag{106}
\end{equation*}
$$

which further can be written as,

$$
\begin{equation*}
\left(1+\frac{M-P}{P N}\right) P L \leq \sum_{n=1}^{N} H\left(A_{n}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right) \tag{107}
\end{equation*}
$$

which leads to the desired converse result,

$$
\begin{align*}
\sum_{i=1}^{P} R_{i} & =\frac{P L}{\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}]}\right)} \\
& \leq \frac{P L}{\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}]} \mid \mathcal{Q}\right)} \\
& \leq \frac{1}{1+\frac{M-P}{P N}} \tag{108}
\end{align*}
$$

## B. Converse Proof for the Case $\frac{M}{P}>2$

In the sequel, we derive an inductive relation that can be used in addition to the base induction step of $1 \leq \frac{M}{P} \leq 2$ derived in the previous sub-section to obtain an upper bound for the MPIR problem. The idea we pursue here is similar in spirit to the one in [8], where the authors developed a base converse step for $M=2$ messages, and developed an induction over the number of messages $M$ for the case $M>2$. Here, we have developed a base converse step for $1 \leq \frac{M}{P} \leq 2$, and now develop an induction over $\left\lfloor\frac{M}{P}\right\rfloor$ for the case $\frac{M}{P}>2$.

The following lemma upper bounds the remaining uncertainty of the answer strings after knowing a subset of size $P$ of the interference messages.

Lemma 3 (Interference Conditioning Lemma): The remaining uncertainty in the answer strings $A_{2: N}^{\left[\mathcal{P}_{2}\right]}$ after conditioning on the messages indexed by $\mathcal{P}_{1}$, such that $\mathcal{P}_{1} \cap \mathcal{P}_{2}=\phi$, $\left|\mathcal{P}_{1}\right|=\left|\mathcal{P}_{2}\right|=P$ is upper bounded by,

$$
\begin{equation*}
H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]} \mid W_{\mathcal{P}_{1}}, \mathcal{Q}\right) \leq(N-1)\left[N H\left(A_{1} \mid \mathcal{Q}\right)-P L\right] \tag{109}
\end{equation*}
$$

Proof: We begin with

$$
\begin{align*}
H & \left(A_{2: N}^{\left[\mathcal{P}_{2}\right]} \mid W_{\mathcal{P}_{1}}, \mathcal{Q}\right) \\
& \leq \sum_{n=2}^{N} H\left(A_{n}^{\left[\mathcal{P}_{2}\right]} \mid W_{\mathcal{P}_{1}}, \mathcal{Q}\right)  \tag{110}\\
& \leq \sum_{n=2}^{N} H\left(A_{1: n-1}^{\left[\mathcal{P}_{1}\right]}, A_{n}^{\left[\mathcal{P}_{2}\right]}, A_{n+1: N}^{\left[\mathcal{P}_{1}\right]} \mid W_{\mathcal{P}_{1}}, \mathcal{Q}\right)  \tag{111}\\
& =\sum_{n=2}^{N} H\left(A_{1: n-1}^{\left[\mathcal{P}_{1}\right]}, A_{n}^{\left[\mathcal{P}_{2}\right]}, A_{n+1: N}^{\left[\mathcal{P}_{1}\right]}, W_{\mathcal{P}_{1}} \mid \mathcal{Q}\right)-H\left(W_{\mathcal{P}_{1}} \mid \mathcal{Q}\right) \\
& =\sum_{n=2}^{N} H\left(A_{1: n-1}^{\left[\mathcal{P}_{1}\right]}, A_{n}^{\left[\mathcal{P}_{2}\right]}, A_{n+1: N}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right)  \tag{112}\\
\quad \leq & \sum_{n=2}^{N} N\left(W_{\mathcal{P}_{1}} \mid A_{1: n-1}^{\left[\mathcal{P}_{1}\right]}, A_{n}^{\left[\mathcal{P}_{2}\right]}, A_{n+1: N}^{\left[\mathcal{P}_{1}\right]}\right)-H\left(W_{\mathcal{P}_{1}}\right)  \tag{113}\\
= & (N-1)\left[N H\left(A_{1} \mid \mathcal{Q}\right)-H\left(W_{\mathcal{P}_{1}}\right)\right. \tag{114}
\end{align*}
$$

where (110) follows from the independence bound; (111) follows from the non-negativity of entropy; (113) follows from the statistical independence between the messages and the queries; and (114) follows from the decodability of $W_{\mathcal{P}_{1}}$ given the answer strings $\left(A_{1: n-1}^{\left[\mathcal{P}_{1}\right]}, A_{n}^{\left[\mathcal{P}_{2}\right]}\right.$, $\left.A_{n+1: N}^{\left[\mathcal{P}_{1}\right]}\right)$, which is tantamount to the privacy constraint as in the second simplification.

Now, we derive the inductive relation for $\frac{M}{P}>2$. Without loss of generality, let $\mathcal{P}_{1}=\{1, \cdots, P\}$ and $\mathcal{P}_{2}=\{P+$ $1, \cdots, 2 P\}$. Then, starting from (99), we write

$$
\begin{align*}
M L= & H\left(A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, A_{2: N}^{\left[\mathcal{P}_{2}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid \mathcal{Q}\right)  \tag{116}\\
= & H\left(A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]} \mid \mathcal{Q}\right)+H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]} \mid A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, \mathcal{Q}\right) \\
& +H\left(A_{2: N}^{\left[\mathcal{P}_{3}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, A_{2: N}^{\left[\mathcal{P}_{2}\right]}, \mathcal{Q}\right)  \tag{117}\\
\leq & N H\left(A_{1} \mid \mathcal{Q}\right)+H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]} \mid A_{1}, A_{2: N}^{\left[\mathcal{P}_{2}\right]}, W_{1: P}, \mathcal{Q}\right) \\
& +H\left(A_{2: N}^{\left[\mathcal{P}_{3}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, A_{2: N}^{\left[\mathcal{P}_{2}\right]}, W_{1: 2 P}, \mathcal{Q}\right) \\
\leq & N H\left(A_{1} \mid \mathcal{Q}\right)+H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]} \mid W_{1: P}, \mathcal{Q}\right)  \tag{118}\\
& +H\left(A_{2: N}^{\left[\mathcal{P}_{3}\right]}, \cdots, A_{2: N}^{\left[\mathcal{P}_{\beta}\right]} \mid A_{1}, W_{1: 2 P}, \mathcal{Q}\right)  \tag{119}\\
= & N H\left(A_{1} \mid \mathcal{Q}\right)+H\left(A_{2: N}^{\left[\mathcal{P}_{2}\right]} \mid W_{1: P}, \mathcal{Q}\right)-H\left(A_{1} \mid W_{1: 2 P}, \mathcal{Q}\right) \\
& +H\left(A_{1: N}^{\left[\mathcal{P}_{3}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]} \mid W_{1: 2 P}, \mathcal{Q}\right)  \tag{120}\\
= & N H\left(A_{1} \mid \mathcal{Q}\right)+H\left(A_{2: N}^{\left[\mathcal{P}_{1}\right]} \mid W_{1: P}, \mathcal{Q}\right)-H\left(A_{1} \mid W_{1: 2 P}, \mathcal{Q}\right) \\
& +(M-2 P) L  \tag{121}\\
\leq & N H\left(A_{1} \mid \mathcal{Q}\right)+(N-1)\left[N H\left(A_{1} \mid \mathcal{Q}\right)-P L\right] \\
& -H\left(A_{1} \mid W_{1: 2 P}, \mathcal{Q}\right)+(M-2 P) L \tag{122}
\end{align*}
$$

where (118) follows from the decodability of $W_{1: 2 P}$ given $\left(A_{1}, A_{2: N}^{\left[\mathcal{P}_{1}\right]}, A_{2: N}^{\left[\mathcal{P}_{2}\right]}\right)$, the symmetry lemma and the independence bound; (119) follows from the fact that conditioning does not increase entropy. In (121), we note that subsets $\left(\mathcal{P}_{3}, \cdots, \mathcal{P}_{\beta}\right)$ include all messages $\left(W_{1}, \cdots, W_{M}\right)$ because every message appears in $\binom{M-1}{P-1}$ subsets. Hence, $H\left(A_{1: N}^{\left[\mathcal{P}_{3}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]} \mid\right.$ $\left.W_{1: 2 P}, \mathcal{Q}\right)=(M-2 P) L$ since $W_{2 P+1: M}$ is decodable from $\left(A_{1: N}^{\left[\mathcal{P}_{3}\right]}, \cdots, A_{1: N}^{\left[\mathcal{P}_{\beta}\right]}\right)$ after knowing $W_{1: 2 P}$. Finally, (122) follows from the interference conditioning lemma.

Consequently, (122) can be written as

$$
\begin{equation*}
N^{2} H\left(A_{1} \mid \mathcal{Q}\right) \geq(N+1) P L+H\left(A_{1} \mid W_{1: 2 P}, \mathcal{Q}\right) \tag{123}
\end{equation*}
$$

which is equivalent to

$$
\begin{equation*}
N H\left(A_{1} \mid \mathcal{Q}\right) \geq\left(1+\frac{1}{N}\right) P L+\frac{1}{N} H\left(A_{1} \mid W_{1: 2 P}, \mathcal{Q}\right) \tag{124}
\end{equation*}
$$

Now, (124) constructs an inductive relation, since evaluating $N H\left(A_{1} \mid W_{1: 2 P}, \mathcal{Q}\right)$ is the same as $N H\left(A_{1} \mid \mathcal{Q}\right)$ with $(M-2 P)$ messages, i.e., the problem of MPIR with $M$ messages for fixed $P$ is reduced to an MPIR problem with $(M-2 P)$ messages for the same fixed $P$. We note that (124) generalizes the inductive relation in [8] for $P=1$.

We can write the induction hypothesis for MPIR with $M$ messages as

$$
\begin{equation*}
N H\left(A_{1} \mid \mathcal{Q}\right) \geq P L\left[\sum_{i=0}^{\left\lfloor\frac{M}{P}\right\rfloor-1} \frac{1}{N^{i}}+\left(\frac{M}{P}-\left\lfloor\frac{M}{P}\right\rfloor\right) \frac{1}{N^{\left\lfloor\frac{M}{P}\right\rfloor}}\right] \tag{125}
\end{equation*}
$$

Next, we proceed with proving this relation for $M+1$ messages. From the induction hypothesis, we have

$$
\begin{align*}
& N H\left(A_{1} \mid W_{1: 2 P}, \mathcal{Q}\right) \\
& \geq P L\left[\sum_{i=0}^{\left\lfloor\frac{M-2 P+1}{P}\right\rfloor-1} \frac{1}{N^{i}}+\left(\frac{M-2 P+1}{P}\right.\right. \\
& \left.\quad-\left\lfloor\frac{M-2 P+1}{P}\right\rfloor\right) \frac{1}{\left.N^{\left\lfloor\frac{M-2 P+1}{P}\right\rfloor}\right]}  \tag{126}\\
& =P L\left[\sum_{i=0}^{\left\lfloor\frac{\lfloor+1}{P}\right\rfloor-3} \frac{1}{N^{i}}+\left(\frac{M+1}{P}-\left\lfloor\frac{M+1}{P}\right\rfloor\right) \frac{1}{N^{\left\lfloor\frac{M+1}{P}\right\rfloor-2}}\right] \tag{127}
\end{align*}
$$

substituting this in (124),

$$
\begin{align*}
& N H\left(A_{1} \mid \mathcal{Q}\right) \\
& \begin{aligned}
& \geq\left(1+\frac{1}{N}\right) P L+\frac{P L}{N^{2}}\left[\sum_{i=0}^{\left\lfloor\frac{M+1}{P}\right\rfloor-3} \frac{1}{N^{i}}+\left(\frac{M+1}{P}\right.\right. \\
&\left.-\left\lfloor\frac{M+1}{P}\right\rfloor\right) \frac{1}{\left.N^{\left\lfloor\frac{M+1}{P}\right\rfloor-2}\right]} \\
&=P L\left[\sum_{i=0}^{\left\lfloor\frac{M+1}{P}\right\rfloor-1} \frac{1}{N^{i}}+\left(\frac{M+1}{P}-\left\lfloor\frac{M+1}{P}\right\rfloor\right) \frac{1}{\left.N^{\left\lfloor\frac{M+1}{P}\right\rfloor}\right]}\right.
\end{aligned}
\end{align*}
$$

which concludes the induction argument.
Consequently, the upper bound for the MPIR problem can be obtained as,

$$
\begin{align*}
\sum_{i=1}^{P} R_{i} & =\frac{P L}{\sum_{n=1}^{N} H\left(A_{n}^{[\mathcal{P}]}\right)}  \tag{130}\\
& \leq \frac{P L}{N H\left(A_{1} \mid \mathcal{Q}\right)}  \tag{131}\\
& =\frac{1}{\sum_{i=0}^{\left\lfloor\frac{M}{P}\right\rfloor-1} \frac{1}{N^{i}}+\left(\frac{M}{P}-\left\lfloor\frac{M}{P}\right\rfloor\right) \frac{1}{N^{\left\lfloor\frac{M}{P}\right\rfloor}}}  \tag{132}\\
& =\left(\frac{1-\left(\frac{1}{N}\right)^{\left\lfloor\frac{M}{P}\right\rfloor}}{1-\frac{1}{N}}+\left(\frac{M}{P}-\left\lfloor\frac{M}{P}\right\rfloor\right) \frac{1}{N^{\left\lfloor\frac{M}{P}\right\rfloor}}\right)^{-1} \tag{133}
\end{align*}
$$

where (132) follows from (129); and (133) follows from evaluating the sum in (132).

## VII. Conclusions

In this paper, we introduced the multi-message private information retrieval (MPIR) problem from an informationtheoretic perspective. The problem generalizes the PIR problem in [8] which retrieves a single message privately. We determined the exact sum capacity for this problem when the number of desired messages is at least half of the number of total stored messages to be $C_{s}^{P}=\frac{1}{1+\frac{M-P}{P N}}$. We showed
that joint retrieval of the desired messages strictly outperforms repeating the single-message capacity achieving scheme for each message. Furthermore, we showed that if the total number of messages is an integer multiple of the number of desired messages, then the sum capacity is $C_{s}^{P}=\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{M / P}}$, which resembles the single-message PIR capacity expression when the number of messages is $\frac{M}{P}$. For the remaining cases, we derived lower and upper bounds. We observed numerically that the gap between the lower and upper bounds decreases monotonically in $N$, and the worst case gap is 0.0082 which occurs for the case $N=2$ when $M=5, P=2$.

The MPIR problem can be extended in several interesting directions. First, we recall from earlier remarks in the paper that the sum capacity for $M / P \notin \mathbb{N}$ is still an open problem, in addition to characterizing the optimal capacity region. Second, the MDS-coded MPIR as an extension of [12] is an interesting open problem, as the contents of the databases are themselves coded via an MDS code in [12]. This is a challenging problem, in particular if $P \geq \frac{M}{2}$, because our achievable scheme here uses a $P \times M$ MDS code; it would be interesting to see how the storage MDS code and the retrieval MDS code would interact. Similar difficulties would exist in the MPIR problem with colluding databases (extending [13]), robust MPIR problem (extending [13]), and MPIR problem with Byzantine databases (extending [28]), as all these problems adopt some version of MDS coding for retrieval purposes. Furthermore, one can examine whether multiround MPIR enhances the MPIR retrieval rate or not (extending the case of single-message retrieval in [29]), and study the effects of limited message size on MPIR (extending [25]). Our converse techniques may be generalized to be applicable to these scenarios. Some progress in these MPIR problems has been made recently in [30].

## REFERENCES

[1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," J. ACM, vol. 45, no. 6, pp. 965-981, 1998.
[2] S. Yekhanin, "Private information retrieval," Commun. ACM, vol. 53, no. 4, pp. 68-73, Apr. 2010.
[3] W. Gasarch, "A survey on private information retrieval," in Proc. Bull. EATCS, vol. 82. 2004, pp. 72-107.
[4] R. Ostrovsky and W. E. Skeith, III, "A survey of single-database private information retrieval: Techniques and applications," in Proc. Int. Workshop Public Key Cryptogr., 2007, pp. 393-411.
[5] C. Cachin, S. Micali, and M. Stadler, "Computationally private information retrieval with polylogarithmic communication," in Proc. Int. Conf. Theory Appl. Cryptogr. Techn., 1999, pp. 402-414.
[6] H. Sun and S. A. Jafar. (Jan. 2016). "Blind interference alignment for private information retrieval." [Online]. Available: https://arxiv.org/abs/ 1601.07885
[7] S. A. Jafar, "Blind interference alignment," IEEE J. Sel. Topics Signal Process., vol. 6, no. 3, pp. 216-227, Jun. 2012.
[8] H. Sun and S. A. Jafar, "The capacity of private information retrieval," IEEE Trans. Inf. Theory, vol. 63, no. 7, pp. 4075-4088, Jul. 2017.
[9] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in Proc. IEEE ISIT, Jun./Jul. 2014, pp. 856-860.
[10] R. Tajeddine and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," in Proc. IEEE ISIT, Jul. 2016, pp. 1411-1415.
[11] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in Proc. IEEE ISIT, Jun. 2015, pp. 2842-2846.
[12] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," IEEE Trans. Inf. Theory, vol. 64, no. 3, pp. 1945-1956, Mar. 2018.
[13] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," IEEE Trans. Inf. Theory, vol. 64, no. 4, pp. 2361-2370, Apr. 2018.
[14] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," SIAM J. Appl. Algebra Geometry, vol. 1, no. 1, pp. 647-664, 2017.
[15] H. Sun and S. A. Jafar. (Jun. 2016). "The capacity of symmetric private information retrieval." [Online]. Available: https://arxiv.org/abs/1606. 08828
[16] Q. Wang and M. Skoglund. (Oct. 2016). "Symmetric private information retrieval for MDS coded distributed storage." [Online]. Available: https://arxiv.org/abs/1610.04530
[17] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York, NY, USA: Wiley, 2012.
[18] R. Henry, Y. Huang, and I. Goldberg, "One (block) size fits all: PIR and SPIR with variable-length records via multi-block queries," in Proc. NDSS, 2013.
[19] D. Demmler, A. Herzberg, and T. Schneider, "RAID-PIR: Practical multi-server PIR," in Proc. ACM Workshop Cloud Comput. Secur., 2014, pp. 45-56.
[20] L. Wang, T. K. Kuppusamy, Y. Liu, and J. Cappos, "A fast multiserver, multi-block private information retrieval protocol," in Proc. IEEE Globecom, Dec. 2015, pp. 1-6.
[21] J. Groth, A. Kiayias, and H. Lipmaa, "Multi-query computationallyprivate information retrieval with constant communication rate," in Proc. Int. Workshop Public Key Cryptogr., 2010, pp. 107-123.
[22] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Batch codes and their applications," in Proc. 36th Annu. ACM Symp. Theory Comput., 2004, pp. 262-271.
[23] W. Lueks and I. Goldberg, "Sublinear scaling for multi-client private information retrieval," in Proc. Int. Conf. Financial Cryptogr. Data Secur. 2015, pp. 168-186.
[24] A. V. Oppenheim and R. W. Schafer, Discrete Time Signal Processing. London, U.K.: Pearson Education, 2010.
[25] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," IEEE Trans. Inf. Forensics Security, vol. 12, no. 12, pp. 2920-2932, Dec. 2017.
[26] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," J. Soc. Ind. Appl. Math., vol. 8, no. 2, pp. 300-304, Jun. 1960.
[27] S. B. Wicker and V. K. Bhargava, Reed-Solomon Codes and Their Applications. New York, NY, USA: Wiley, 1999.
[28] K. Banawan and S. Ulukus. (Jun. 2017). "The capacity of private information retrieval from Byzantine and colluding databases." [Online]. Available: https://arxiv.org/abs/1706.01442
[29] H. Sun and S. A. Jafar. (Nov. 2016). "Multiround private information retrieval: Capacity and storage overhead." [Online]. Available: https://arxiv.org/abs/1611.02257
[30] Y. Zhang and G. Ge. (Oct. 2017). "Private information retrieval from MDS coded databases with colluding servers under several variant models." [Online]. Available: https://arxiv.org/abs/1705.03186

Karim Banawan (S'13) received the B.Sc. and M.Sc. degrees (Highest Hons.) in electrical engineering from Alexandria University, Alexandria, Egypt, in 2008, 2012, respectively, M.Sc. degree in electrical engineering from University of Maryland, College Park, MD, USA in 2017. He is currently pursuing the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. His research interests include information theory, wireless communications, physical layer security and private information retrieval.

Sennur Ulukus (S'90-M'98-SM'15-F'16) is a Professor of Electrical and Computer Engineering at the University of Maryland at College Park, where she also holds a joint appointment with the Institute for Systems Research (ISR). Prior to joining UMD, she was a Senior Technical Staff Member at AT\&T Labs-Research. She received her Ph.D. degree in Electrical and Computer Engineering from Wireless Information Network Laboratory (WINLAB), Rutgers University, and B.S. and M.S. degrees in Electrical and Electronics Engineering from Bilkent University. Her research interests are in wireless communications, information theory, signal processing, and networks, with recent focus on private information retrieval, timely status updates over networks, energy harvesting communications, information theoretic physical layer security, and wireless energy and information transfer.

Dr. Ulukus is a fellow of the IEEE, and a Distinguished Scholar-Teacher of the University of Maryland. She received the 2003 IEEE Marconi Prize Paper Award in Wireless Communications, an 2005 NSF CAREER Award, the 2010-2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Education Award. She is a Distinguished Lecturer of the Infomation Theory Society for 2018-2019. She is on the Editorial Board of the IEEE Transactions on Green Communications and Networking since 2016. She was an Editor for the IEEE Journal on Selected Areas in Communications Series on Green Communications and Networking (2015-2016), IEEE Transactions on Information Theory (2007-2010), and IEEE Transactions on Communications (2003-2007). She was a Guest Editor for the IEEE Journal on Selected Areas in Communications (2015 and 2008), Journal of Communications and Networks (2012), and IEEE Transactions on Information Theory (2011). She was a general TPC co-chair of 2017 IEEE ISIT, 2016 IEEE Globecom, 2014 IEEE PIMRC, and 2011 IEEE CTW.


[^0]:    Manuscript received February 5, 2017; revised December 20, 2017; accepted April 4, 2018. Date of publication April 19, 2018; date of current version September 13, 2018. This work was supported by NSF under Grant CNS 1314733, Grant CCF 14-22111, Grant CCF 14-22129, and Grant CNS 15-26608. This paper was presented in part at the IEEE ISIT 2017.
    The authors are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: kbanawan@umd.edu; ulukus@umd.edu).
    Communicated by A. Ramamoorthy, Associate Editor for Coding Techniques.
    Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.
    Digital Object Identifier 10.1109/TIT.2018.2828310

[^1]:    ${ }^{1}$ We note that using $q=\min \left\{p^{m} \geq M: p\right.$ is a prime, $\left.m \in \mathbb{N}\right\}$ is sufficient to ensure the existence of the $P \times \bar{M}$ MDS generator matrix in Section 4. Furthermore, binary field suffices for the achievable scheme in Section 5.

[^2]:    ${ }^{2}$ Note that in the degenerate case, when $P=M$, the privacy constraint is trivially satisfied as $H(\mathcal{P})=H\left(\mathcal{P} \mid Q_{n}^{[\mathcal{P}]}\right)=0$ as there is no uncertainty about the identity of the desired messages if $P=M$ either with or without the knowledge of the queries. Thus, the optimal sum retrieval rate is 1 which is achieved by downloading all the messages.

[^3]:    ${ }^{4}$ We note that this is similar to the TPIR problem when $\frac{N}{T} \in \mathbb{N}$, in which case one cannot simply lump every $T$ databases together and apply the capacity-achieving scheme of PIR with non-colluding databases for the new system that consists of $\frac{N}{T}$ databases. In both problems, the use of MDS codes is important to induce symmetry across the group of messages/databases.

[^4]:    ${ }^{5}$ Although it seems at first that $C_{s}^{P} \rightarrow 1-\frac{1}{N}$ if $M \rightarrow \infty$, we emphasize that this is true if only $P=o(M)$, i.e., $P$ does not scale with $M$. If $P=\gamma M$, then as $M \rightarrow \infty$, we have $C_{S}^{P}=\frac{1}{1+\frac{1-\gamma}{\gamma N}}>1-\frac{1}{N}$, when $\gamma>\frac{1}{2}$, and $C_{s}^{P}=\frac{1-\frac{1}{N}}{1-\left(\frac{1}{N}\right)^{1 / \gamma}}>1-\frac{1}{N}$, when $\frac{1}{\gamma} \geq 2 \in \mathbb{N}$.

[^5]:    ${ }^{6}$ Check for instance in Table IV that all of the downloads (equations) involving undesired symbols from database 2 are used in database 1: singles $c_{6}, d_{6}, e_{6}, c_{7}, d_{7}, e_{7}, c_{8}, d_{8}, e_{8}, c_{9}, d_{9}, e_{9}, c_{10}, d_{10}, e_{10}$; sums of twos $c_{15}+$ $d_{15}, c_{16}+e_{15}, d_{16}+e_{16}, c_{17}+d_{17}, c_{18}+e_{17}, d_{18}+e_{18}$; sum of threes $c_{20}+d_{20}+e_{20}$, all downloaded from database 2 are all used as side information in database 1 .

[^6]:    ${ }^{7}$ We note that the assumption that $W_{i} \in \mathbb{F}_{q}^{L}$ is indeed unnecessary in terms of converse arguments. Consequently, our converse proof is valid for any storage alphabet.

