Noisy Private Information Retrieval: On Separability of Channel Coding and Information Retrieval

Karim Banawan, *Member, IEEE*, and Sennur Ulukus, *Fellow, IEEE*

***Abstract*— We consider the problem of noisy private information retrieval (NPIR) from N non-communicating databases, each storing the same set of M messages. In this model, the answer strings are not returned through noiseless bit pipes, but rather through *noisy* memoryless channels. We aim at characterizing the PIR capacity for this model as a function of the statistical information measures of the noisy channels such as entropy and mutual information. We derive a general upper bound for the retrieval rate in the form of a max-min optimization. We use the achievable schemes for the PIR problem under asymmetric traffic constraints and random coding arguments to derive a general lower bound for the retrieval rate. The upper and lower bounds match for $M = 2$ and $M = 3$, for any N, and any noisy channel. The lower and upper bounds show a separation between channel coding and retrieval scheme except for adapting the traffic ratio from the databases. We refer to this as *almost separation*. Next, we consider the private information retrieval problem from multiple access channels (MAC-PIR). In MAC-PIR, the database responses reach the user through a multiple access channel (MAC) that mixes the responses together in a stochastic way. We show that for the additive MAC and the conjunction/disjunction MAC, channel coding and retrieval scheme are *inseparable* unlike in NPIR. We show that the retrieval scheme depends on the properties of the MAC, in particular on the linearity aspect. For both cases, we provide schemes that achieve the full capacity without any loss due to the privacy constraint, which implies that the user can exploit the nature of the channel to improve privacy. Finally, we show that the full unconstrained capacity is not always attainable by determining the capacity of the selection channel.**

***Index Terms*— Private information retrieval (PIR), noisy channels, multiple access channels (MACs), separation of channel coding and information retrieval.**

I. Introduction

IN THE era of big data, efficient data-mining techniques are present everywhere, from social media to online-shopping and search history. These new challenges motivate studying the privacy issues that arise in modern networks. Private information retrieval (PIR), introduced by Chor *et al.* [1] and remained an important research avenue in computer science community (see for example [1]–[5]), is a canonical problem to study the privacy of the downloaded content from public databases. In the classical PIR, a user wishes to retrieve a file privately from N distributed and non-colluding databases each storing the same set of M messages (files), in a way that no database can learn the identity of the user's desired file. To that end, the user submits queries for the databases that do not reveal the user's interest in the desired file. The databases respond with *correct* answer strings via *noiseless orthogonal links*, from which the user reconstructs the desired file. PIR schemes are designed to be more efficient than the trivial scheme of downloading all the files stored in the databases in terms of the retrieval rate, which is defined as the ratio between the number of downloaded bits from the desired message and the total download.

Recently, the PIR problem has attracted a renewed interest within the information theory community [6]–[10]. In order to characterize the fundamental limits of the problem, Sun-Jafar introduced the notion of PIR capacity C_{PIR} in [11], which is defined as the supremum of all PIR rates over all achievable retrieval schemes. Reference [11] proved that for the classical PIR model, $C_{\text{PIR}} = (1+\frac{1}{N}+\cdots+\frac{1}{N^{M-1}})^{-1}$. The achievability scheme is a greedy algorithm that employs a *symmetric query* structure for all databases. Following [11], the capacities of many interesting variants of the classical PIR problem have been considered [12]–[43].

In all previous works, the links from the databases to the user are assumed to be noiseless. Furthermore, these works assume that the answer strings are returned via orthogonal links, i.e., the user receives N separate answer strings, which are not mixed. There are many practical settings where these assumptions may not be valid. For instance, while browsing (retrieving information on) the internet, some packets may be dropped randomly. This scenario can be abstracted out as passing the answer strings through an erasure channel. Alternatively, the data packets may be randomly corrupted, which can be modeled as a binary symmetric channel that flips randomly some symbols in the answer strings. Consequently, a more realistic retrieval model may be to assume that the databases return their answer strings through memoryless noisy channels with known transition probabilities. The noisy

Manuscript received July 14, 2018; revised March 2, 2019; accepted July 31, 2019. Date of publication August 15, 2019; date of current version November 20, 2019. This work was supported by NSF under Grant CNS 13-14733, Grant CCF 14-22111, Grant CNS 15-26608, and Grant CCF 17-13977. This article was presented in part at the IEEE Asilomar 2018 and the IEEE ITW 2018.

K. Banawan was with the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD 20742 USA. He is now with the Electrical Engineering Department, Faculty of Engineering, Alexandria University, Alexandria 21544, Egypt (e-mail: kbanawan@alexu.edu.eg).

S. Ulukus is with the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD 20742 USA (e-mail: ulukus@umd.edu).

Communicated by M. R. Bloch, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2935440

nature of the channel induces random errors along the received answer strings.

Yet, in other applications, the answer strings may be mixed before reaching the user. For example: if the user is retrieving the desired file from wireless base stations, the answer strings would be combined on the air before reaching the user. Another example is retrieval from a cloud, where the returned packets may collide and superimpose each other. These practical settings can be represented with another abstract model, which is the cooperative multiple access channel[1] (MAC) model, where the databases cooperate to convey the desired message to the user, while the user receives a stochastic mapping from the database responses in general. These two cases, namely, *noisy* and *multiple access* nature of retrieval channels, pose many interesting questions, such as: How to devise schemes that mitigate the errors introduced by the channel with a small sacrifice from the private retrieval rate? Is there a *separation* between the channel coding needed for reliable transmission over noisy channels and the private retrieval scheme, or if there is a necessity for joint processing? How do the statistical properties of the noisy channels fundamentally affect the private retrieval rate?

In this paper, we introduce noisy PIR with orthogonal links (NPIR) and PIR problem from multiple access channel (MAC-PIR). We first focus on the NPIR problem and then consider the MAC-PIR problem in Section VI. In NPIR, the nth database is connected to the user via a discrete memoryless channel with known transition probability distribution $p(y_n|x_n)$. Hence, the user needs to decode the desired message *reliably* by observing the noisy versions of the returned answer strings. Intuitively, since a channel with worse channel condition needs a lower code rate to combat the channel errors, we do not expect the lengths of the answer strings to be the same from all the databases. Therefore, in this work, we allow the traffic from each database to be *asymmetric* as in [38] and [39]. In this work, we aim at characterizing the capacity of the NPIR problem in terms of the statistical information measures of the noisy channels such as mutual information, the number of messages M, and the number of databases N. To that end, we first derive a general upper bound for the retrieval rate in the form of a max-min problem. The converse proof is inspired by the converse proof in [38], in particular in the way the asymmetry is handled. We show the achievability proof by random coding arguments and enforcing the uncoded responses to operate at one of the corner points of the PIR problem under asymmetric traffic constraints. The upper and lower bounds match for $M = 2$ and $M = 3$ messages, for arbitrary N databases, and any noisy channel. Our results show that the channel coding needed to mitigate the channel errors and the retrieval scheme are *almost separable* in the sense that the noisy channels affect only the traffic ratio requested from each database and not the explicit coding technique. Interestingly, the upper and lower bounds depend only on the *capacity* of the noisy channels and not on the explicit transition probability of the channels.

In the MAC-PIR problem, the responses of the databases reach the user through a discrete memoryless MAC with a known transition probability $p(y|x_1, \cdots, x_N)$. In this case, the output of the channel is a mixture (possibly noisy mixture) of all database responses. The user needs to decode the desired message with vanishingly small probability of error from the output of the channel. Interestingly, for this model, we show that channel coding and retrieval strategy are *inseparable* unlike in the NPIR problem. We show this fact by deriving the PIR capacity of two simple MACs, namely: additive MAC, and logical conjunction/disjunction MAC. In these two cases, we show that *privacy for free* can be attained by designing retrieval strategies that exploit the properties of the channel to maximize the retrieval rate. Interestingly, we show that for the additive MAC, the optimal PIR scheme is linear, while for the logical conjunction/disjunction MAC we show that a non-linear PIR scheme, that requires $N \geq 2^{M-1}$ is needed to achieve $C_{PIR} = 1$. We conclude this discussion by showing that full unconstrained capacity may not be attainable for all MACs by giving a counterexample, which is the selection MAC, which has a capacity of $C_{PIR} = \frac{1}{M}$. The exact PIR capacity of the MAC-PIR for an arbitrary transition probability distribution remains an open problem in general.

II. System Model

We consider a classical PIR model with N replicated and non-communicating databases storing M messages. Each database stores the same set of messages $W_{1:M} = \{W_1, \cdots, W_M\}$. The mth message W_m is an L-length binary (without loss of generality) vector picked uniformly from $\mathbb{F}_2^L$. The messages $W_{1:M}$ are independent and identically distributed, i.e.,

$$H(W_m) = L, \quad m \in \{1, \cdots, M\} \tag{1}$$

$$H(W_{1:M}) = ML \tag{2}$$

In PIR, a user wants to retrieve a message W_i reliably and privately. To that end, the user submits N queries $Q_{1:N}^{[i]} = \{Q_1^{[i]}, \cdots, Q_N^{[i]}\}$, one for each database. Since the user does not have any information about the message set in advance, the queries and the messages are statistically independent,

$$I(W_{1:M}; Q_{1:N}^{[i]}) = 0, \quad i \in \{1, \cdots, M\} \tag{3}$$

The nth database responds to $Q_n^{[i]}$ with a t_n-length answer string $A_n^{[i]} = (X_{n,1}^{[i]}, \cdots, X_{n,t_n}^{[i]})$. The nth answer string is a deterministic function of the messages $W_{1:M}$ and the query $Q_n^{[i]}$, hence,

$$H(A_n^{[i]}|W_{1:M}, Q_n^{[i]}) = 0, \ n \in \{1, \cdots, N\}, \ i \in \{1, \cdots, M\} \tag{4}$$

In noisy PIR with orthogonal links (NPIR, see Fig. 1), the user receives the nth answer string via a discrete memoryless channel (response channel) with a transition probability $p(y_n|x_n)$. In this model, the noisy channels are *orthogonal*,

[1]We note that by *cooperative MAC*, we do not mean *database collusion* which refers to database communication to try to figure out the identity of the desired file. Instead, by *cooperative MAC*, we mean the implicit cooperation created by the user (retriever) through a careful design of queries. In our system, there is no explicit communication between the databases, however, since the user can jointly design the queries to all databases and the databases respond truthfully to user queries, the responses to the user can be thought of as codewords from cooperative users in a multiple access channel, i.e., there is an implicit cooperation in the responses through the joint design of the queries.

in the sense that the noisy answer strings do not interact (mix). Thus, the user receives a noisy answer string $\tilde{A}_n^{[i]} = (Y_{n,1}^{[i]}, \cdots, Y_{n,t_n}^{[i]})$. Therefore, we have[2],

$$P\left(\tilde{A}_n^{[i]} = (y_{n,1}^{[i]}, \cdots, y_{n,t_n}^{[i]}) | A_n^{[i]} = (x_{n,1}^{[i]}, \cdots, x_{n,t_n}^{[i]})\right) = \prod_{\eta_n=1}^{t_n} p\left(y_{n,\eta_n}^{[i]} | x_{n,\eta_n}^{[i]}\right) \quad (5)$$

Consequently, $(W_{1:M}, Q_n^{[i]}) \to A_n^{[i]} \to \tilde{A}_n^{[i]}$ forms a Markov chain. Let us denote the channel capacity of the nth response channel by C_n, denote,

$$C_n = \max_{p(x_n)} \; I(X_n; Y_n) \quad (6)$$

where X_n, Y_n are the single-letter input and output pair for the nth response channel. Without loss of generality, assume that the channel capacities are ordered such that $C_1 \geq C_2 \geq \cdots \geq C_N$, i.e., the channel capacities form a non-increasing sequence. Let $\mathbf{C} = (C_1, \cdots, C_N)$ be the vector of the channel capacities.

We note that, in general, the user and the databases can agree on suitable lengths $\{t_n\}_{n=1}^N$ for the answer strings, which may not be equal in general, such that they maximize the retrieval rate. Let us define the traffic ratio vector $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_N)$ as,

$$\tau_n = \frac{t_n}{\sum_{j=1}^N t_j}, \quad n \in \{1, \cdots, N\} \quad (7)$$

To ensure privacy, the queries $Q_{1:N}^{[i]}$ should be designed such that the query to the nth database does not reveal any information about i. We can write the privacy constraint as

$$(Q_n^{[i]}, A_n^{[i]}, W_{1:M}) \sim (Q_n^{[j]}, A_n^{[j]}, W_{1:M}), \quad \forall i, j \in \{1, \cdots, M\} \quad (8)$$

We note that from privacy constraint and due to the Markov chain $(W_{1:M}, Q_n^{[i]}) \to A_n^{[i]} \to \tilde{A}_n^{[i]}$, we may write that $(Q_n^{[i]}, A_n^{[i]}, \tilde{A}_n^{[i]}, W_{1:M}) \sim (Q_n^{[j]}, A_n^{[j]}, \tilde{A}_n^{[j]}, W_{1:M})$, $\forall i, j \in \{1, \cdots, M\}$.

In addition, the user should be able to reconstruct the desired message W_i by observing the noisy answer strings $\tilde{A}_{1:N}^{[i]}$ with arbitrarily small probability of error $P_e(L)$, i.e., $P_e(L) \to 0$ as $L \to \infty$. Hence, from Fano's inequality, we have,

$$H(W_i | Q_{1:N}^{[i]}, \tilde{A}_{1:N}^{[i]}) \leq 1 + P_e(L) \cdot L = o(L) \quad (9)$$

where $\frac{o(L)}{L} \to 0$ as $L \to \infty$.

[2]Here, we comment on the differences between the NPIR model introduced in this paper and the BPIR model (PIR from Byzantine databases) in [24]. In BPIR, there are B databases that respond arbitrarily untruthfully to the user queries, i.e., $H(\tilde{A}_n^{[i]} | W_{1:M}, Q_n^{[i]}) > 0$ for all $n \in \mathcal{B}$ such that $|\mathcal{B}| = B$. The members of the Byzantine database set $\mathcal{B}$ are unknown to the user. The Byzantine databases can respond by any possible error pattern they wish, i.e., they can use any transition probability distribution for their response channels. Thus, in BPIR, the user does not have any knowledge about the transition probability distribution of the response channels unlike in NPIR. Nevertheless, since the user in BPIR has the knowledge that there are exactly B Byzantine databases, that setting results in a more structured error pattern than NPIR here. For these reasons, the two problems are entirely different and we cannot think of the BPIR problem as an NPIR problem with channel capacity vector $\boldsymbol{C} = \{1, 1, \cdots, 1, 0, 0, \cdots, 0\}$ with B zeros and $N - B$ ones, and arbitrary placement of ones and zeros. The techniques and the results for these two problems are different as well.

For a fixed traffic ratio vector $\boldsymbol{\tau}$, the retrieval rate $R(\boldsymbol{\tau}, \mathbf{C})$ is achievable if there exists a sequence of retrieval schemes, indexed by the message length L, that satisfy the privacy constraint (8) and the reliability constraint (9) with answer string lengths $\{t_n\}_{n=1}^N$ that conform with (7), thus,

$$R(\boldsymbol{\tau}, \mathbf{C}) = \lim_{L \to \infty} \frac{L}{\sum_{n=1}^N t_n} \quad (10)$$

Consequently, the retrieval rate $R(\mathbf{C})$ is the supremum of $R(\boldsymbol{\tau}, \mathbf{C})$ over all traffic ratio vectors in $\mathbb{T} = \{(\tau_1, \cdots, \tau_N) : \tau_n \geq 0 \;\; \forall n, \sum_{n=1}^N \tau_n = 1\}$. The PIR capacity for this model $C_{\text{PIR}}(\mathbf{C})$ is given by

$$C_{\text{PIR}}(\mathbf{C}) = \sup \; R(\mathbf{C}) \quad (11)$$

where the supermum is over all achievable retrieval schemes.

III. Main Results and Discussions on NPIR

In this section, we present the main results of the NPIR problem. The first result gives an upper bound for the NPIR problem.

Theorem 1 (Upper bound) *For NPIR with noisy links of capacities $\mathbf{C} = (C_1, \cdots, C_N)$, the retrieval rate is upper bounded by,*

$$\begin{aligned} C_{PIR}(\mathbf{C}) &\leq \bar{C}_{PIR}(\mathbf{C}) \\ &= \max_{\boldsymbol{\tau} \in \mathbb{T}} \min_{n_i \in [N]} \frac{\theta(0) + \frac{\theta(n_1)}{n_1} + \frac{\theta(n_2)}{n_1 n_2} + \cdots + \frac{\theta(n_{M-1})}{\prod_{i=1}^{M-1} n_i}}{1 + \frac{1}{n_1} + \frac{1}{n_1 n_2} + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}} \end{aligned} \quad (12)$$

where $\mathbb{T} = \left\{\boldsymbol{\tau} : \tau_n \geq 0 \;\; \forall n \in [1:N], \;\; \sum_{n=1}^N \tau_n = 1\right\}$, $[N] = \{1, \cdots, N\}$ *and* $\theta(\ell) = \sum_{n=\ell+1}^N \tau_n C_n$.

The proof of this upper bound is given in Section IV. The second result gives an achievability scheme for the NPIR problem.

Theorem 2 (Lower bound) *For NPIR with noisy links of capacities $\mathbf{C} = (C_1, \cdots, C_N)$, for a monotone non-decreasing sequence $\mathbf{n} = \{n_i\}_{i=0}^{M-1} \subset \{1, \cdots, N\}^M$, let $n_{-1} = 0$, and $\mathcal{S} = \{i \geq 0 : n_i - n_{i-1} > 0\}$. Denote $y_\ell[k]$ to be the number of stages of the achievable scheme that downloads k-sums from the nth database in one repetition of the scheme, such that $n_{\ell-1} \leq n \leq n_\ell$, and $\ell \in \mathcal{S}$. Let $\xi_\ell = \prod_{s \in \mathcal{S} \setminus \{\ell\}} \binom{M-2}{s-1}$. The number of stages $y_\ell[k]$ is characterized by the following system of difference equations*:

$$\begin{aligned} y_0[k] &= (n_0 - 1) y_0[k-1] + \sum_{j \in \mathcal{S} \setminus \{0\}} (n_j - n_{j-1}) y_j[k-1] \\ y_1[k] &= (n_1 - n_0 - 1) y_1[k-1] + \sum_{j \in \mathcal{S} \setminus \{1\}} (n_j - n_{j-1}) y_j[k-1] \\ y_\ell[k] &= n_0 \xi_\ell \delta[k - \ell - 1] + (n_\ell - n_{\ell-1} - 1) y_\ell[k-1] \\ &\quad + \sum_{j \in \mathcal{S} \setminus \{\ell\}} (n_j - n_{j-1}) y_j[k-1], \quad \ell \geq 2 \end{aligned} \quad (13)$$

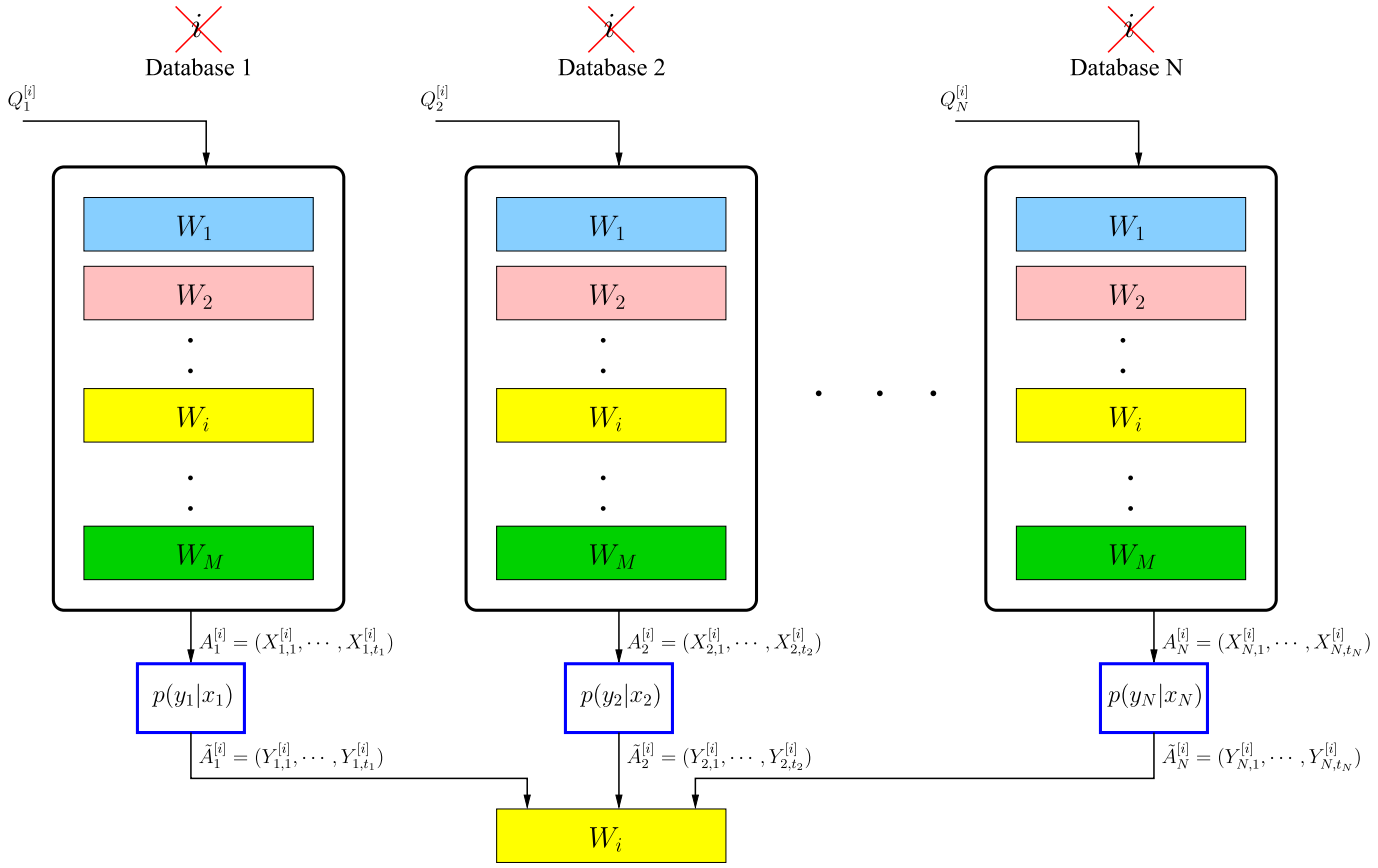

Fig. 1. The noisy PIR (NPIR) problem.

where $\delta[\cdot]$ *is the Kronecker delta function. The initial conditions of* (13) *are* $y_0[1] = \prod_{s\in\mathcal{S}} \binom{M-2}{s-1}$, *and* $y_j[k] = 0$ *for* $k \leq j$. *Then, the achievable rate corresponding to* $\mathbf{n}$ *is given by*:

$$R(\mathbf{n}, \mathbf{C}) = \frac{\sum_{\ell\in\mathcal{S}} \sum_{k=1}^{M} \binom{M-1}{k-1} y_\ell[k](n_\ell - n_{\ell-1})}{\sum_{\ell\in\mathcal{S}} \sum_{n=n_{\ell-1}+1}^{n_\ell} \frac{\sum_{k=1}^{M} \binom{M}{k} y_\ell[k]}{C_n}} \tag{14}$$

Consequently, the capacity $C_{PIR}(\mathbf{C})$ *is lower bounded by*:

$$\begin{aligned} C_{PIR}(\mathbf{C}) &\geq R(\mathbf{C}) \\ &= \max_{n_i\in[N]} R(\mathbf{n}, \mathbf{C}) \qquad (15) \\ &= \max_{n_i\in[N]} \frac{\sum_{\ell\in\mathcal{S}} \sum_{k=1}^{M} \binom{M-1}{k-1} y_\ell[k](n_\ell - n_{\ell-1})}{\sum_{\ell\in\mathcal{S}} \sum_{n=n_{\ell-1}+1}^{n_\ell} \frac{\sum_{k=1}^{M} \binom{M}{k} y_\ell[k]}{C_n}} \qquad (16) \end{aligned}$$

where $n_0 \leq n_1 \leq \cdots \leq n_{M-1}$

The proof of this lower bound is given in Section V. We have the following remarks[3].

[3]We note that variable n_i refers to the index of the last database that exploits i-sum as side information in its initial round of download. This means that we have $n_i - n_{i-1}$ databases that exploit i-sum side information symbols in the initial round of download and have the same uncoded traffic ratio (before applying the channel code). The user optimizes over $n_i \in \{1, \cdots, N\}$ to maximize the PIR retrieval rate. For more details about the notation, please refer to [38].

Remark 1 *The upper and lower bounds for the retrieval rate are similar to the corresponding bounds for the PIR-WTC-II problem [39] after replacing the secrecy capacity of WTC-II,* $1-\mu_n$, *with the capacity of the noisy link* C_n. *Thus, the NPIR problem inherits all the structural remarks of the PIR-WTC-II problem.*

Remark 2 *The upper and lower bounds for the retrieval rate do not depend explicitly on the transition probabilities of the noisy channels* $p(y_n|x_n)$, *but rather depend on the capacities of the noisy channels* C_n.

Remark 3 *Theorem 1 and Theorem 2 imply that the channel coding needed for combating channel errors is "almost seperable" from the retrieval scheme. More specifically, the expressions for the upper and lower bounds are separable in terms of the privacy and channel effects, as both expressions depend directly on* C_n *for all* n, *which suggests applying the capacity achieving code for each noisy channel for the output of the private retrieval scheme. The channel coding problem and the retrieval problem are coupled only through agreeing on a traffic ratio vector* $\boldsymbol{\tau}$. *Other than* $\boldsymbol{\tau}$, *the channel coding acts as an outer code for the responses of the databases to the user queries. Interestingly, the result implies that our schemes work even for heterogeneous channels, e.g., if* $N = 2$, *the channel from one database can be a BSC, and the channel from the other database can be a BEC.*

Remark 4 *Our results imply that randomized strategies for PIR cannot increase the retrieval rate. We can view the noisy channel between the user and the database as a randomizer for the actions of the databases, which is available to the databases but not available to the user. Since the capacity expression does not depend on $p(y_n|x_n)$ and is always maximized by $C_n = 1$, any randomizing strategy $p(y_n|x_n)$ cannot enhance the retrieval rate, even if the databases can choose the transition probability distributions $p(y_n|x_n)$.*

Corollary 1 (Exact capacity for $M = 2$ and $M = 3$ messages) *For NPIR, the capacity $C_{PIR}(\mathbf{C})$ for $M = 2$, and an arbitrary N is given by*:

$$C(\mathbf{C}) = \max_{n_i \in [N]} \frac{n_0 n_1}{\sum_{n=1}^{n_0} \frac{n_0+1}{C_n} + \sum_{n=n_0+1}^{n_1} \frac{n_0}{C_n}} \tag{17}$$

and for $M = 3$, $C(\mathbf{C})$ is given by,

$$\max_{n_i \in [N]} \frac{n_0 n_1 n_2}{\sum_{n=1}^{n_0} \frac{n_0 n_1 + n_0 + 1}{C_n} + \sum_{n=n_0+1}^{n_1} \frac{n_0 n_1 + n_0}{C_n} + \sum_{n=n_1+1}^{n_2} \frac{n_0 n_1}{C_n}} \tag{18}$$

Remark 5 *As we will show in Section V, the retrieval scheme operates at one of the corner points of the PIR problem with asymmetric traffic constraints [38]. Here, we state the uncoded traffic (before applying the channel code) returned from the databases. The uncoded traffic ratio τ_n is a function of the sequence $\mathbf{n} = (n_0, n_1, n_2)$, which is specified by carrying out the optimization problem in* (15)*,* (17) *and* (18)*. Therefore, referring to [38, Section 6], for the case of $M = 2$, the uncoded traffic ratios from the databases are*:

$$\tau_n = \begin{cases} \frac{n_0}{n_0(n_1+1)}, & 1 \le n \le n_0 \\ \frac{1}{n_1+1}, & n_0 + 1 \le n \le n_1 \\ 0, & n > n_1 \end{cases} \tag{19}$$

For $M = 3$, the uncoded traffic ratios are:

$$\tau_n = \begin{cases} \frac{n_0 n_1 + n_0 + 1}{n_0(n_2 n_1 + n_1 + 1)}, & 1 \le n \le n_0 \\ \frac{n_1+1}{n_2 n_1 + n_1 + 1}, & n_0 + 1 \le n \le n_1 \\ \frac{n_1}{n_2 n_1 + n_1 + 1}, & n_1 + 1 \le n \le n_2 \\ 0, & n > n_2 \end{cases} \tag{20}$$

The proof of Corollary 1 follows from the optimality of the PIR-WTC-II scheme in [39] for $M = 2$ and $M = 3$ messages by replacing $1 - \mu_n$ by C_n.

a) Example: The capacity for NPIR from BSC(p_1), BSC(p_2), $N = 2$, $M = 3$: To show how Theorem 1 reduces to Corollary 1 for $M = 3$, we apply Theorem 1 to the case of $M = 3$, $N = 2$, and the links to the user are BSC(p_1), and BSC(p_2). From Theorem 1, we can write the upper bound for the achievable retrieval rate as:

$$\max_{\boldsymbol{\tau} \in \mathbb{T}} \min_{n_i \in \{1,2\}} \frac{\sum_{n=1}^{N} \tau_n C_n + \frac{\sum_{n=n_1+1}^{N} \tau_n C_n}{n_1} + \frac{\sum_{n=n_2+1}^{N} \tau_n C_n}{n_1 n_2}}{1 + \frac{1}{n_1} + \frac{1}{n_1 n_2}} \tag{21}$$

where $C_n = 1 - H(p_n)$.

By observing $\tau_2 = 1 - \tau_1$ and the fact that C_n is monotonically decreasing in p_n for $p_n \in (0, \frac{1}{2})$ (which implies that $p_1 \le p_2$ satisfies $C_1 \ge C_2$), (21) can be explicitly written as the following linear program:

$$\begin{aligned} \max_{\tau_2, R} \ & R \\ \text{s.t. } & R \le \frac{1}{3}(1-H(p_1)) + \left[(1-H(p_2)) - \frac{1}{3}(1-H(p_1))\right]\tau_2 \\ & R \le \frac{2}{5}(1-H(p_1)) + \left[\frac{4}{5}(1-H(p_2)) - \frac{2}{5}(1-H(p_1))\right]\tau_2 \\ & R \le \frac{4}{7}(1-H(p_1)) + \left[\frac{4}{7}(1-H(p_2)) - \frac{4}{7}(1-H(p_1))\right]\tau_2 \\ & 0 \le \tau_2 \le 1 \end{aligned} \tag{22}$$

The bound corresponding to $n_1 = 2$, $n_2 = 1$ is inactive for all values of (p_1, p_2). Since (22) is a linear program, its solution resides at the corner points of the feasible region. The first corner point occurs at $\tau_2^{(1)} = 0$, which corresponds to the upper bound $R \le \frac{1-H(p_1)}{3}$. The second corner point is at the intersection of the first two constraints, i.e.,

$$\begin{aligned} &\frac{1}{3}(1 - H(p_1)) + \left[(1 - H(p_2)) - \frac{1}{3}(1 - H(p_1))\right]\tau_2^{(2)} \\ &= \frac{2}{5}(1 - H(p_1)) + \left[\frac{4}{5}(1 - H(p_2)) - \frac{2}{5}(1 - H(p_1))\right]\tau_2^{(2)} \end{aligned} \tag{23}$$

which leads to,

$$\tau_2^{(2)} = \frac{1 - H(p_1)}{3(1 - H(p_2)) + (1 - H(p_1))} \tag{24}$$

which corresponds to the upper bound $R \le \frac{2}{\frac{3}{1-H(p_1)} + \frac{1}{1-H(p_2)}}$. Similarly, by observing the intersection between the last two constraints, we have the following upper bound $R \le \frac{4}{\frac{4}{1-H(p_1)} + \frac{3}{1-H(p_2)}}$, which is achieved at $\tau_2^{(3)} = \frac{3(1-H(p_1))}{4(1-H(p_2)) + 3(1-H(p_1))}$. Consequently, an explicit upper bound for the retrieval rate is:

$$\max\left\{\frac{1-H(p_1)}{3}, \frac{2}{\frac{3}{1-H(p_1)} + \frac{1}{1-H(p_2)}}, \frac{4}{\frac{4}{1-H(p_1)} + \frac{3}{1-H(p_2)}}\right\} \tag{25}$$

In Section V-A, we will show how these rates can be achieved, hence (25) is the exact capacity. This capacity result is illustrated in Fig. 3. The figure shows the partitioning of the (p_1, p_2) (by convention $p_1 \le p_2$) space according to the active capacity expression. When the ratio $2 < \frac{1-H(p_1)}{1-H(p_2)} \le 3$, $C_{\text{PIR}}(p_1, p_2) = \frac{2}{\frac{3}{1-H(p_1)} + \frac{1}{1-H(p_2)}}$. When the ratio $\frac{1-H(p_1)}{1-H(p_2)} \le 2$, $C_{\text{PIR}}(p_1, p_2) = \frac{4}{\frac{4}{1-H(p_1)} + \frac{3}{1-H(p_2)}}$, otherwise, $C_{\text{PIR}}(p_1, p_2) = \frac{1-H(p_1)}{3}$. Interestingly, Fig. 3 shows that the dominant strategy for most (p_1, p_2) pairs is to rely only on database 1 for the retrieval process. The capacity function $C_{\text{PIR}}(p_1, p_2)$ is shown in Fig. 4. The figure shows that the maximum value for the capacity is $C_{\text{PIR}}(0, 0) = \frac{4}{7}$, which is consistent with [11]. The figure also shows that $C_{\text{PIR}}(0.5, 0.5) = 0$, as the answer strings become independent of the user queries. We observe that $C_{\text{PIR}}(0, p_2) = \frac{1}{3}$ for $p_2 \ge H^{-1}(\frac{2}{3}) = 0.1737$, since

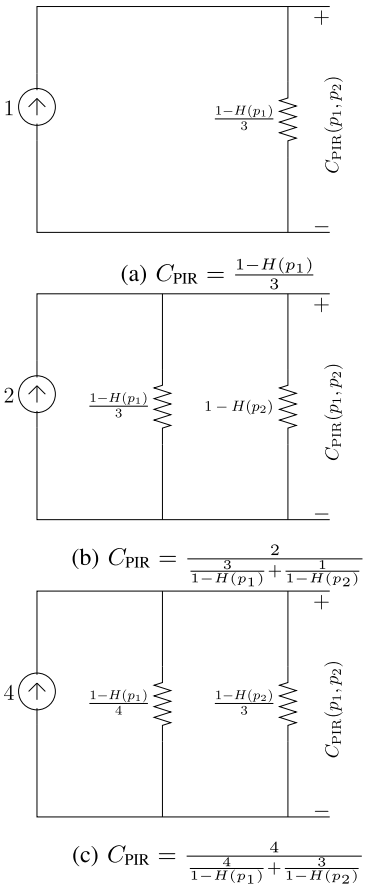

Fig. 2. Circuit analogy for the capacity expression of PIR from BSC(p_1), BSC(p_2).

the retrieval is performed only from database 1, which is connected to the user via a noiseless link.

Remark 6 *We will show in Section V that channel coding and retrieval schemes for NPIR are almost separable. Nevertheless, the final capacity expression couples the capacity of the noisy channels and the retrieval rates from databases with noiseless links in a non-trivial way. We illustrate the capacity expression in* (25) *by means of circuit theory analogy in Fig. 2. The current from the current source represents the number of desired bits, the voltage across the current source corresponds to the achievable retrieval rate, and the channel effect of the link connected to the nth database is abstracted via a parallel resistor, whose value depends on the capacity of the channel and the total download from the nth database. We note that the ratio between the denominators of the resistors corresponds to the ratio between the uncoded traffic (before applying the channel code) from the databases (namely, zero traffic from database 2 in Fig 2a,* $3:1$ *in Fig. 2b, and* $4:3$ *in Fig. 2c; see also the motivating example in Section V-A). Intuitively, to maximize the retrieval rate, the user chooses one of the three circuits in Fig. 2. The circuits are arranged ascendingly in the number of the desired bits (namely, 1, 2,* 4 *bits), while the values of the resistors decrease, as the total download increases and/or due to adding extra parallel branch. This results in a tension between conveying more desired bits and decreasing the equivalent resistor of the circuit. The capacity-achieving scheme is the one which maximizes the product of these contradictory effects (i.e., the voltage).*

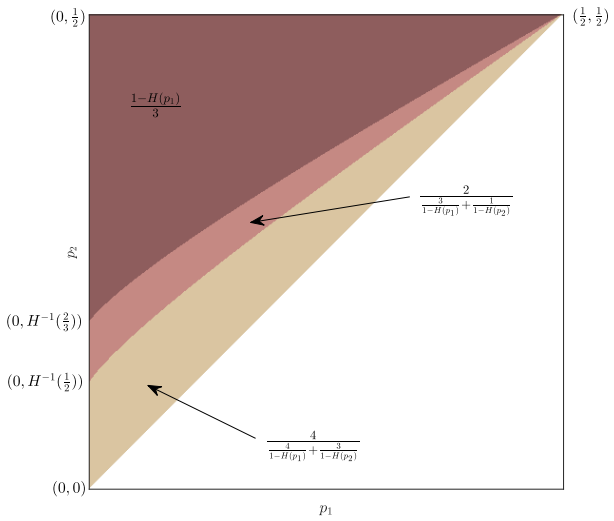

Fig. 3. Partitions of (p_1, p_2) space according to retrieval rate expression for $M = 3$, $N = 2$.

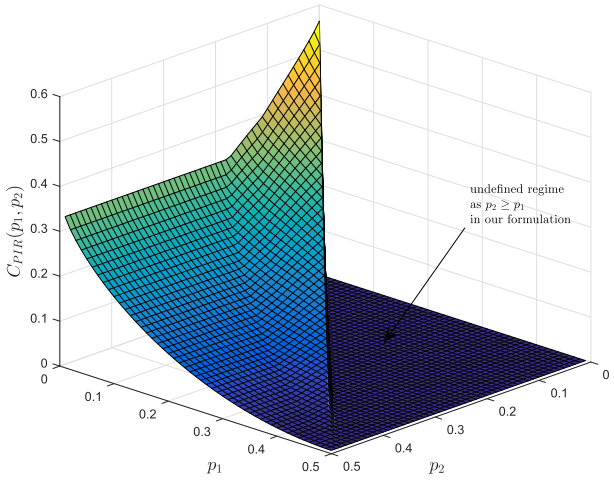

Fig. 4. Capacity function $C_{\text{PIR}}(p_1, p_2)$ for $M = 3$, $N = 2$.

IV. Converse Proof for NPIR

In this section, we derive a general upper bound for the NPIR problem. The main idea of the converse hinges on the fact that the traffic from the databases should be dependent on the relative channel qualities (i.e., channel capacities) of the response channels. Thus, we extend the converse proof in [38] to account for the noisy observations.

We will need the following lemma, which characterizes the channel effect on the noisy answer strings. The lemma states that the remaining uncertainty on a subset of answer strings after revealing the queries and the message set is a sum of single-letter conditional entropies of the noisy channels over the lengths of the answer strings. The lemma is a consequence of the Markov chain $(W_{1:M}, Q_{1:N}^{[m]}, \tilde{A}_{1:n-1}^{[m]}) \to A_n^{[m]} \to \tilde{A}_n^{[m]}$.

Lemma 1 (Channel effect) *For any subset $\mathcal{S} \subseteq \{1, \cdots, N\}$ for all $m \in \{1, \cdots, M\}$, the remaining uncertainty on the noisy answer strings $\tilde{A}_{\mathcal{S}}^{[m]}$ given $(W_{1:M}, Q_{1:N}^{[m]})$ is given by,*

$$H(\tilde{A}_{\mathcal{S}}^{[m]}|W_{1:M}, Q_{1:N}^{[m]}) = \sum_{n\in\mathcal{S}} \sum_{\eta_n=1}^{t_n} H\left(Y_{n,\eta_n}^{[m]}|X_{n,\eta_n}^{[m]}\right) \tag{26}$$

Furthermore, (26) *is true if conditioned on the complementary subset of the noisy answer strings $\tilde{A}_{\bar{\mathcal{S}}}^{[m]}$, i.e.,*

$$H(\tilde{A}_{\mathcal{S}}^{[m]}|W_{1:M}, Q_{1:N}^{[m]}, \tilde{A}_{\bar{\mathcal{S}}}^{[m]}) = \sum_{n\in\mathcal{S}} \sum_{\eta_n=1}^{t_n} H\left(Y_{n,\eta_n}^{[m]}|X_{n,\eta_n}^{[m]}\right) \tag{27}$$

where $\bar{\mathcal{S}} = \{1, \cdots, N\} \setminus \mathcal{S}$.

Proof: We start with the left hand side of (26),

$$\begin{aligned}
&H(\tilde{A}_{\mathcal{S}}^{[m]}|W_{1:M}, Q_{1:N}^{[m]}) \\
&= \sum_{n\in\mathcal{S}} H(\tilde{A}_n^{[m]}|\tilde{A}_{1:n-1}^{[m]}, W_{1:M}, Q_{1:N}^{[m]}) && (28)\\
&\overset{(4)}{=} \sum_{n\in\mathcal{S}} H(\tilde{A}_n^{[m]}|\tilde{A}_{1:n-1}^{[m]}, W_{1:M}, Q_{1:N}^{[m]}, A_n^{[m]}) && (29)\\
&= \sum_{n\in\mathcal{S}} H(\tilde{A}_n^{[m]}|A_n^{[m]}) && (30)\\
&= \sum_{n\in\mathcal{S}} \sum_{\eta_n=1}^{t_n} H(Y_{n,\eta_n}^{[m]}|X_{n,1}^{[m]}, \cdots, X_{n,t_n}^{[m]}, Y_{n,1}, \cdots, Y_{n,\eta_n-1}^{[m]}) && (31)\\
&\overset{(5)}{=} \sum_{n\in\mathcal{S}} \sum_{\eta_n=1}^{t_n} H(Y_{n,\eta_n}^{[m]}|X_{n,\eta_n}^{[m]}) && (32)
\end{aligned}$$

where (29) follows from the fact that $A_n^{[m]}$ is a deterministic function of $(W_{1:M}, Q_n^{[m]})$, (30) follows from the fact that $(W_{1:M}, Q_{1:N}^{[m]}, \tilde{A}_{1:n-1}^{[m]}) \to A_n^{[m]} \to \tilde{A}_n^{[m]}$ is a Markov chain, (32) follows from the fact that the channel is memoryless.

The proof of (27) follows similarly by observing that $(W_{1:M}, Q_{1:N}^{[m]}, \tilde{A}_{1:n-1}^{[m]}, \tilde{A}_{\bar{\mathcal{S}}}^{[m]}) \to A_n^{[m]} \to \tilde{A}_n^{[m]}$ is a Markov chain as well. ■

We need the following lemma which upper bounds the mutual information between the noisy answer strings and the interfering messages with a linear function of the channel capacities.

Lemma 2 (Noisy interference bound) *For NPIR, the mutual information between the interfering messages $W_{2:M}$ and the noisy answer strings $\tilde{A}_{1:N}^{[1]}$ given the desired message W_1 is upper bounded by,*

$$I\left(W_{2:M}; Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]}|W_1\right) \le \sum_{n=1}^{N} t_n C_n - L + o(L) \tag{33}$$

Proof: We start with the left hand side of (33),

$$\begin{aligned}
&I(W_{2:M}; Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]}|W_1) \\
&\overset{(2)}{=} I\left(W_{2:M}; W_1, Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]}\right) && (34)\\
&= I\left(W_{2:M}; Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]}\right) + I\left(W_{2:M}; W_1|Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]}\right) && (35)\\
&\overset{(9)}{\le} I\left(W_{2:M}; Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]}\right) + o(L) && (36)\\
&\overset{(3)}{=} I\left(W_{2:M}; \tilde{A}_{1:N}^{[1]}|Q_{1:N}^{[1]}\right) + o(L) && (37)\\
&= H\left(\tilde{A}_{1:N}^{[1]}|Q_{1:N}^{[1]}\right) - H\left(\tilde{A}_{1:N}^{[1]}|W_{2:M}, Q_{1:N}^{[1]}\right) + o(L) && (38)\\
&= H\left(\tilde{A}_{1:N}^{[1]}|Q_{1:N}^{[1]}\right) - H\left(\tilde{A}_{1:N}^{[1]}, W_1|W_{2:M}, Q_{1:N}^{[1]}\right) \\
&\quad + H\left(W_1|W_{2:M}, Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]}\right) + o(L) && (39)\\
&\overset{(9)}{\le} H\left(\tilde{A}_{1:N}^{[1]}|Q_{1:N}^{[1]}\right) - H\left(\tilde{A}_{1:N}^{[1]}, W_1|W_{2:M}, Q_{1:N}^{[1]}\right) + o(L) && (40)\\
&= H\left(\tilde{A}_{1:N}^{[1]}|Q_{1:N}^{[1]}\right) - H\left(W_1|W_{2:M}, Q_{1:N}^{[1]}\right) \\
&\quad - H\left(\tilde{A}_{1:N}^{[1]}|W_{1:M}, Q_{1:N}^{[1]}\right) + o(L) && (41)\\
&\overset{(26)}{\le} \sum_{n=1}^{N} \sum_{\eta_n=1}^{t_n} \left[H\left(Y_{n,\eta_n}^{[1]}\right) - H\left(Y_{n,\eta_n}^{[1]}|X_{n,\eta_n}^{[1]}\right)\right] - L + o(L) && (42)\\
&= \sum_{n=1}^{N} \sum_{\eta_n=1}^{t_n} I\left(X_{n,\eta_n}^{[1]}; Y_{n,\eta_n}^{[1]}\right) - L + o(L) && (43)\\
&\le \sum_{n=1}^{N} t_n C_n - L + o(L) && (44)
\end{aligned}$$

where (34) follows from the independence of the messages, (36), (40) follow from the decodability of W_1 given $(Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]})$, (37) follows from the independence of $(W_{2:M}, Q_{1:N}^{[1]})$, (42) follows from the independence of $(W_1, W_{2:M}, Q_{1:N}^{[1]})$, Lemma 1, and the fact that conditioning cannot increase entropy, (44) follows from the fact that $I\left(X_{n,\eta_n}^{[m]}; Y_{n,\eta_n}^{[m]}\right) \le C_n$ by the definition of the nth channel capacity. ■

Finally, in order to capture the recursive structure of the problem in terms of the messages and to express the potential asymmetry of the optimal scheme, we will need the following lemma, which inductively lower bounds the mutual information term in Lemma 2. The lemma implies that n_{m-1} databases can apply a symmetric scheme when the retrieval problem is reduced to retrieving message W_{m-1} from the set of $W_{m-1:M}$ messages. For the remaining answer strings, we directly bound them by their corresponding length of the unobserved portion $\sum_{n=n_{m-1}+1}^{N} t_n C_n$.

Lemma 3 (Noisy induction lemma) *For all $m \in \{2, \ldots, M\}$ and for an arbitrary $n_{m-1} \in \{1, \cdots, N\}$, the mutual information term in Lemma 2 can be inductively lower bounded as,*

$$\begin{aligned}
&I\left(W_{m:M}; Q_{1:N}^{[m-1]}, \tilde{A}_{1:N}^{[m-1]}|W_{1:m-1}\right) \\
&\quad \ge \frac{1}{n_{m-1}} I\left(W_{m+1:M}; Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]}|W_{1:m}\right) \\
&\quad\quad + \frac{1}{n_{m-1}} \left(L - \sum_{n=n_{m-1}+1}^{N} t_n C_n\right) - \frac{o(L)}{n_{m-1}} && (45)
\end{aligned}$$

Proof: We start with the left hand side of (45) after multiplying by n_{m-1},

$$
\begin{aligned}
&n_{m-1}\, I\left(W_{m:M};\, Q_{1:N}^{[m-1]}, \tilde{A}_{1:N}^{[m-1]} | W_{1:m-1}\right) \\
&\geq n_{m-1}\, I\left(W_{m:M};\, Q_{1:n_{m-1}}^{[m-1]}, \tilde{A}_{1:n_{m-1}}^{[m-1]} | W_{1:m-1}\right) && (46)\\
&\geq \sum_{n=1}^{n_{m-1}} I\left(W_{m:M};\, Q_{n}^{[m-1]}, \tilde{A}_{n}^{[m-1]} | W_{1:m-1}\right) && (47)\\
&\overset{(8)}{=} \sum_{n=1}^{n_{m-1}} I\left(W_{m:M};\, Q_{n}^{[m]}, \tilde{A}_{n}^{[m]} | W_{1:m-1}\right) && (48)\\
&\overset{(3)}{=} \sum_{n=1}^{n_{m-1}} I\left(W_{m:M};\, \tilde{A}_{n}^{[m]} | Q_{n}^{[m]}, W_{1:m-1}\right) && (49)\\
&= \sum_{n=1}^{n_{m-1}} H\left(\tilde{A}_{n}^{[m]} | Q_{n}^{[m]}, W_{1:m-1}\right) - H\left(\tilde{A}_{n}^{[m]} | Q_{n}^{[m]}, W_{1:M}\right) && (50)\\
&\geq \sum_{n=1}^{n_{m-1}} H\left(\tilde{A}_{n}^{[m]} | \tilde{A}_{1:n-1}^{[m]}, Q_{1:n_{m-1}}^{[m]}, W_{1:m-1}\right) \\
&\qquad - H\left(\tilde{A}_{n}^{[m]} | \tilde{A}_{1:n-1}^{[m]}, Q_{1:n_{m-1}}^{[m]}, W_{1:M}\right) && (51)\\
&= \sum_{n=1}^{n_{m-1}} I\left(W_{m:M};\, \tilde{A}_{n}^{[m]} | \tilde{A}_{1:n-1}^{[m]}, Q_{1:n_{m-1}}^{[m]}, W_{1:m-1}\right) && (52)\\
&= I\left(W_{m:M};\, \tilde{A}_{1:n_{m-1}}^{[m]} | Q_{1:n_{m-1}}^{[m]}, W_{1:m-1}\right) && (53)\\
&\overset{(3)}{=} I\left(W_{m:M};\, Q_{1:n_{m-1}}^{[m]}, \tilde{A}_{1:n_{m-1}}^{[m]} | W_{1:m-1}\right) && (54)\\
&\overset{(3),(4)}{=} I\left(W_{m:M};\, Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m-1}\right) \\
&\quad - I\left(W_{m:M};\, \tilde{A}_{n_{m-1}+1:N}^{[m]} | Q_{1:N}^{[m]}, \tilde{A}_{1:n_{m-1}}^{[m]}, W_{1:m-1}\right) && (55)\\
&= I\left(W_{m:M};\, Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m-1}\right) \\
&\quad - H\left(\tilde{A}_{n_{m-1}+1:N}^{[m]} | Q_{1:N}^{[m]}, \tilde{A}_{1:n_{m-1}}^{[m]}, W_{1:m-1}\right) \\
&\quad + H\left(\tilde{A}_{n_{m-1}+1:N}^{[m]} | Q_{1:N}^{[m]}, \tilde{A}_{1:n_{m-1}}^{[m]}, W_{1:M}\right) && (56)\\
&\overset{(27)}{\geq} I\left(W_{m:M};\, Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m-1}\right) \\
&\quad - \sum_{n=n_{m-1}+1}^{N} \sum_{\eta_n=1}^{t_n} \left[H\left(Y_{n,\eta_n}^{[m]}\right) - H\left(Y_{n,\eta_n}^{[m]} | X_{n,\eta_n}^{[m]}\right)\right] && (57)\\
&\overset{(9)}{\geq} I\left(W_{m:M};\, W_m, Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m-1}\right) \\
&\quad - \sum_{n=n_{m-1}+1}^{N} \sum_{\eta_n=1}^{t_n} I\left(X_{n,\eta_n}^{[m]};\, Y_{n,\eta_n}^{[m]}\right) - o(L) && (58)\\
&= I\left(W_{m:M};\, W_m | W_{1:m-1}\right) + I\left(W_{m:M};\, Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m}\right) \\
&\quad - \sum_{n=n_{m-1}+1}^{N} \sum_{\eta_n=1}^{t_n} I\left(X_{n,\eta_n}^{[m]};\, Y_{n,\eta_n}^{[m]}\right) - o(L) && (59)\\
&= I\left(W_{m+1:M};\, Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m}\right) \\
&\quad + \left(L - \sum_{n=n_{m-1}+1}^{N} \sum_{\eta_n=1}^{t_n} I\left(X_{n,\eta_n}^{[m]};\, Y_{n,\eta_n}^{[m]}\right)\right) - o(L) && (60)\\
&\geq I\left(W_{m+1:M};\, Q_{1:N}^{[m]}, \tilde{A}_{1:N}^{[m]} | W_{1:m}\right) \\
&\quad + \left(L - \sum_{n=n_{m-1}+1}^{N} t_n C_n\right) - o(L) && (61)
\end{aligned}
$$

where (46), (47) follow from the non-negativity of mutual information, (48) follows from the privacy constraint, (49) follows from the independence of $\left(W_{m:M}, Q_{n}^{[m]}\right)$, (51) follows from the fact that conditioning cannot increase entropy and from the fact that $(W_{1:M}, Q_{1:n_{m-1}}^{[m]}, \tilde{A}_{1:n-1}^{[m]}) \rightarrow (W_{1:M}, Q_{n}^{[m]}) \rightarrow \tilde{A}_{n}^{[m]}$ forms a Markov chain, (54) follows from the independence of the messages and the queries, (55) follows from the chain rule, the independence of the queries and the messages, and the fact that $Q_{1:N}^{[m]} \rightarrow Q_{1:n_{m-1}}^{[m]} \rightarrow \tilde{A}_{1:n_{m-1}}^{[m]}$ forms a Markov chain by (4), (57) follows from the fact that conditioning reduces entropy and Lemma 1, (58) follows from the reliability constraint, (61) follows from the definition of the channel capacity. Finally, dividing both sides by n_{m-1} leads to (45). ■

Now, we are ready to derive an explicit upper bound for the retrieval rate from noisy channels. Fixing the length of the nth answer string to t_n and applying Lemma 2 and Lemma 3 successively for an arbitrary sequence $\{n_i\}_{i=1}^{M-1} \subset \{1, \cdots, N\}^{M-1}$, we have the following,

$$
\begin{aligned}
&\sum_{n=1}^{N} t_n C_n - L + \tilde{o}(L) \\
&\overset{(33)}{\geq} I\left(W_{2:M};\, Q_{1:N}^{[1]}, \tilde{A}_{1:N}^{[1]} | W_1\right) && (62)\\
&\overset{(45)}{\geq} \frac{1}{n_1}\left(L - \sum_{n=n_1+1}^{N} t_n C_n\right) + \frac{1}{n_1} I\left(W_{3:M};\, Q_{1:N}^{[2]}, \tilde{A}_{1:N}^{[2]} | W_{1:2}\right) && (63)\\
&\overset{(45)}{\geq} \frac{1}{n_1}\left(L - \sum_{n=n_1+1}^{N} t_n C_n\right) + \frac{1}{n_1 n_2}\left(L - \sum_{n=n_2+1}^{N} t_n C_n\right) \\
&\quad + \frac{1}{n_2} I\left(W_{4:M};\, Q_{1:N}^{[3]}, \tilde{A}_{1:N}^{[3]} | W_{1:3}\right) && (64)\\
&\overset{(45)}{\geq} \cdots \\
&\overset{(45)}{\geq} \frac{1}{n_1}\left(L - \sum_{n=n_1+1}^{N} t_n C_n\right) + \frac{1}{n_1 n_2}\left(L - \sum_{n=n_2+1}^{N} t_n C_n\right) \\
&\quad + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}\left(L - \sum_{n=n_{M-1}+1}^{N} t_n C_n\right) && (65)
\end{aligned}
$$

where $\tilde{o}(L) = \left(1 + \frac{1}{n_1} + \frac{1}{n_1 n_2} + \cdots + \frac{1}{\prod_{i=1}^{M-1} n_i}\right) o(L)$, (62) follows from Lemma 2, and the remaining bounding steps follow from successive application of Lemma 3.

Ordering terms, we have,

$$\left(1+\frac{1}{n_1}+\frac{1}{n_1 n_2}+\cdots+\frac{1}{\prod_{i=1}^{M-1} n_i}\right)L \leq \left(\theta(0)+\frac{\theta(n_1)}{n_1}+\cdots+\frac{\theta(n_{M-1})}{\prod_{i=1}^{M-1} n_i}\right)\sum_{n=1}^{N} t_n+\tilde{o}(L) \quad (66)$$

where $\theta(\ell)=\sum_{n=\ell+1}^{N}\tau_n C_n$.

We conclude the proof by taking $L\to\infty$. Thus, for an arbitrary sequence $\{n_i\}_{i=1}^{M-1}$, we have

$$R(\boldsymbol{\tau},\mathbf{C})=\frac{L}{\sum_{n=1}^{N} t_n} \leq \frac{\theta(0)+\frac{\theta(n_1)}{n_1}+\frac{\theta(n_2)}{n_1 n_2}+\cdots+\frac{\theta(n_{M-1})}{\prod_{i=1}^{M-1} n_i}}{1+\frac{1}{n_1}+\frac{1}{n_1 n_2}+\cdots+\frac{1}{\prod_{i=1}^{M-1} n_i}} \quad (67)$$

Finally, we get the tightest bound by minimizing over the sequence $\{n_i\}_{i=1}^{M-1}$ over the set $\{1,\cdots,N\}$, as

$$R(\boldsymbol{\tau},\mathbf{C}) \leq \min_{n_i\in[N]} \frac{\theta(0)+\frac{\theta(n_1)}{n_1}+\frac{\theta(n_2)}{n_1 n_2}+\cdots+\frac{\theta(n_{M-1})}{\prod_{i=1}^{M-1} n_i}}{1+\frac{1}{n_1}+\frac{1}{n_1 n_2}+\cdots+\frac{1}{\prod_{i=1}^{M-1} n_i}} \quad (68)$$

$$= \min_{n_i\in[N]} \frac{\sum_{n=1}^{N}\tau_n C_n+\frac{\sum_{n=n_1+1}^{N}\tau_n C_n}{n_1}+\cdots+\frac{\sum_{n=n_{M-1}+1}^{N}\tau_n C_n}{\prod_{i=1}^{M-1} n_i}}{1+\frac{1}{n_1}+\cdots+\frac{1}{\prod_{i=1}^{M-1} n_i}} \quad (69)$$

The user and the databases can agree on a traffic ratio vector $\boldsymbol{\tau}\in\mathbb{T}=\{(\tau_1,\cdots,\tau_N):\tau_n\geq 0\ \forall n,\sum_{n=1}^{N}\tau_n=1\}$ that maximizes $R(\boldsymbol{\tau},\mathbf{C})$, hence the retrieval rate $R(\mathbf{C})$ is upper bounded by,

$$R(\mathbf{C})\leq \max_{\boldsymbol{\tau}\in\mathbb{T}} R(\boldsymbol{\tau},\mathbf{C}) \quad (70)$$

leading to the upper bound in Theorem 1.

V. Achievability Proof for NPIR

In this section, we present the achievability proof for the NPIR problem. We show that by means of the random coding argument, each database can independently encode its response such that the probability of error can be made vanishingly small. The databases use the uncoded responses as an indexing mechanism for choosing codewords from a randomly generated codebook. The uncoded responses, which are the truthful responses to the user queries, vary in length to maximize the retrieval rate. The query structure builds on the achievability proofs for PIR under asymmetric traffic constraints [38].

A. Motivating Example: $M=3$, $N=2$, via BSC(p_1), BSC(p_2)

We illustrate the retrieval scheme for $N=2$ databases, $M=3$ messages when the answer strings pass through BSC(p_1) and BSC(p_2). We show that the channel coding (using linear block codes) is *almost separable* from the retrieval scheme (which hinges on the result of [38]). We begin with the case when $(p_1,p_2)=(0.1,0.2)$, then we extend this technique for all (p_1,p_2) pairs. We will need the following lemma, which shows the achievability of Shannon's channel coding theorem for BSC using linear block codes [44, Theorem 4.17, Corollary 4.18].

Lemma 4 (Shannon's coding theorem for BSC [44]) *For BSC(p) with crossover probability $p\in(0,\frac{1}{2})$. Let n, k be integers such that $R=\frac{k}{n}<1-H(p)$, and let $\mathbb{E}_{\mathcal{C}}[P_e(\mathcal{C})]$ denote the expected probability of error $P_e(\mathcal{C})$ calculated over all linear $[n,k]$ codes $\mathcal{C}$, assuming a nearest-codeword decoder. Then,*

$$\mathbb{E}_{\mathcal{C}}[P_e(\mathcal{C})]<2\cdot 2^{-n\Delta(p,R)} \quad (71)$$

for some $\Delta(p,R)>0$. Moreover, for all $\rho\in(0,1]$, all but less than ρ of the linear $[n,k]$ codes satisfy,

$$P_e(\mathcal{C})<\frac{2}{\rho}\cdot 2^{-n\Delta(p,R)} \quad (72)$$

The result implies that as long as the rate of the linear $[n,k]$ code is strictly less than the capacity, then there exists a linear $[n,k]$ code with exponentially decreasing probability of error in n with high probability.

1) Achievable Scheme for BSC(0.1), BSC(0.2): Now, we focus on the case when $(p_1,p_2)=(0.1,0.2)$. Using the explicit upper bound in (25), we infer that $R\leq\frac{4}{\frac{4}{1-H(p_1)}+\frac{3}{1-H(p_2)}}$ which is 0.2183 for $p_1=0.1$, $p_2=0.2$. To operate at $\tau_2=\tau_2^{(3)}=\frac{3(1-H(p_1))}{4(1-H(p_2))+3(1-H(p_1))}$, we enforce the ratio between the uncoded traffic, i.e., before channel coding, to be 4 : 3. This results in coded traffic ratio of $\frac{4}{1-H(p_1)}:\frac{3}{1-H(p_2)}$, which appears in the denominator of the upper bound. Concurrently, this results in retrieving 4 desired bits per scheme repetition, which appears in the numerator.

To that end, the user repeats the following retrieval scheme for ν times. Each repetition of the scheme operates over blocks of $L^*=4$ bits from all messages $W_{1:3}$. The user permutes the indices of the bits of each message independently and uniformly. Let $a_i(j)$, $b_i(j)$, $c_i(j)$ denote the ith bit of block j from the permuted message W_1, W_2, W_3, respectively. Assume without loss of generality that the desired file is W_1. In block j, the user requests to download a single bit from each message from database 1, i.e., the user requests to download $a_1(j)$, $b_1(j)$, and $c_1(j)$ from database 1. From database 2, the user exploits the side information generated from database 1 by requesting to download the sums $a_2(j)+b_1(j)$, $a_3(j)+c_1(j)$, and $b_2(j)+c_2(j)$. Finally, the user exploits the side information generated from database 1 by downloading $a_4(j)+b_2(j)+c_2(j)$ from database 2. The query table for the jth block is summarized in Table I. Denote the number of uncoded bits requested from the nth database by D_n, then $D_1=4$, $D_2=3$. This guarantees that the ratio between the uncoded traffic is 4 : 3 (for any number of repetitions ν). This query structure is private, as all combinations of the sums are included in the queries and the indices of the message bits are uniformly and independently permuted for each block of

TABLE I

THE QUERY TABLE FOR THE jTH BLOCK OF $M = 3$, $N = 2$, $p_1 = 0.1$, $p_2 = 0.2$

Database 1	Database 2
$a_1(j)$	$a_2(j) + b_1(j)$
$b_1(j)$	$a_3(j) + c_1(j)$
$c_1(j)$	$b_2(j) + c_2(j)$
$a_4(j) + b_2(j) + c_2(j)$	

messages (which operate on different set of bits), the privacy constraint is satisfied.

After receiving the queries of the user, the nth database concatenates the uncoded binary answer strings into a vector $U_n^{[1]}$ of length νD_n, i.e.,

$$U_1^{[1]} = [a_1(1) \quad b_1(1) \quad c_1(1) \quad a_4(1) + b_2(1) + c_2(1) \quad \cdots \quad a_1(\nu) \quad b_1(\nu) \quad c_1(\nu) \quad a_4(\nu) + b_2(\nu) + c_2(\nu)]^T \tag{73}$$

$$U_2^{[1]} = [a_2(1) + b_1(1) \quad a_3(1) + c_1(1) \quad b_2(1) + c_2(1) \quad \cdots \quad a_2(\nu) + b_1(\nu) \quad a_3(\nu) + c_1(\nu) \quad b_2(\nu) + c_2(\nu)]^T \tag{74}$$

The nth database encodes the vector $U_n^{[1]}$ to a coded answer string $A_n^{[1]}$ of length t_n using a $(t_n, \nu D_n)$ linear block code (which belongs to the set of good codes that satisfy (72)) such that:

$$t_n = \left\lceil \frac{\nu D_n}{1 - H(p_n)} \right\rceil \tag{75}$$

This ensures that $\frac{\nu D_n}{t_n} < 1 - H(p_n)$. The nth database responds with $A_n^{[1]}$ via the noisy channel BSC(p_n). The user receives the noisy answer string $\tilde{A}_n^{[1]}$ from the nth database.

To perform the decoding, the user employs the nearest-codeword decoder to find an estimate of $A_n^{[1]}$ based on the observation $\tilde{A}_n^{[1]}$. Since $\frac{\nu D_n}{t_n} < 1 - H(p_n)$, using Lemma 4 and the union bound, the probability of error in decoding is upper bounded by:

$$P_e(L) \leq P_e(\mathcal{C}_1) + P_e(\mathcal{C}_2) \tag{76}$$

$$\leq \frac{2}{\rho} \left[2^{-t_1 \Delta\left(p_1, \frac{\nu D_1}{t_1}\right)} + 2^{-t_2 \Delta\left(p_2, \frac{\nu D_2}{t_2}\right)} \right] \tag{77}$$

As $\nu \to \infty$, $L \to \infty$ and $t_n \to \infty$, we have $P_e(L) \to 0$. This ensures the decodability of $U_n^{[1]}$ with high probability. Since the vectors $U_1^{[1]}$, $U_2^{[2]}$ are designed to exploit the side information, the user can cancel the effect of the undesired messages and be left only with the correct W_1 with probability of error $P_e(L)$. This satisfies the reliability constraint.

Finally, we calculate the achievable retrieval rate. The retrieval scheme decodes $L = \nu L^* = 4\nu$ bits from the desired messages. The retrieval scheme downloads $t_n = \left\lceil \frac{\nu D_n}{1 - H(p_n)} \right\rceil$ bits from the nth database, hence as $\nu \to \infty$, we have

$$R = \frac{L}{t_1 + t_2} \tag{78}$$

$$= \frac{\nu L^*}{\frac{\nu D_1}{1 - H(p_1)} + \frac{\nu D_2}{1 - H(p_2)}} \tag{79}$$

$$= \frac{4}{\frac{4}{1 - H(p_1)} + \frac{3}{1 - H(p_2)}} = 0.2183 \tag{80}$$

which matches the upper bound.

TABLE II

THE QUERY TABLE FOR THE jTH BLOCK OF $M = 3$, $N = 2$ TO ACHIEVE $R = \frac{2}{\frac{3}{1-H(p_1)} + \frac{1}{1-H(p_2)}}$

Database 1	Database 2
$a_1(j), b_1(j), c_1(j)$	$a_2(j) + b_1(j) + c_1(j)$

2) Achieving the Upper Bound for Arbitrary (p_1, p_2)*:* Now, we show that the upper bound in (25) is achievable for any (p_1, p_2). The idea is to design the uncoded response vectors $U_1^{[1]}$, $U_2^{[2]}$ such that the ratio of their traffic matches one of the corner points of the PIR problem under asymmetric traffic constraints [38].

a) For $R = \frac{1 - H(p_1)}{3}$*:* For this rate, the user requests to download from database 1 only and does not access database 2. Thus, the user downloads all the contents of database 1 to satisfy the privacy constraint. Specifically, the user downloads $a_1(j), b_1(j), c_1(j)$ at the jth block of the retrieval process. Database 1 encodes the responses $U_1^{[1]}$ into t_1-length answer string using $(t_1, \nu D_1)$, where $D_1 = 3$, and $t_1 = \left\lceil \frac{\nu D_1}{1 - H(p_1)} \right\rceil$. The user decodes ν desired symbols from ν repetitions with vanishingly small probability of error. Consequently, $R = \frac{1 - H(p_1)}{3}$.

b) For $R = \frac{2}{\frac{3}{1-H(p_1)} + \frac{1}{1-H(p_2)}}$*:* For this rate, the user designs the queries such that the traffic ratio between the uncoded responses is $3 : 1$. Thus, in the jth block, the user requests to download one bit from each message, i.e., the user requests to download $a_1(j), b_1(j), c_1(j)$ from database 1. The user mixes the undesired information obtained from database 1 into one combined symbol $b_1(j) + c_1(j)$ and uses this symbol as a side information in database 2 by requesting to download $a_2(j) + b_1(j) + c_1(j)$. The query table for the jth block of the scheme is depicted in Table II.

After repeating the retrieval process ν times, database 1 encodes the responses using a linear $(t_1, \nu D_1) = \left(\left\lceil \frac{3\nu}{1 - H(p_1)} \right\rceil, 3\nu \right)$ code, while database 2 encodes its responses using a linear $(t_2, \nu D_2) = \left(\left\lceil \frac{\nu}{1 - H(p_2)} \right\rceil, \nu \right)$ code. Using Lemma 4, the user can decode the correct W_1 with vanishingly small probability of error. The user decodes $L = 2\nu$ bits from W_1, hence, as $\nu \to \infty$

$$R = \frac{L}{t_1 + t_2} = \frac{2}{\frac{3}{1 - H(p_1)} + \frac{1}{1 - H(p_2)}} \tag{81}$$

c) For $R = \frac{4}{\frac{4}{1-H(p_1)}+\frac{3}{1-H(p_2)}}$*:* An instance for this scheme is the $(p_1, p_2) = (0.1, 0.2)$ example. Please refer to Section V-A.1 for the details.

Therefore, the capacity of the PIR problem from BSC(p_1), BSC(p_2) is given by:

$$C_{\text{PIR}}(p_1, p_2) = \max\left\{\frac{1-H(p_1)}{3}, \frac{2}{\frac{3}{1-H(p_1)}+\frac{1}{1-H(p_2)}}, \frac{4}{\frac{4}{1-H(p_1)}+\frac{3}{1-H(p_2)}}\right\}$$

B. General Achievable Scheme

In this section, we present a general achievable scheme for the NPIR problem. The main idea of the scheme is to use the uncoded response from the nth database to user's query as an *index* for choosing the transmitted codeword from a codebook generated according to the optimal probability distribution. The query structure maps to one of the corner points of PIR under asymmetric traffic constraints [38] in order to maximize the retrieval rate.

Following the notations in [38], we denote the number of side information symbols that are used simultaneously in the initial round of downloads at the nth database by $s_n \in \{0, 1, \cdots, M-1\}$, e.g., if $s_n = 1$, then the user requests to download a sum of 1 desired symbol and 1 undesired symbol as a side information in the form of $a + b$, $a + c$,... etc., while $s_n = 2$ implies that the user mixes every two undesired symbols to form one side information symbol, i.e., the user requests to download $a+b+c$, $a+c+d$,... etc. For a given non-decreasing sequence $\{n_i\}_{i=0}^{M-1} \subset \{1, \cdots, N\}^M$, the databases are divided into groups, such that group 0 contains database 1 through database n_0, group 1 contains $n_1 - n_0$ databases starting from database $n_0 + 1$, and so on.

Hence, let $s_n = i$ for all $n_{i-1}+1 \le n \le n_i$ with $n_{-1} = 0$ by convention. Denote $\mathcal{S} = \{i : s_n = i \text{ for some } n \in \{1, \cdots, N\}\}$. We follow the round and stage definitions in [21]. The kth round is the download queries that admit a sum of k different messages (k-sum in [11]). A stage of the kth round is a query block of the kth round that exhausts all $\binom{M}{k}$ combinations of the k-sum. Denote $y_\ell[k]$ to be the number of stages in round k downloaded from the nth database, such that $n_{\ell-1} + 1 \le n \le n_\ell$. Our scheme is repeated for ν repetitions. Each repetition has the same query structure and operates over a block of message symbols of length L^*. Denote the total requested symbols from the nth database in one repetition of the scheme by $D_n(\mathbf{n})$. The details of the achievable scheme are as follows:

1) *Codebook construction:* According to the optimal probability distribution $p^*(x_n)$ (that maximizes the mutual information $I(X_n; Y_n)$), the nth database constructs a $\left(2^{\nu D_n(\mathbf{n})}, t_n(\mathbf{n})\right)$ codebook $\mathcal{C}_n$ at random[4], i.e., $p(x_{n,1}, \cdots, x_{n,t_n(\mathbf{n})}) = \prod_{\eta_n=1}^{t_n(\mathbf{n})} p^*(x_{n,\eta_n})$. Specifically, the codebook $\mathcal{C}_n$ can be written in the form of a $2^{\nu D_n(\mathbf{n})} \times t_n(\mathbf{n})$ matrix as:

$$\begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_{t_n(\mathbf{n})}(1) \\ x_1(2) & x_2(2) & \cdots & x_{t_n(\mathbf{n})}(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(2^{\nu D_n(\mathbf{n})}) & \cdots & \cdots & x_{t_n(\mathbf{n})}(2^{\nu D_n(\mathbf{n})}) \end{bmatrix} \tag{82}$$

where

$$t_n(\mathbf{n}) = \left\lceil \frac{\nu D_n(\mathbf{n})}{C_n} \right\rceil \tag{83}$$

This ensures that the rate of $\mathcal{C}_n$, $\frac{\nu D_n(\mathbf{n})}{t_n(\mathbf{n})} < C_n$ to ensure reliable transmission over the noisy channel. The nth database reveals the codebook $\mathcal{C}_n$ to the user.

2) *Initialization at the user side:* The user permutes each message independently and uniformly using a random interleaver, i.e.,
$$\omega_m(i) = W_m(\pi_m(i)), \quad i \in \{1, \cdots, L\} \tag{84}$$
where $\omega_m(i)$ is the ith symbol of the permuted W_m, $\pi_m(\cdot)$ is a random interleaver for the mth message that is chosen independently, uniformly, and privately at the user's side.
3) *Initial download:* From the nth database where $1 \le n \le n_0$, the user requests to download $\prod_{s\in\mathcal{S}} \binom{M-2}{s-1}$ symbols from the desired message. The user sets the round index $k = 1$. I.e., the user requests the desired symbols from $y_0[1] = \prod_{s\in\mathcal{S}} \binom{M-2}{s-1}$ different stages.
4) *Message symmetry:* To satisfy the privacy constraint, for each stage initiated in the previous step, the user completes the stage by requesting the remaining $\binom{M-1}{k-1}$ k-sum combinations that do not include the desired symbols, in particular, if $k = 1$, the user requests $\prod_{s\in\mathcal{S}} \binom{M-2}{s-1}$ individual symbols from each undesired message.
5) *Database symmetry:* We divide the databases into groups. Group $\ell \in \mathcal{S}$ corresponds to databases $n_{\ell-1} + 1$ to n_ℓ. Database symmetry is applied within each group only. Consequently, the user repeats step 2 over each group of databases, in particular, if $k = 1$, the user downloads $\prod_{s\in\mathcal{S}} \binom{M-2}{s-1}$ individual symbols from each message from the first n_0 databases (group 1).
6) *Exploitation of side information:* The undesired symbols downloaded within the kth round (the k-sums that do not include the desired message) are used as side information in the $(k+1)$th round. This exploitation of side information is performed by requesting to download $(k+1)$-sum consisting of 1 desired symbol and a k-sum of undesired symbols only that were generated in the kth round. Note that for the nth database, if $s_n > k$, then this database does not exploit the side information generated in the kth round. Consequently, the nth database belonging to the ℓth group exploits the side information generated in the kth round from all databases except itself if $s_n \le k$. Moreover, for $s_n = k$, extra side information can be used in the nth database. This is due to the fact that the user can form $n_0 \prod_{s\in\mathcal{S}\setminus\{s_n\}} \binom{M-2}{s-1}$ extra stages of side information by constructing k-sums of the undesired symbols in round 1 from the databases in group 0.
7) *Repeat* steps 4, 5, 6 after setting $k = k + 1$ until $k = M$.

[4]We note that the databases should have the knowledge about $\mathbf{n}$ (or equivalently the traffic ratio vector $\boldsymbol{\tau}$). This information should be conveyed in the queries. The user and the databases exchange the codebooks for every $\mathbf{n}$ prior to the retrieval process.

8) *Repetition of the scheme:* Repeat steps $3, \cdots, 7$ for a total of ν repetitions.
9) *Shuffling the order of the queries:* By shuffling the order of the queries uniformly, all possible queries can be made equally likely regardless of the message index. This guarantees the privacy.
10) *Encoding the responses to the user's queries:* The nth database responds to the user queries truthfully. The nth database concatenates all the responses to the user's queries in a vector $U_n^{[i]}$ of length $\nu D_n(\mathbf{n})$. The nth database uses $U_n^{[i]}$ as an index for choosing a codeword from $\mathcal{C}_n$, i.e., the index of the codeword and $U_n^{[i]}$ should be in bijection (e.g., by transforming $U_n^{[i]}$ into a decimal value). Consequently, the nth database responds with,

$$A_n^{[i]} = [x_1(U_n^{[i]}) \quad x_1(U_n^{[i]}) \quad \cdots \quad x_{t_n(\mathbf{n})}(U_n^{[i]})]^T \tag{85}$$

C. Privacy, Reliability, and Achievable Rate

a) Privacy: The privacy of the scheme follows from the privacy of the inherent PIR scheme under asymmetric traffic constraints. Specifically, for every stage of the kth round initiated, all $\binom{M}{k}$ combinations of the k-sum are included at each round. Thus, the structure of the queries is the same for any desired message at any repetition of the achievable scheme. Due to the random and independent permutation of each message and the random shuffling of the order of the queries, all queries are equally likely independent of the desired message index, and thus the privacy constraint in (8) is guaranteed.

b) Reliability: The user employs *joint typicality decoder*[5] for every noisy answer string $\tilde{A}_n^{[i]}$ to decode the codeword index. From the channel coding theorem [45, Theorem 7.7.1], for every rate $\frac{\nu D_n(\mathbf{n})}{t_n(\mathbf{n})} < C_n$, there exists a sequence of $(2^{\nu D_n(\mathbf{n})}, t_n(\mathbf{n}))$ with maximum probability of error $P_e(\mathcal{C}_n) \to 0$ as $t_n(\mathbf{n}) \to \infty$. By letting $\nu \to \infty$, we have $t_n(\mathbf{n}) \to \infty$, $\frac{\nu D_n(\mathbf{n})}{t_n(\mathbf{n})} < C_n$ and hence we ensure the existence of a good code such that $P_e(\mathcal{C}_n) \to 0$. By union bound, the probability of error in decoding the indices of the codewords from every database is upper bounded by $P_e \leq \sum_{n=1}^N P_e(\mathcal{C}_n) \to 0$.

Since the index of the codeword is bijective to $U_n^{[i]}$, the probability of error in decoding $U_n^{[i]}$ for $n = 1, \cdots, N$ is vanishingly small. Now, by construction of the queries as in [38], all side information symbols used in the $(k+1)$th round are decodable in the kth round or from round 1, the user cancels out these side information and is left with symbols from the desired message. Consequently, there is no error in the decoding given that $U_n^{[i]}$ is correct for every n.

c) Achievable Rate: The structure of one repetition of our scheme is exactly as [38]. The recursive structure is described using the following system of difference equations that relate the number of stages in the databases belonging to a specific group as shown in [38, Theorem 2]:

$$\begin{aligned}
y_0[k] &= (n_0-1)y_0[k-1] + \sum_{j\in\mathcal{S}\setminus\{0\}} (n_j - n_{j-1})y_j[k-1] \\
y_1[k] &= (n_1-n_0-1)y_1[k-1] + \sum_{j\in\mathcal{S}\setminus\{1\}} (n_j - n_{j-1})y_j[k-1] \\
y_\ell[k] &= n_0\xi_\ell\delta[k-\ell-1] + (n_\ell - n_{\ell-1}-1)y_\ell[k-1] \\
&\quad + \sum_{j\in\mathcal{S}\setminus\{\ell\}} (n_j - n_{j-1})y_j[k-1], \quad \ell \geq 2
\end{aligned} \tag{86}$$

where $y_\ell[k]$ is the number of stages in the kth round in a database belonging to the ℓth group, i.e., for the nth database, such that $n_{\ell-1}+1 \leq n \leq n_\ell$.

To calculate $D_n(\mathbf{n})$ where $n_{\ell-1} \leq n \leq n_\ell$, we note that for any stage in the kth round, the user downloads $\binom{M-1}{k-1}$ desired symbols from a total of $\binom{M}{k}$ downloads. Therefore,

$$D_n(\mathbf{n}) = \sum_{k=1}^M \binom{M}{k} y_\ell[k], \quad n_{\ell-1} \leq n \leq n_\ell \tag{87}$$

Thus, the total download $\sum_{n=1}^N t_n(\mathbf{n})$ from all databases from all repetitions is calculated by observing (83) and ignoring the ceiling operator as $\nu \to \infty$,

$$\sum_{n=1}^N t_n(\mathbf{n}) = \sum_{n=1}^N \frac{\nu D_n(\mathbf{n})}{C_n} \tag{88}$$

$$= \nu \left[\sum_{n=1}^{n_0} \frac{\sum_{k=1}^M \binom{M}{k} y_0[k]}{C_n} + \sum_{n=n_0+1}^{n_1} \frac{\sum_{k=1}^M \binom{M}{k} y_1[k]}{C_n} + \cdots \right] \tag{89}$$

$$= \nu \sum_{\ell\in\mathcal{S}} \sum_{n=n_{\ell-1}+1}^{n_\ell} \frac{\sum_{k=1}^M \binom{M}{k} y_\ell[k]}{C_n} \tag{90}$$

Furthermore, the total desired symbols from all databases from all repetitions is given by,

$$L(\mathbf{n}) = \nu \sum_{\ell\in\mathcal{S}} \sum_{k=1}^M \binom{M-1}{k-1} y_\ell[k](n_\ell - n_{\ell-1}) \tag{91}$$

Consequently, the following rate is achievable corresponding to the sequence $\mathbf{n}$,

$$R(\mathbf{n}, \mathbf{C}) = \frac{\sum_{\ell\in\mathcal{S}} \sum_{k=1}^M \binom{M-1}{k-1} y_\ell[k](n_\ell - n_{\ell-1})}{\sum_{\ell\in\mathcal{S}} \sum_{n=n_{\ell-1}+1}^{n_\ell} \frac{\sum_{k=1}^M \binom{M}{k} y_\ell[k]}{C_n}} \tag{92}$$

Since this scheme is achievable for every monotone non-decreasing sequence $\mathbf{n} = \{n_i\}_{i=0}^{M-1}$, the following rate is achievable,

$$R(\mathbf{C}) = \max_{n_i\in[N]} \frac{\sum_{\ell\in\mathcal{S}} \sum_{k=1}^M \binom{M-1}{k-1} y_\ell[k](n_\ell - n_{\ell-1})}{\sum_{\ell\in\mathcal{S}} \sum_{n=n_{\ell-1}+1}^{n_\ell} \frac{\sum_{k=1}^M \binom{M}{k} y_\ell[k]}{C_n}} \tag{93}$$

VI. PIR from Multiple Access Channel

In this section, we consider the MAC-PIR problem. This problem is an extension of the NPIR model presented in Section II which consists of N non-colluding and replicated databases storing M messages. In MAC-PIR (see Fig. 5),

[5]We note that the random coding arguments presented here for the general case can be readily applied to any noisy channel including the one in the motivating example, which is BSC. We choose to use the linear block coding argument for BSC to simplify the presentation of the example.

the user sends a query $Q_n^{[i]}$ for the nth database to retrieve W_i privately and correctly. The nth database responds with an answer string $A_n^{[i]} = (X_{n,1}^{[i]}, \cdots, X_{n,t}^{[i]})$. The user receives a noisy observation $\tilde{A}_n^{[i]} = (Y_1^{[i]}, \cdots, Y_t^{[i]})$, where the responses of the databases $(A_1^{[i]}, A_2^{[i]}, \cdots, A_N^{[i]})$ pass through a discrete memoryless channel with a transition probability distribution $p(y|x_1, \cdots, x_N)$, i.e.,

$$P\left(\tilde{A}^{[i]}|A_1^{[i]}, A_2^{[i]}, \cdots, A_N^{[i]}\right) = \prod_{\eta=1}^{t} p\left(y_\eta^{[i]}|x_{1,\eta}^{[i]}, x_{2,\eta}^{[i]}, \cdots, x_{N,\eta}^{[i]}\right) \tag{94}$$

In this sense, the retrieval is performed via a *cooperative multiple access channel*, as the databases cooperate to convey the message W_i to a common receiver (the user). The full cooperation is realized via the user queries. Furthermore, in MAC-PIR, the database responses are mixed together to have the noisy observation $\tilde{A}^{[i]}$ in contrast to the noisy PIR problem with orthogonal links presented in Section II.

In MAC-PIR, the user should be able to reconstruct W_i with vanishingly small probability of error by observing the noisy and mixed output $\tilde{A}^{[i]}$, i.e., the reliability constraint is written as:

$$H(W_i|Q_{1:N}^{[i]}, \tilde{A}^{[i]}) \leq o(L) \tag{95}$$

and the privacy constraint is written as:

$$(Q_n^{[i]}, A_n^{[i]}, W_{1:M}) \sim (Q_n^{[j]}, A_n^{[j]}, W_{1:M}), \quad \forall i, j \in \{1, \cdots, M\} \tag{96}$$

We observe that we cannot claim that $\tilde{A}^{[i]} \sim \tilde{A}^{[j]}$ in the MAC-PIR problem as we claimed for the NPIR problem. This is due to the fact that the user cannot statistically differentiate between the responses corresponding to each message and hence the user cannot decode the desired message. This is in contrast to the NPIR problem with orthogonal links, where $\tilde{A}^{[i]} \sim \tilde{A}^{[j]}$ due to the Markov chain $(W_{1:M}, Q_n^{[i]}) \rightarrow A_n^{[i]} \rightarrow \tilde{A}_n^{[i]}$.

The retrieval rate for the MAC-PIR is given by:

$$R = \frac{L}{t} \tag{97}$$

and the MAC-PIR capacity is $C_{\text{PIR}} = \sup R$ over all retrieval schemes. We note that, without loss of generality, we can assume that all responses from the databases have the same length t in contrast to the NPIR problem with orthogonal links. The reason is that the retrieval rate depends only on the output of the channel and not on the individual responses of the databases. Hence, even if the database responses are different in lengths, we can choose $t = \max_{n\in[N]} t_n$ by appending the remaining responses by dummy symbols.

In the sequel, we discuss the issue of separability of channel coding and the information retrieval in MAC-PIR via some examples. Interestingly, we show that the optimal PIR scheme for the additive MAC and logic conjunction/disjunction MAC, the channel coding and the retrieval scheme are dependent on the channel transition probability, and hence channel coding and retrieval procedure are inseparable.

A. Additive MAC

In the first special case, we consider the additive MAC. In the additive MAC, at each time instant η, the responses of the databases are added together (in modulo-2) in addition to a random variable $Z_\eta \sim$ Bernoulli(p), which is independent of $(W_{1:M}, Q_{1:N}^{[i]})$ and corresponds to a random additive noise, i.e.,

$$Y_\eta = \sum_{n=1}^{N} X_{n,\eta} + Z_\eta \tag{98}$$

The following theorem characterizes the capacity of the MAC-PIR problem if the channel is restricted to additive MACs.

Theorem 3 *The additive MAC-PIR capacity is,*

$$C_{PIR} = 1 - H(p) \tag{99}$$

where $p \in [0, 0.5)$ is the flipping probability of the additive noise.

We have the following remarks.

Remark 7 *For noiseless additive MAC, i.e., $p = 0$ and $Y_\eta = \sum_{n=1}^{N} X_{n,\eta}$, the MAC-PIR capacity is $C_{PIR} = 1$. This implies that there is no penalty due to the privacy constraint, i.e., the user can have privacy for free. Interestingly, this is the first instance where the PIR capacity is independent of the number of databases N and the number of messages M.*

Remark 8 *For noiseless additive MAC, i.e., $p = 0$, separation between channel coding and retrieval process is not optimal unlike the NPIR problem with orthogonal links. In fact, the retrieval scheme is dependent on the structure of the channel. To see this, the user generates a random binary vector $\mathbf{h} = [h_1 h_2 \cdots h_M] \in \{0, 1\}^M$. The user sends $\mathbf{h}$ to database 1, flips the ith position of $\mathbf{h}$ and sends it to database 2, and does not send anything to the remaining databases. Thus, the responses of the databases are,*

$$A_1^{[i]} = \sum_{m=1}^{M} h_m W_m \tag{100}$$

$$A_2^{[i]} = \sum_{m=1}^{M} h_m W_m + W_i \tag{101}$$

This is exactly the retrieval scheme in [1]. Since the channel is additive and noiseless, $\tilde{A}^{[i]} = A_1^{[i]} + A_2^{[i]} = W_i$. Hence, the user downloads 1 bit from the channel in order to get 1 bit from the desired file and $R = 1$. Here, we note that, the channel performs the processing at the user for free. This implies that by careful design of queries, the user can exploit the channel in its favor to maximize the retrieval rate.

Proof: We prove the converse and achievability.

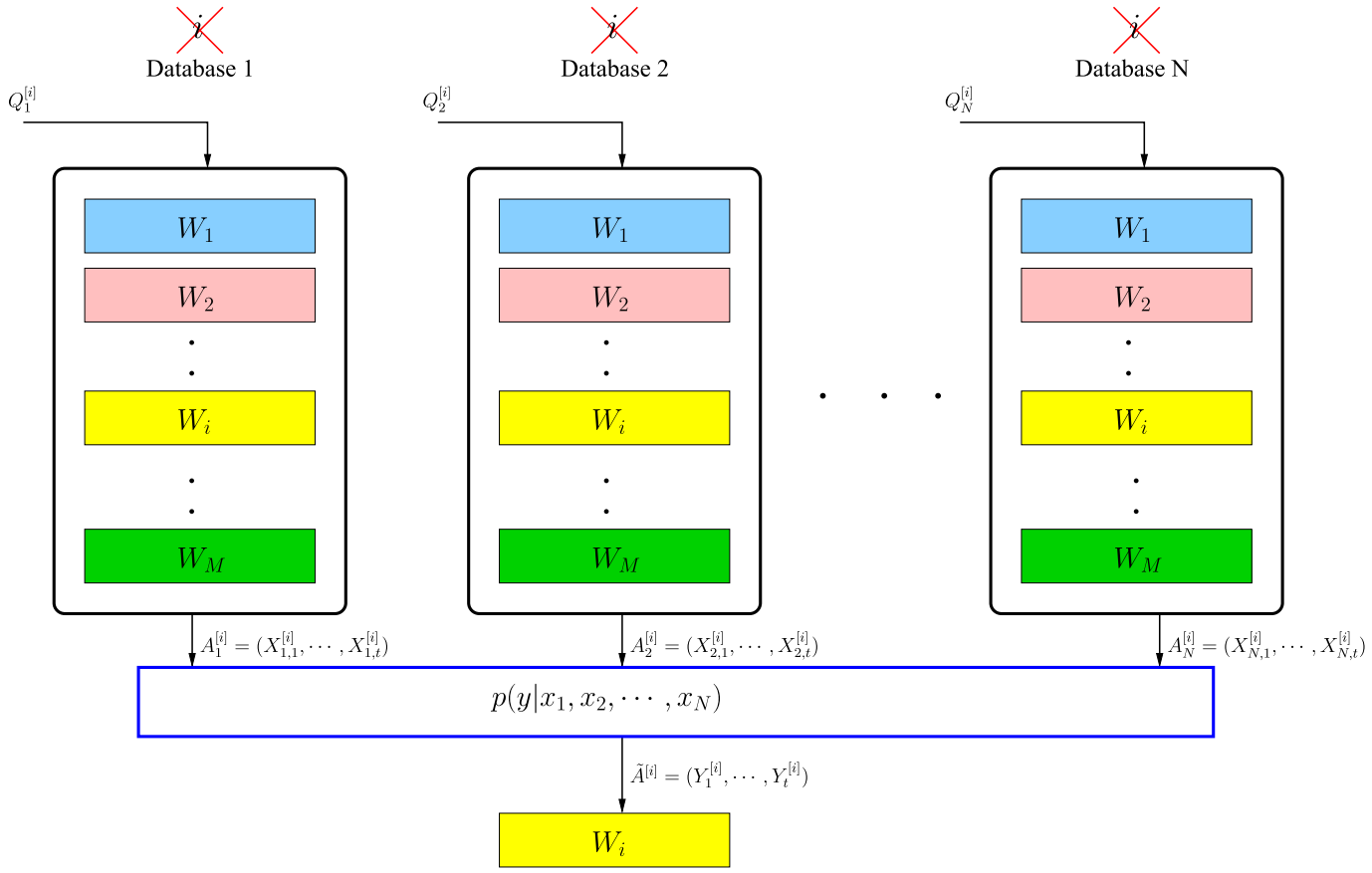

Fig. 5. The MAC-PIR problem.

a) The converse proof: To show the converse, we assume that W_1 is the desired message without loss of generality. Then, we have the following implications,

$$
\begin{aligned}
L &= H(W_1) && (102)\\
&\overset{(2),(3)}{=} H(W_1|W_{2:M}, Q_{1:N}^{[1]}) && (103)\\
&\overset{(95)}{\leq} H(W_1|W_{2:M}, Q_{1:N}^{[1]}) \\
&\quad - H(W_1|W_{2:M}, Q_{1:N}^{[1]}, \tilde{A}^{[1]}) + o(L) && (104)\\
&= I(W_1; \tilde{A}^{[1]}|Q_{1:N}^{[1]}, W_{2:M}) + o(L) && (105)\\
&= H(\tilde{A}^{[1]}|Q_{1:N}^{[1]}, W_{2:M}) \\
&\quad - H(\tilde{A}^{[1]}|Q_{1:N}^{[1]}, W_{1:M}) + o(L) && (106)\\
&\overset{(4)}{\leq} H(\tilde{A}^{[1]}) - H(\tilde{A}^{[1]}|Q_{1:N}^{[1]}, W_{1:M}, A_{1:N}^{[1]}) + o(L) && (107)\\
&= t - H(\tilde{A}^{[1]}|A_{1:N}^{[1]}) + o(L) && (108)\\
&= t - \sum_{\eta=1}^{t} H(Y_\eta^{[1]}|X_{1,\eta}^{[1]}, X_{2,\eta}^{[1]}, \cdots, X_{N,\eta}^{[1]}) + o(L) && (109)\\
&= t - \sum_{\eta=1}^{t} H\left(\sum_{n=1}^{N} X_{n,\eta}^{[1]} + Z_\eta | X_{1,\eta}^{[1]}, X_{2,\eta}^{[1]}, \cdots, X_{N,\eta}^{[1]}\right) \\
&\quad + o(L) && (110)\\
&= t - \sum_{\eta=1}^{t} H(Z_\eta|X_{1,\eta}^{[1]}, X_{2,\eta}^{[1]}, \cdots, X_{N,\eta}^{[1]}) + o(L) && (111)\\
&= t(1 - H(p)) + o(L) && (112)
\end{aligned}
$$

where (103) follows from the independence of the messages and the queries, (104) follows from the reliability constraint, (107) follows from the fact that the answer string $A_n^{[1]}$ is a deterministic function of the messages and the queries, (108) follows from the fact that $(W_{1:M}, Q_{1:N}^{[1]}) \to A_{1:N}^{[1]} \to \tilde{A}^{[1]}$ is a Markov chain, (109) follows from the fact that the channel is memoryless, and (112) follows from the independence of Z_η and $(X_{1,\eta}^{[1]}, X_{2,\eta}^{[1]}, \cdots, X_{N,\eta}^{[1]})$ as a consequence of the independence of $(Z_\eta, W_{1:M}, Q_{1:N}^{[1]})$.

Hence, by reordering terms and taking $L \to \infty$, we have $R = \frac{L}{t} \leq 1 - H(p)$. Note that we can interpret the upper bound as the cooperative MAC bound, i.e., $R \leq I(Y; X_1, X_2, \cdots, X_N) = 1 - H(p)$.

b) The achievability proof: To show the general achievability, the user submits queries to database 1 and database 2 only and ignores the remaining databases. We note that the additive MAC in this case boils down to $Y_\eta = X_{1,\eta} + X_{2,\eta} + Z_\eta$, which means that the channel $p(y|x_1, x_2)$ is BSC(p). Consequently, we use again Shannon's coding theorem for BSC in Lemma 4.

To that end, let the mth message be a vector $W_m = [W_m(1) \quad W_m(2) \quad \cdots \quad W_m(L)]$ of length L. The user repeats

the following scheme L times. For the jth repetition of the scheme, the user generates a random binary vector $\mathbf{h}(j) = [h_1(j) \;\; h_2(j) \;\; \cdots \;\; h_M(j)] \in \{0,1\}^M$. The user sends the following queries to the databases:

$$Q_1^{[i]}(j) = \mathbf{h}(j) \tag{113}$$

$$Q_2^{[i]}(j) = \mathbf{h}(j) + \mathbf{e}_i \tag{114}$$

where $\mathbf{e}_i$ is the unit vector containing 1 only at the ith position. The queries are private since $Q_n^{[i]}$ is a vector picked uniformly from $\{0,1\}^M$ for any message i.

For the jth repetition of the scheme, the database uses the received query vector as a combining vector for the jth element of all messages. The nth database concatenates all responses in a vector $U_n^{[i]}$ of length L, hence

$$U_1^{[i]} = \Bigg[\sum_{m=1}^{M} h_m(1)W_m(1) \quad \sum_{m=1}^{M} h_m(2)W_m(2) \quad \cdots \quad \sum_{m=1}^{M} h_m(L)W_m(L)\Bigg] \tag{115}$$

$$U_2^{[i]} = \Bigg[\sum_{m=1}^{M} h_m(1)W_m(1) + W_i(1) \quad \sum_{m=1}^{M} h_m(2)W_m(2) + W_i(2) \quad \cdots \quad \sum_{m=1}^{M} h_m(L)W_m(L) + W_i(L)\Bigg] \tag{116}$$

From Lemma 4, for $p \in (0, 0.5)$, all but ρ linear $[t, L]$ block codes $\mathcal{C}$, where $\frac{L}{t} = R < 1 - H(p)$ that have $P_e(\mathcal{C}) < \frac{2}{\rho}.2^{-t\,\Delta(p,R)}$. Then, the databases agree on the same $[t, L]$ code from the family of good codes, where $t = \frac{L}{[1-H(p)]}$. The nth database encodes $U_n^{[i]}$ independently by the same $[t, L]$ linear block code to output $A_n^{[i]}$.

After passing through the noisy channel, the noisy observation is given by:

$$\tilde{A}^{[i]} = A_1^{[i]} + A_2^{[i]} + Z_{1:t} \tag{117}$$

$$= \hat{A}^{[i]} + Z_{1:t} \tag{118}$$

Since the two databases employ the same linear block code, the sum of the two codewords $\hat{A}^{[i]} = A_1^{[i]} + A_2^{[i]}$ is also a valid codeword corresponding to the sum $U_1^{[i]} + U_2^{[i]}$.

Consequently, as $L \to \infty$, $t \to \infty$, the probability of error in decoding the sum $U_1^{[i]} + U_2^{[i]}$ is $P_e(L) \to 0$. By observing that $U_1^{[i]} + U_2^{[i]} = W_i$, the reliability proof follows. ■

Remark 9 *In the achievability proof, the PIR scheme relies on the additivity of the channel. In particular, the scheme uses a linear block code to exploit the fact that the sum of two codewords from a linear block code is also a valid codeword. Consequently, the retrieval process depends on the channel transition probability explicitly as opposed to the NPIR problem with orthogonal links.*

B. Logic Conjunction/Disjunction MACs

In this section, we show that we can achieve privacy for free for MACs other than the additive MACs. We illustrate this result by considering the MAC-PIR problem through channels that output the logical conjunctions (logic AND)/disjunctions (logic OR) of the inputs. Let $\wedge$ denote the logical conjunction operator, $\vee$ denote the logical disjunction operator, and $\neg$ denote the logical negation operator. The input-output relation of the discrete memoryless logical conjunction channel is given as:

$$Y_\eta = \bigwedge_{n=1}^{N} X_{n,\eta} \tag{119}$$

For the logical conjunction channel, we have the following capacity result.

Theorem 4 *In the logical conjunction MAC-PIR problem, if $N \geq 2^{M-1}$, then the MAC-PIR capacity is $C_{PIR} = 1$, where M is the number of messages.*

We have the following observations:

Remark 10 *Similar to the additive MAC, there is no loss due to the privacy constraint for the conjunction MAC. In this case, the capacity depends on the number of messages M, and the number of databases N unlike the additive MAC. Interestingly, the result shows the first instance of a threshold for the number of databases at which the full unconstrained capacity can be achieved $N = 2^{M-1}$, which is dependent on the number of messages M.*

Remark 11 *We note that the minimum number of databases N that results in $C_{PIR} = 1$ is still an open problem. In fact, the capacity for $N < 2^{M-1}$ is also an interesting open problem.*

Proof: It suffices to show only the achievability for this problem as the retrieval rate is trivially upper bounded by 1. To that end, the user submits queries to 2^{M-1} databases and submits nothing to the remaining databases. The user generates the random variables $(Z_1, \cdots, Z_M)$ independently, privately, and uniformly from $\{0, 1\}$. The random variable $Z_m \sim \text{Bernoulli}(\frac{1}{2})$ is a Bernoulli random variable that represents the negation state of the mth message literal in the first query $Q_1^{[i]}$, i.e., if $Z_m = 1$, this means that the user requests W_m in $Q_1^{[i]}$, while $Z_m = 0$ means that the user requests $\neg W_m$ in $Q_1^{[i]}$. Let $\tilde{W}_m$ be the requested literal from the mth message in $Q_1^{[i]}$, hence,

$$\tilde{W}_m = \begin{cases} W_m, & Z_m = 1 \\ \neg W_m, & Z_m = 0 \end{cases} \tag{120}$$

Now, without loss of generality, assume that W_1 is the desired message. From database 1, the user requests to download the disjunction $X_1 = \bigvee_{m=1}^{M} \tilde{W}_m$. From every other database, the user requests the same literal $\tilde{W}_1$ with a new disjunction of the remaining messages with different negation pattern than what is requested from database 1. I.e., from

database 2, the user requests the disjunction $X_2 = \tilde{W}_1 \vee \neg\tilde{W}_2 \vee \bigvee_{m\in[M]\setminus\{1,2\}} \tilde{W}_m$. From database 3, the user requests the disjunction $X_3 = \tilde{W}_1 \vee \neg\tilde{W}_3 \vee \bigvee_{m\in[M]\setminus\{1,3\}} \tilde{W}_m$, $\cdots$ etc. Denote the disjunction of messages $W_{2:M}$ requested from the nth database by F_n, where $n \in \{1, \cdots, 2^{M-1}\}$, then the received observation at the user is

$$\begin{aligned} Y &= \left(\bigvee_{m=1}^{M} \tilde{W}_m\right) \\ &\quad \wedge \left(\tilde{W}_1 \vee \neg\tilde{W}_2 \vee \bigvee_{m\in[M]\setminus\{1,2\}} \tilde{W}_m\right) \\ &\quad \wedge \left(\tilde{W}_1 \vee \neg\tilde{W}_3 \vee \bigvee_{m\in[M]\setminus\{1,3\}} \tilde{W}_m\right) \wedge \cdots \qquad (121) \\ &= \tilde{W}_1 \vee \bigwedge_{i=1}^{2^{M-1}} F_i \qquad (122) \\ &= \tilde{W}_1 \qquad (123) \end{aligned}$$

where (122) follows from successively applying the Boolean relation $(\tilde{W}_1 \vee G_1) \wedge (\tilde{W}_1 \vee G_2) = \tilde{W}_1 \vee (G_1 \wedge G_2)$ for any logical expressions G_1, G_2. (123) follows from the fact that there exist 2^{M-1} different negation states for the literals from $W_{2:M}$, each negation state is requested from one database in the form of logical expression F_i, hence the conjunction of all these logical expressions $\bigwedge_{i=1}^{2^{M-1}} F_i = 0$ as all possible product of sums of $W_{2:M}$ exist in the conjunction. This satisfies the reliability constraint. Another way to see this result is that the queries are designed such that they cover *exactly half* the M-dimensional Karnaugh map, which can be reduced to either W_1 or $\neg W_1$.

Furthermore, since the negation state for every message is chosen uniformly, independently, and uniformly for each message, the probability of receiving specific query from the user is $\frac{1}{2^M}$ irrespective to the desired message, which guarantees the privacy. ■

a) Illustrative example: $M = 3$ messages, $N = 4$ databases with conjunction channel: As an explicit example, let $M = 3$, $N = 2^{M-1} = 4$, then the user requests the following:

$$X_1 = \tilde{W}_1 \vee \tilde{W}_2 \vee \tilde{W}_3 \qquad (124)$$

$$X_2 = \tilde{W}_1 \vee \neg\tilde{W}_2 \vee \tilde{W}_3 \qquad (125)$$

$$X_3 = \tilde{W}_1 \vee \tilde{W}_2 \vee \neg\tilde{W}_3 \qquad (126)$$

$$X_4 = \tilde{W}_1 \vee \neg\tilde{W}_2 \vee \neg\tilde{W}_3 \qquad (127)$$

Hence, the output of the channel is,

$$\begin{aligned} Y &= X_1 \wedge X_2 \wedge X_3 \wedge X_4 \qquad (128) \\ &= (\tilde{W}_1 \vee \tilde{W}_2 \vee \tilde{W}_3) \wedge (\tilde{W}_1 \vee \neg\tilde{W}_2 \vee \tilde{W}_3) \\ &\quad \wedge (\tilde{W}_1 \vee \tilde{W}_2 \vee \neg\tilde{W}_3) \wedge (\tilde{W}_1 \vee \neg\tilde{W}_2 \vee \neg\tilde{W}_3) \qquad (129) \\ &= (\tilde{W}_1 \vee (\tilde{W}_2 \vee \tilde{W}_3) \wedge (\neg\tilde{W}_2 \vee \tilde{W}_3)) \\ &\quad \wedge (\tilde{W}_1 \vee (\tilde{W}_2 \vee \neg\tilde{W}_3) \wedge (\neg\tilde{W}_2 \vee \neg\tilde{W}_3)) \qquad (130) \\ &= (\tilde{W}_1 \vee W_3) \wedge (\tilde{W}_1 \vee \neg\tilde{W}_3) \qquad (131) \\ &= \tilde{W}_1 \qquad (132) \end{aligned}$$

Thus, the user can decode W_1 from Y as the user knows the correct negation pattern for $\tilde{W}_1$ privately. The scheme is private as all queries are equally likely with probability $\frac{1}{8}$ irrespective to the desired message. Since the user downloads 1 bit to retrieve 1 bit from the desired message, the retrieval rate $R = 1$.

Remark 12 *We note that the result is still valid if the channel is replaced by a disjunction channel, i.e.,*

$$Y_\eta = \bigvee_{n=1}^{N} X_{n,\eta} \qquad (133)$$

In this case, the user submits the same queries for the databases with replacing every disjunction operator with a conjunction operator. The proof of reliability follows from the duality of the product-of-sum and the sum-of-product.

Remark 13 *The achievable scheme for the conjunction channel is a non-linear retrieval scheme that depends on the non-linear characteristics of the channel in contrast to the linear retrieval scheme used for the additive channel. This confirms the non-separability between the retrieval scheme and the channel coding needed for reliable communication through the channel.*

C. Selection Channel

In this example, we illustrate the fact that the *privacy for free* phenomenon may not be always feasible for any arbitrary channel in the MAC-PIR problem. To illustrate this, we consider the selection channel. In this channel, the user selects to connect to one database only at random and sticks to it throughout the transmission, i.e.,

$$Y_\eta = X_{n,\eta}, \quad n \sim \text{uniform}\{1, \cdots, N\} \qquad (134)$$

In this channel, the user is connected to the same database at every channel use. This implies that the user faces a single-database ($N = 1$) PIR problem at every channel use. The optimal PIR strategy for $N = 1$ is to download all the messages (M messages) from the connected database. Thus, the PIR capacity is given by $C_{\text{PIR}} = \frac{1}{M}$.

It is worth noting that there is another slight variant of the selection channel, in which the user selects to connect to one database at random at every channel use, i.e.,

$$Y_\eta = X_{n(\eta),\eta}, \quad n(\eta) \sim \text{uniform}\{1, \cdots, N\} \qquad (135)$$

where $n(\eta)$ corresponds to the database index at channel use η. Then, $C_{PIR} \leq C = (1 + \frac{1}{N} + \cdots + \frac{1}{N^{M-1}})^{-1}$ trivially as the capacity of the classical PIR C [11], in which all the databases are connected to the user, is an upper bound for this problem, as the user can choose to ignore all the responses except the ones in the classical PIR problem. For the achievability, the user can repeat the achievable scheme in [11] ν times, which results in using the selection channel $t = \nu\frac{L}{C} = \nu\frac{N(N^M-1)}{N-1}$. At channel use η, the user chooses a new query element from $Q^{[i]}_{n(\eta)}$ and submits it to database $n(\eta)$. As $\nu \to \infty$, by strong law of large numbers, each database

will be visited t_n times, where $t_n \to \frac{t}{N}$ in the limit for every n. Hence, all bits are decodable by the decodability of the scheme in [11] and $C_{PIR} = C = (1 + \frac{1}{N} + \cdots + \frac{1}{N^{M-1}})^{-1} < 1$ as well.

VII. Conclusion

In this paper, we introduced noisy PIR with orthogonal links (NPIR), and PIR from multiple access channels (MAC-PIR). We focused on the issue of the separability of the channel coding and the retrieval scheme. For the NPIR problem, we proved that the channel coding and the retrieval scheme are *almost separable* in the sense that every database implements its own channel coding independently from other databases. The problem is coupled only through agreeing on a suitable traffic ratio vector to maximize the retrieval rate. On the other hand, these conclusions are not valid for the MAC-PIR problem. We showed two examples, namely: PIR from additive MAC and PIR from logical conjunction/disjunction MAC. In these examples, we showed that the channel coding and retrieval schemes are indeed *inseparable* unlike in the NPIR problem. In both cases, we showed that by careful design of joint retrieval and coding schemes, we can attain the full capacity $C_{PIR} = 1 - H(p)$ and $C_{PIR} = 1$, respectively, with no loss due to the privacy constraint.

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] W. Gasarch, "A survey on private information retrieval," *Bull. EATCS*, vol. 82, pp. 72–107, 2004.

[3] C. Cachin, S. Micali, and M. Stadler, "Computationally private information retrieval with polylogarithmic communication," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn.* Berlin, Germany: Springer, 1999, pp. 402–414.

[4] R. Ostrovsky and W. Skeith, III, "A survey of single-database private information retrieval: Techniques and applications," in *Proc. Int. Workshop Public Key Cryptogr.* Berlin, Germany: Springer, 2007, pp. 393–411.

[5] S. Yekhanin, "Private information retrieval," *Commun. ACM*, vol. 53, no. 4, pp. 68–73, Apr. 2010.

[6] 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.

[7] G. Fanti and K. Ramchandran, "Efficient private information retrieval over unsynchronized databases," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 7, pp. 1229–1239, Oct. 2015.

[8] T. H. Chan, S. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *Proc. IEEE ISIT*, Jun. 2015, pp. 2842–2846.

[9] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *Proc. IEEE ISIT*, Jun. 2015, pp. 2852–2856.

[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] 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.

[12] 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.

[13] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[14] 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.

[15] 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.

[16] Q. Wang and M. Skoglund, "Symmetric private information retrieval for MDS coded distributed storage," in *Proc. IEEE ICC*, May 2017, pp. 1–6.

[17] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.

[18] 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.

[19] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[20] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. El Rouayheb, "Private information retrieval schemes for coded data with arbitrary collusion patterns," in *Proc. IEEE ISIT*, Jun. 2017, pp. 1908–1912.

[21] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[22] Y. Zhang and G. Ge, "A general private information retrieval scheme for MDS coded databases with colluding servers," 2017, *arXiv:1704.06785*. [Online]. Available: https://arxiv.org/abs/1704.06785

[23] Y. Zhang and G. Ge, "Private information retrieval from MDS coded databases with colluding servers under several variant models," 2017, *arXiv:1705.03186*. [Online]. Available: https://arxiv.org/abs/1705.03186

[24] K. Banawan and S. Ulukus, "The capacity of private information retrieval from Byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[25] R. Tajeddine and S. El Rouayheb, "Robust private information retrieval on coded data," in *Proc. IEEE ISIT*, Jun. 2017, pp. 1903–1907.

[26] Q. Wang and M. Skoglund, "Secure symmetric private information retrieval from colluding databases with adversaries," in *Proc. 55th Annu. Conf. Commun. Control Comput. (Allerton)*, Oct. 2017, pp. 1083–1090.

[27] R. Tandon, "The capacity of cache aided private information retrieval," in *Proc. 55th Annu. Conf. Commun. Control Comput. (Allerton)*, Oct. 2017, pp. 1078–1082.

[28] Q. Wang and M. Skoglund, "Linear symmetric private information retrieval for MDS coded distributed storage with colluding servers," 2017, *arXiv:1708.05673*. [Online]. Available: https://arxiv.org/abs/1708.05673

[29] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," 2017, *arXiv:1709.00112*. [Online]. Available: https://arxiv.org/abs/1709.00112

[30] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3215–3232, May 2019.

[31] Z. Chen, Z. Wang, and S. Jafar, "The capacity of T-private information retrieval with private side information," 2019, *arXiv:1709.03022*. [Online]. Available: https://arxiv.org/abs/1709.03022

[32] Y.-P. Wei, K. Banawan, and S. Ulukus, "The capacity of private information retrieval with partially known private side information," *IEEE Trans. Info. Theory*, to be published.

[33] Q. Wang and M. Skoglund, "Secure private information retrieval from colluding databases with eavesdroppers," 2017, *arXiv:1710.01190*. [Online]. Available: https://arxiv.org/abs/1710.01190

[34] H. Sun and S. A. Jafar, "The capacity of private computation," 2017, *arXiv:1710.11098*. [Online]. Available: https://arxiv.org/abs/1710.11098

[35] M. Mirmohseni and M. A. Maddah-Ali, "Private function retrieval," 2017, *arXiv:1711.04677*. [Online]. Available: https://arxiv.org/abs/1711.04677

[36] M. Abdul-Wahid, F. Almoualem, D. Kumar, and R. Tandon, "Private information retrieval from storage constrained databases—Coded caching meets PIR," 2017, *arXiv:1711.05244*. [Online]. Available: https://arxiv.org/abs/1711.05244

[37] Y.-P. Wei, K. Banawan, and S. Ulukus, "Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1126–1139, Jun. 2018.

[38] K. Banawan and S. Ulukus, "Asymmetry hurts: Private information retrieval under asymmetric traffic constraints," *IEEE Trans. Inf. Theory*, to be published.

[39] K. Banawan and S. Ulukus, "Private information retrieval through wiretap channel II: Privacy meets security," *IEEE Trans. Inf. Theory*, under review.

[40] Z. Chen, Z. Wang, and S. Jafar, "The asymptotic capacity of private search," in *Proc. IEEE ISIT*, Jun. 2018, pp. 2122–2126.

[41] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti, "Robust private information retrieval from coded systems with byzantine and colluding servers," in *Proc. IEEE ISIT*, Jun. 2018, pp. 2451–2455.
[42] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from Uncoded storage constrained databases," 2018, *arXiv:1805.04104*. [Online]. Available: https://arxiv.org/abs/1805.04104
[43] Y.-P. Wei and S. Ulukus, "The capacity of private information retrieval with private side information under storage constraints," *IEEE Trans. Inf. Theory*, under review.
[44] R. Roth, *Introduction to Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

Karim Banawan (S'13–M'18) Karim Banawan received the B.Sc. and M.Sc. degrees, with highest honors, in electrical engineering from Alexandria University, Alexandria, Egypt, in 2008, 2012, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Maryland at College Park, MD, USA, in 2017 and 2018, respectively, with his Ph.D. thesis on private information retrieval and security in networks. He was the recipient of the Distinguished Dissertation Fellowship from the Department of Electrical and Computer Engineering, at the University of Maryland College Park, for his Ph.D. thesis work.

In 2019, he joined the department of electrical engineering, Alexandria University, as an assistant professor. 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) Sennur Ulukus is the Anthony Ephremides Professor in Information Sciences and Systems in the Department 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 information theory, wireless communications, machine learning, signal processing and networks, with recent focus on private information retrieval, age of information, distributed coded computation, energy harvesting communications, 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, the 2019 IEEE Communications Society Best Tutorial Paper Award, an 2005 NSF CAREER Award, the 2010-2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Outstanding Teaching Award. She is a Distinguished Lecturer of the IEEE Information 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 is a TPC co-chair of 2019 ITW, 2017 IEEE ISIT, 2016 IEEE Globecom, 2014 IEEE PIMRC, and 2011 IEEE CTW.