CHALMERS
UNIVERSITY OF TECHNOLOGY

Private Information Retrieval in Wireless Coded Caching

Citation for the original published paper (version of record):
Kumar, S., Graell I Amat, A., Rosnes, E. (2019). Private Information Retrieval in Wireless Coded Caching. IEEE Workshop on Signal Processing Advances in Wireless Communications, SPAWC. http://dx.doi.org/10.1109/SPAWC.2019.8815548

Private Information Retrieval in Wireless Coded Caching

(Invited Paper)

Siddhartha Kumar[†], Alexandre Graell i Amat[‡†], and Eirik Rosnes[†]
[†]Simula UiB, N-5020 Bergen, Norway
[‡]Department of Signals and Systems, Chalmers University of Technology, SE-41296 Gothenburg, Sweden

***Abstract*—We consider private information retrieval (PIR) in a content delivery scenario where, to reduce the backhaul usage, data is cached using maximum distance separable codes in a number of small-cell base stations (SBSs). We present a PIR protocol that allows the user to retrieve files of different popularities from the network without revealing the identity of the desired file to *curious* SBSs that potentially collaborate. We formulate an optimization problem to optimize the content placement and the number of queries of the protocol such that the backhaul rate is minimized. We further prove that, contrary to the case of no PIR, uniform content placement is optimal. Compared to a recently proposed protocol by Kumar *et al.* the presented protocol gives a reduced backhaul rate.**

I. Introduction

Distributed caching is a promising technology to enable low-latency content delivery in wireless networks [1]. The key idea is to bring content closer to the end user by caching it in a number of small-cell base stations (SBSs) [2], [3] or directly in the mobile devices [4]. For the first scenario, it was shown in [2] that encoding content using erasure correcting codes prior of being cached significantly improves performance compared to popular content caching, where the most popular files are cached across all SBSs. This approach was further studied in [3], where the authors assumed that content is cached using maximum distance separable (MDS) codes and optimized the rates of the codes used for caching to minimize the usage of the backhaul link. Alternatively, content can be cached to facilitate index-coded broadcasts [5].

The explosion of distributed information systems has been accompanied by increasing concerns about security and privacy. In particular, in recent years, private information retrieval (PIR) has attracted significant attention in the research community. In PIR one would like to retrieve data from a distributed database, where some of the servers are *curious* (and can potentially collude), without revealing the identity of the requested piece of data to the curious servers. PIR was first introduced by Chor *et al.* in [6] for the noncolluding scenario (i.e., the servers do not collaborate), assuming that the data is replicated among the servers. The efficiency of a PIR protocol is typically given in terms of its PIR rate, defined as the ratio between the requested file size and the amount of downloaded data. The maximum achievable rate by any PIR protocol, i.e., the PIR capacity, was derived in [7] and [8] for the noncolluding case when data is replicated across servers and stored using a single MDS code, respectively. Several PIR protocols have been introduced in [9]–[11].

This work was partially funded by the Research Council of Norway (grant 240985/F20) and the Swedish Research Council (grant 2016-04253).

In this paper, we consider the private retrieval of content from a cellular network where, similar to the scenario in [2], [3], a number of SBSs with some cache capacity are deployed to make content delivery more efficient. In particular, as in [3] we assume that, prior of being cached, content is encoded using MDS codes. We assume that some of the SBSs are curious and may collaborate to identify the requested content. Users wish to download files of different popularities without leaking any information of the requested files to the curious SBSs. We present a PIR protocol that yields the desired privacy for this scenario. The proposed protocol is an extension of the protocols in [10], [11] to the multiple code rate case that supports the fact that files may be stored using different code rates. Based on this protocol, we then derive the backhaul rate of the system and formulate an optimization problem to optimize the content placement and the protocol parameters such that the backhaul rate is minimized. We further prove that uniform content allocation is optimal, in contrast to the case of no PIR, where uniform content allocation is suboptimal in general [3]. Compared to our recently proposed protocol [12], the presented protocol gives a reduced backhaul rate. We give results for a Poisson point process (PPP) deployment model where SBSs are distributed over the plane according to a PPP.

II. System Model

We consider a cellular network where a macro-cell is served by a macro base station (MBS) and N_{SBS} SBSs are deployed to offload traffic from the MBS. The MBS has access to a library of F files $\boldsymbol{X}^{(i)}$, $i = 1, \ldots, F$, through a backhaul link. In particular, each file $\boldsymbol{X}^{(i)}$ consists of βL bits and is represented as a matrix composed of β stripes of L bits each,

$$\boldsymbol{X}^{(i)} = \begin{pmatrix} \tilde{\boldsymbol{x}}_1^{(i)} \\ \vdots \\ \tilde{\boldsymbol{x}}_\beta^{(i)} \end{pmatrix},$$

where $\tilde{\boldsymbol{x}}_a^{(i)}$ represents the a-th stripe, $a = 1, \ldots, \beta$, of $\boldsymbol{X}^{(i)}$. The file library has popularity distribution $\boldsymbol{p} = (p_1, \ldots, p_F)$, where file $\boldsymbol{X}^{(i)}$ is requested with probability p_i. We assume that each SBS has a cache size equivalent to M files.

A. Content Placement

File $\boldsymbol{X}^{(i)}$ is partitioned into βk_i packets of size L/k_i bits and encoded before being cached in the SBSs. In particular, each

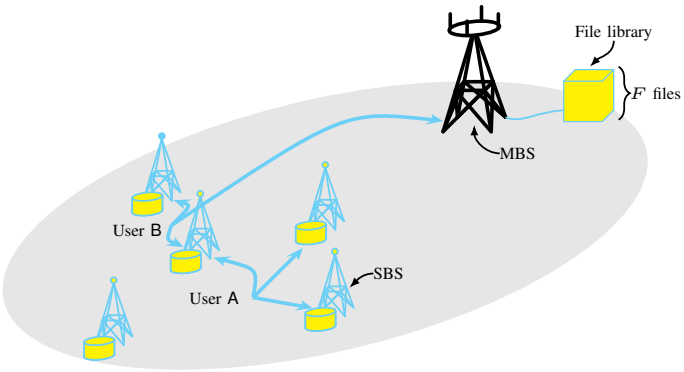

Fig. 1. A cellular network consisting of an MBS and five SBSs. The most popular files are cached on the SBSs using a $(5, 3)$ MDS code. User A can download a cached file from the three nearby SBSs, whereas User B, which is close to only two SBSs, downloads a cached file from the SBSs and the MBS.

packet is mapped onto a symbol of the field $\mathrm{GF}(q^{\delta_i})$, with $\delta_i \geq \frac{L}{k_i \log_2 q}$. For simplicity, we assume that $\frac{L}{k_i \log_2 q}$ is integer and set $\delta_i = \frac{L}{k_i \log_2 q}$. Thus, stripe $\tilde{\boldsymbol{x}}_a^{(i)}$ can be equivalently represented by a stripe $\boldsymbol{x}_a^{(i)}$, $a = 1, \ldots, \beta$, of symbols over $\mathrm{GF}(q^{\delta_i})$. Each stripe $\boldsymbol{x}_a^{(i)}$ is then encoded using an (N_{SBS}, k_i) MDS code $\mathcal{C}_i$ over $\mathrm{GF}(q)$, $q > N_{\mathsf{SBS}}$, into a codeword $\boldsymbol{c}_a^{(i)} = (c_{a,1}^{(i)}, \ldots, c_{a,N_{\mathsf{SBS}}}^{(i)})$, where code symbols $c_{a,j}^{(i)}$, $j = 1, \ldots, N_{\mathsf{SBS}}$, are over $\mathrm{GF}(q^{\delta_i})$. For later use, we define $k_{\mathsf{min}} \triangleq \min\{k_i\}$, $k_{\mathsf{max}} \triangleq \max\{k_i\}$, and $\delta_{\mathsf{max}} \triangleq \frac{L}{k_{\mathsf{min}} \log_2 q}$.

The encoded file can be represented by a $\beta \times N_{\mathsf{SBS}}$ matrix $\boldsymbol{C}^{(i)} = (c_{a,j}^{(i)})$. Code symbols $c_{a,j}^{(i)}$ are then stored in the j-th SBS (the ordering is unimportant). Thus, for each file $\boldsymbol{X}^{(i)}$, each SBS caches one coded symbol of each stripe of the file. We define $\mu_i \triangleq 1/k_i$. As $k_i \in \{1, \ldots, N_{\mathsf{SBS}} - 1\}$,

$$\mu_i \in \mathcal{M} \triangleq \{0, 1/(N_{\mathsf{SBS}} - 1), \ldots, 1/2, 1\},$$

where $\mu_i = 0$ implies that file $\boldsymbol{X}^{(i)}$ is not cached. Note that, to achieve privacy in a nontrivial manner (i.e., without downloading everything), $k_i < N_{\mathsf{SBS}}$, i.e., files need to be cached with redundancy. As a result, $\mu_i = 1/N_{\mathsf{SBS}}$ is not allowed. This is in contrast to the case of no PIR, where $k_i = N_{\mathsf{SBS}}$ (and hence $\mu_i = 1/N_{\mathsf{SBS}}$) is possible. Since each SBS can cache the equivalent of M files, $\sum_{i=1}^{F} \mu_i \leq M$. We define the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_F)$ and refer to it as the *content placement*. Note that $\boldsymbol{\mu}$ defines the rates of the codes used for caching. Also, we denote by $\mathcal{C}_{\mathsf{MDS}}^{\boldsymbol{\mu}}$ the caching scheme that uses MDS codes $\{\mathcal{C}_i\}$ according to content placement $\boldsymbol{\mu}$. For later use, we define $\mu_{\mathsf{min}} \triangleq \min\{\mu_i | \mu_i \neq 0\}$ and $\mu_{\mathsf{max}} \triangleq \max\{\mu_i\}$, and let $\boldsymbol{H}^{\mathcal{C}}$ denote a parity-check matrix of a code $\mathcal{C}$. The considered scenario is depicted in Fig. 1.

B. File Request

Mobile devices request files according to the popularity distribution $\boldsymbol{p} = (p_1, \ldots, p_F)$. Without loss of generality, we assume $p_1 \geq p_2 \geq \ldots \geq p_F$. The user request is initially served by the SBSs within communication range. We denote by γ_b the probability that the user is served by b SBSs and define $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{N_{\mathsf{SBS}}})$. Additional required symbols are fetched from the MBS. The efficiency of a (PIR) caching scheme is measured in terms of the so-called backhaul rate [3], defined as

$$\mathsf{R} \triangleq \frac{\text{average no. of bits downloaded from the MBS}}{\beta L},$$

and one would like to design a PIR caching scheme so that R is minimized. Note that for the case of no caching $\mathsf{R} = 1$.

C. Private Information Retrieval and Problem Formulation

We assume that some of the SBSs are curious and potentially collaborate with each other. On the other hand, we assume that the MBS can be trusted. The users wish to retrieve files from the cellular network, but do not want the curious SBSs to learn any information about which file is requested by the user. The goal is to retrieve data from the network privately while minimizing the use of the backhaul link, i.e., while minimizing R. Thus, the goal is to optimize the content placement $\boldsymbol{\mu}$ to minimize R.

III. Private Information Retrieval Protocol

We are now ready to present the proposed PIR protocol. The protocol can be seen as an extension of the protocols in [10], [11] to the case where data is stored using erasure correcting codes of different rates.

Consider that the user wants to retrieve file $\boldsymbol{X}^{(i)}$. According to the protocol, the user generates $n \leq N_{\mathsf{SBS}}$ queries $\boldsymbol{Q}^{(l)}$, $l = 1, \ldots, n$, of which b queries are sent to b SBSs within communication range and the remaining $n - b$ queries are sent to the MBS, unless $b < b^{\mathsf{th}}$ in which case the file is downloaded directly from the MBS. Since by assumption the MBS can be trusted, the protocol does not leak any information by downloading the file directly from the MBS. The number of queries n and the threshold b^{th} need be optimized. This is in contrast to the protocol in [12] where a fixed $b^{\mathsf{th}} = 1$ was used. The intuition is that when no SBSs are within communication range, it is beneficial to download the file with no redundancy directly from the MBS. Each query $\boldsymbol{Q}^{(l)}$ is a $d_l \times \beta F$ matrix over $\mathrm{GF}(q)$ with the following structure,

$$\boldsymbol{Q}^{(l)} = \begin{pmatrix} \boldsymbol{q}_1^{(l)} \\ \vdots \\ \boldsymbol{q}_{d_l}^{(l)} \end{pmatrix} = \begin{pmatrix} q_{1,1}^{(l)} & \cdots & q_{1,\beta F}^{(l)} \\ \vdots & \cdots & \vdots \\ q_{d_l,1}^{(l)} & \cdots & q_{d_l,\beta F}^{(l)} \end{pmatrix}.$$

The query matrix $\boldsymbol{Q}^{(l)}$ consists of d_l subqueries $\boldsymbol{q}_j^{(l)}$, $j = 1, \ldots, d_l$, of length βF symbols each. More specifically, the $n - b$ queries sent to the MBS contain $d_l = d_{\mathsf{MBS}}$ subqueries, while the queries sent to b SBSs within communication range contain $d_l = d_{\mathsf{SBS}}$ subqueries. This is in contrast to the protocol in [12] where a fixed $d_l = k_{\mathsf{max}}$ was used for all queries. In response to each query, a SBS (or the MBS) sends back a response vector $\boldsymbol{r}^{(l)} = (r_1^{(l)}, \ldots, r_{d_l}^{(l)})^{\mathsf{T}} = \boldsymbol{Q}^{(l)} (c_{1,l}^{(1)}, \ldots, c_{\beta,l}^{(1)}, \ldots, c_{1,l}^{(F)}, \ldots, c_{\beta,l}^{(F)})^{\mathsf{T}}$, where $(\cdot)^{\mathsf{T}}$ denotes the transpose of its argument. Each response vector consists of d_l subresponses, $r_j^{(l)} \in \mathrm{GF}(q^{\delta_{\mathsf{max}}})$, $j = 1, \ldots, d_l$, and as such each subresponse symbol is of L/k_{min} bits.

The concept of information-theoretic PIR is formally defined as follows.

Definition 1. *Consider a wireless caching scenario with N_{SBS} SBSs that cache parts of a library of F files and in which an arbitrary set $\mathcal{T}$ of T SBSs are curious and may collude. A user wishes to retrieve the i-th file and generates queries $\boldsymbol{Q}^{(l)}$, $l = 1, \ldots, n$. In response to the queries the SBSs and (potentially) the MBS send back the responses $\boldsymbol{r}^{(l)}$. This scheme achieves perfect information-theoretic PIR if and only if*

$$\textit{Privacy:} \qquad \mathsf{H}\big(i|\{\boldsymbol{Q}^{(l)}, l \in \mathcal{T}\}\big) = \mathsf{H}(i); \tag{1a}$$

$$\textit{Recovery:} \qquad \mathsf{H}\big(\boldsymbol{X}^{(i)}|\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(n)}\big) = 0. \tag{1b}$$

The privacy condition (1a) means that the curious SBSs gain no additional information about which file is requested from the queries (i.e., the uncertainty about the file requested after observing the queries is identical to the a priori uncertainty determined by the popularity distribution). The recovery condition (1b) guarantees that the user is able to recover the file from the n response vectors.

For the sequel, we define the set of codes $\{\mathcal{C}_i'\}$, of parameters (n, k_i), obtained by puncturing, without loss of generality, the rightmost coordinates of the (N_{SBS}, k_i) storage codes $\mathcal{C}_i$, and by $\mathcal{C}'_{\mathsf{max}}$ the code with parameters (n, k_{max}). We require that k_{min} divides k_i for all i, i.e., $k_{\mathsf{min}} \mid k_i$, which ensures that $\mathrm{GF}(q^{\delta_i}) \subseteq \mathrm{GF}(q^{\delta_{\mathsf{max}}})$. Furthermore, for the protocol we require that $\mathcal{C}_i' \subseteq \mathcal{C}'_{\mathsf{max}}$. The proposed protocol is characterized by the following codes: the (n, k_i) codes $\{\mathcal{C}_i'\}$, which characterize the storage of files on SBSs, an $(n, \bar{k})$ code $\bar{\mathcal{C}}$ that defines the queries and is referred as the query code, and an $(n, \tilde{k})$ code $\tilde{\mathcal{C}}$ that characterizes the retrieval process.

A. Query Construction

The construction starts by choosing βF codewords $\bar{\boldsymbol{c}}_m^{(i)} = (\bar{c}_{m,1}^{(i)}, \ldots, \bar{c}_{m,n}^{(i)}) \in \bar{\mathcal{C}}$, $m = 1, \ldots, \beta$, $i = 1, \ldots, F$, independently and uniformly at random. Next, let $\mathring{\boldsymbol{c}}_l = (\mathring{\boldsymbol{c}}_l^{(1)}, \ldots, \mathring{\boldsymbol{c}}_l^{(F)})$, $l = 1, \ldots, n$, where $\mathring{\boldsymbol{c}}_l^{(i)} = (\bar{c}_{1,l}^{(i)}, \ldots, \bar{c}_{\beta,l}^{(i)})$ is the collection of the l-th coordinates of the β codewords $\bar{\boldsymbol{c}}_m^{(i)}$, $m = 1, \ldots, \beta$. Finally, assuming that the i-th file is requested, the j-th subquery to the l-th node (SBS or MBS) is given as

$$\boldsymbol{q}_j^{(l)} = \mathring{\boldsymbol{c}}_l + \boldsymbol{\delta}_j^{(l)}, \quad \boldsymbol{\delta}_j^{(l)} = \begin{cases} \boldsymbol{\omega}_{\beta(i-1)+s_j^{(l)}} & \text{if } l \in \mathcal{J}_j, \\ \boldsymbol{\omega}_0 & \text{otherwise,} \end{cases} \tag{2}$$

for some set $\mathcal{J}_j$ that will be defined shortly. Vector $\boldsymbol{\omega}_t$, $t = 1, \ldots, \beta F$, denotes the t-th (βF)-dimensional unit vector, i.e., the length-βF vector with a one in the t-th coordinate and zeroes elsewhere. The vector $\boldsymbol{\omega}_0$ denotes the all-zero vector of length βF. The meaning of index $s_j^{(l)}$ will become apparent below.

The queries in (2) are a sum of a random vector and a deterministic vector. The random vector $\mathring{\boldsymbol{c}}_l$ ensures that the protocol achieves privacy, whereas the design of the deterministic vector $\boldsymbol{\delta}_j^{(l)}$ allows for the retrieval of coded symbols pertaining to the requested file $\boldsymbol{X}^{(i)}$. In particular, the set $\mathcal{J}_j$ in (2) is a set that corresponds to the nodes (SBSs or MBS) from which the j-th subquery downloads code symbols. As in [11], these sets are characterized by a $d \times n$ binary matrix $\hat{\boldsymbol{E}}$, where for the proposed protocol $d \triangleq \max\{d_l\}$, and β information sets $\mathcal{I}_m$, $m = 1, \ldots, \beta$, of $\mathcal{C}'_{\mathsf{max}}$. The index $s_j^{(l)}$ in (2) is chosen such that $s_j^{(l)} \in \mathcal{F}_l$ and $s_j^{(l)} \neq s_{j'}^{(l)}$ for $j' \neq j$, $j', j = 1, \ldots, d_l$, where $\mathcal{F}_l = \{m : l \in \mathcal{I}_m\}$ is the set of indices pertaining to the β information sets of $\mathcal{C}'_{\mathsf{max}}$ that contain the l-th coordinate. The matrix $\hat{\boldsymbol{E}}$ should satisfy the following three conditions.

- C1. The j-th row of $\hat{\boldsymbol{E}}$ should have support $\mathcal{J}_j$ of size Γ, for some Γ, which implies that in each subquery the protocol is able to recover Γ code symbols of the requested file.
- C2. Each row, regarded as an erasure pattern (where ones denote erasures) should be correctable by the retrieval code $\tilde{\mathcal{C}}$. This allows the protocol to be able to recover desired code symbols from each subquery.
- C3. To ensure that the downloaded code symbols enable the protocol to retrieve the requested file, the protocol needs to guarantee that they are part of the information sets $\mathcal{I}_m$, $m = 1, \ldots, \beta$. These β information sets correspond to the β stripes of the requested file. Also, the protocol should be able to download $\Gamma d \geq \beta k_i$ unique code symbols across d subqueries. This leads to the following property. Let $\boldsymbol{t}_l$ be the l-th column vector of $\hat{\boldsymbol{E}}$. Then, $w_{\mathsf{H}}(\boldsymbol{t}_l) = |\mathcal{F}_l|$, where $w_{\mathsf{H}}(\boldsymbol{t}_l)$ denotes the Hamming weight of $\boldsymbol{t}_l$.

B. Response Vectors

Corresponding to the j-th subquery of the l-th query, either the MBS or a SBS computes the subresponse $r_j^{(l)} = \langle \boldsymbol{q}_j^{(l)}, (c_{1,l}^{(1)}, \ldots, c_{\beta,l}^{(F)}) \rangle$, which is collected into a length-n vector $\boldsymbol{\rho}_j = (r_j^{(1)}, \ldots, r_j^{(n)})^{\mathsf{T}}$ as follows,

$$\boldsymbol{\rho}_j = \sum_{i=1}^{F} \sum_{m=1}^{\beta} \underbrace{\begin{pmatrix} \bar{c}_{m,1}^{(i)} c_{m,1}^{(i)} \\ \vdots \\ \bar{c}_{m,n}^{(i)} c_{m,n}^{(i)} \end{pmatrix}}_{\in \left\{\boldsymbol{x} \in (\mathrm{GF}(q^{\delta_{\mathsf{max}}}))^n : \boldsymbol{H}^{\mathcal{C}_i' \circ \bar{\mathcal{C}}} \boldsymbol{x} = \boldsymbol{0}\right\}} + \begin{pmatrix} o_j^{(1)} \\ \vdots \\ o_j^{(n)} \end{pmatrix}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. $o_j^{(l)}$ represents the code symbol from the requested file downloaded in the j-th subresponse from the l-th response vector, or zero if no symbol is downloaded in the j-th subresponse from the l-th response vector. From (3), $\boldsymbol{\rho}_j$ is a sum of βF vectors from the nullspaces of $\boldsymbol{H}^{\mathcal{C}_i' \circ \bar{\mathcal{C}}}$, $i = 1, \ldots, F$, β vectors from each nullspace. Consider a retrieval code $\tilde{\mathcal{C}}$ of the form

$$\tilde{\mathcal{C}} = \sum_{i=1}^{F} \mathcal{C}_i' \circ \bar{\mathcal{C}} \overset{(a)}{=} \left(\sum_{i=1}^{F} \mathcal{C}_i' \right) \circ \bar{\mathcal{C}}, \tag{4}$$

where $\circ$ denotes the Hadamard product and where (a) follows due to the fact that the Hadamard product is distributive over addition. The symbols requested by the user are then obtained by solving the system of linear equations defined by $\boldsymbol{H}^{\tilde{\mathcal{C}}} \boldsymbol{\rho}_j = \boldsymbol{H}^{\tilde{\mathcal{C}}} (o_j^{(1)}, \ldots, o_j^{(n)})^{\mathsf{T}}$.

C. Privacy

For the protocol to have nonzero PIR rate, the storage codes $\{\mathcal{C}_i\}$ and the query code $\bar{\mathcal{C}}$ should be such that $\tilde{\mathcal{C}}$ is of rate $R < 1$. We present a family of MDS codes, namely generalized Reed-Solomon (GRS) codes, that work with the protocol. An

$(n, k, \boldsymbol{v}, \boldsymbol{\kappa})$ GRS code over GF(q) of length n and dimension k is a weighted polynomial evaluation code that is defined by a weighting vector $\boldsymbol{v}$ and an evaluation vector $\boldsymbol{\kappa}$ [13, Ch. 5].

Theorem 1. *Let $\mathcal{C}_{\mathsf{MDS}}^{\boldsymbol{\mu}}$ be a caching scheme with GRS codes $\{\mathcal{C}_i\}$ of parameters $(N_{\mathsf{SBS}}, k_i, \boldsymbol{v}, (\kappa_1, \ldots, \kappa_{N_{\mathsf{SBS}}}))$ and let $\mathcal{C}'_i$ be the (n, k_i) code obtained by puncturing $\mathcal{C}_i$. Also, let $\bar{\mathcal{C}}$ be an $(n, T, \bar{\boldsymbol{v}}, (\kappa_1, \ldots, \kappa_n))$ GRS code. Let $\boldsymbol{X}^{(i)}$ denote the requested file, and let $T \leq n - k_{\mathsf{max}}$. Then, for $\beta = \Gamma = n - (k_{\mathsf{max}} + T - 1)$, $d_{\mathsf{MBS}} = k_i$, and $d_{\mathsf{SBS}} = k_{\mathsf{max}}$, the protocol achieves PIR against up to T colluding SBSs.*

Proof: The proof follows the same lines as the proof of [12, Th. 1]. ∎

To illustrate the main principles of the proposed PIR protocol, we now present a brief example.

Example 1. *Consider a cellular network with $N_{\mathsf{SBS}} = 6$ SBSs that cache two files $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ by first encoding them using two GRS codes $\mathcal{C}_1$ and $\mathcal{C}_2$, of parameters $(6, 1, \boldsymbol{v}, \boldsymbol{\kappa})$ and $(6, 5, \boldsymbol{v}, \boldsymbol{\kappa})$, respectively. The user wishes to retrieve $\boldsymbol{X}^{(1)}$ privately, and we assume that $n = N_{\mathsf{SBS}}$ (i.e., no puncturing) and that none of the SBSs collude, i.e., $T = 1$. According to Theorem 1, we choose $\bar{\mathcal{C}}$ as a $(6, 1, \bar{\boldsymbol{v}}, \boldsymbol{\kappa})$ GRS code. The generator matrices of the two storage codes are $\boldsymbol{G}^{\mathcal{C}_1} = \begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{pmatrix}$ and $\boldsymbol{G}^{\mathcal{C}_2} = \boldsymbol{V}^{\mathsf{T}} \cdot \mathrm{diag}(\boldsymbol{v})$, respectively, where $\boldsymbol{V}$ is a 6×5 Vandermonde matrix constructed from the evaluation vector $(\kappa_1, \ldots, \kappa_6)$ and $\mathrm{diag}(\boldsymbol{v})$ is a diagonal matrix with $\boldsymbol{v}$ along the diagonal. It follows that $\mathcal{C}_1 \subset \mathcal{C}_2$. From (4) and [10, Prop. 3], $\bar{\mathcal{C}}$ is a $(6, 5, \boldsymbol{v} \circ \bar{\boldsymbol{v}}, \boldsymbol{\kappa})$ GRS code, and according to Theorem 1, we set $\beta = \Gamma = 1$, $d_{\mathsf{MBS}} = k_1 = 1$, and $d_{\mathsf{SBS}} = k_{\mathsf{max}} = 5$. Consider that the user has access to SBSs 1, 2, and 5 (i.e., $b = 3$) and that $b^{\mathsf{th}} \leq b$. Thus, $d_1 = d_2 = d_5 = d_{\mathsf{SBS}} = 5$, $d_3 = d_4 = d_6 = d_{\mathsf{MBS}} = 1$, and $d = \max\{d_l\} = 5$. Furthermore, we have*

$$\hat{\boldsymbol{E}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad \textit{and} \quad \mathcal{I}_1 = \{1, 2, 3, 4, 5\},$$

where $\mathcal{I}_1$ is an information set of $\mathcal{C}'_{\mathsf{max}} = \mathcal{C}_2$.

The user generates $\beta F = 2$ codewords of $\bar{\mathcal{C}}$ independently and uniformly at random. Without loss of generality assume that $\bar{\boldsymbol{c}}_1^{(1)} = \bar{\boldsymbol{c}}_1^{(2)} = (v_1, \ldots, v_6)$. Then, the subqueries $\boldsymbol{q}_1^{(l)}$, $l = 1, \ldots, 6$, are constructed according to (2) as

$$\boldsymbol{q}_1^{(l)} = \begin{cases} (\bar{c}_{1,l}^{(1)}, \bar{c}_{1,l}^{(2)}) + (1, 0) & \textit{if } l = 1, \\ (\bar{c}_{1,l}^{(1)}, \bar{c}_{1,l}^{(2)}) + (0, 0) & \textit{otherwise}. \end{cases}$$

Each subquery $\boldsymbol{q}_1^{(l)}$, $l = 1, \ldots, 6$, is sent to either the MBS or a SBS. More specifically, $\boldsymbol{q}_1^{(1)}$, $\boldsymbol{q}_1^{(2)}$, and $\boldsymbol{q}_1^{(5)}$ are sent to the first, the second, and the fifth SBS, respectively, while the remaining subqueries are sent to the MBS. The SBSs and the MBS generate the subresponses as

$$r_1^{(l)} = \langle \boldsymbol{q}_1^{(l)}, (c_{1,l}^{(1)}, c_{1,l}^{(2)}) \rangle = \begin{cases} \bar{c}_{1,l}^{(1)} c_{1,l}^{(1)} + \bar{c}_{1,l}^{(2)} c_{1,l}^{(2)} + c_{1,l}^{(1)} & \textit{if } l = 1, \\ \bar{c}_{1,l}^{(1)} c_{1,l}^{(1)} + \bar{c}_{1,l}^{(2)} c_{1,l}^{(2)} & \textit{otherwise}. \end{cases}$$

Since $\boldsymbol{\rho}_1$ is the sum of two vectors from the nullspaces of $\boldsymbol{H}^{\mathcal{C}_1 \circ \bar{\mathcal{C}}}$ and $\boldsymbol{H}^{\mathcal{C}_2 \circ \bar{\mathcal{C}}}$, which is in the nullspace of $\boldsymbol{H}^{\bar{\mathcal{C}}}$, and the vector $\left(c_{1,1}^{(1)}, 0, 0, 0, 0, 0\right)^{\mathsf{T}}$, $\boldsymbol{H}^{\bar{\mathcal{C}}} \boldsymbol{\rho}_1 = \boldsymbol{H}^{\bar{\mathcal{C}}} \left(c_{1,1}^{(1)}, 0, 0, 0, 0, 0\right)^{\mathsf{T}}$. One can trivially solve the above equation to obtain $c_{1,1}^{(1)}$. With this, the user has obtained $\boldsymbol{X}^{(1)}$. However, to ensure privacy at the SBSs, the user sends the remaining four subqueries to the three SBSs within reach (and not to the MBS), and receives corresponding subresponses that are disregarded.

IV. Backhaul Rate Analysis

Proposition 1. *The backhaul rate for the PIR caching scheme $\mathcal{C}_{\mathsf{MDS}}^{\boldsymbol{\mu}}$ in Section II (with GRS codes) is*

$$\mathsf{R}_{\mathsf{PIR}} = \frac{\mu_{\mathsf{max}} \mu_{\mathsf{min}}}{\mu_{\mathsf{min}}(n - T + 1) - 1} \sum_{i=1}^{F} p_i \frac{\lceil \mu_i \rceil}{\mu_i} \sum_{b=b^{\mathsf{th}}}^{n} \gamma_b (n - b) + \sum_{i=1}^{F} p_i \lceil \mu_i \rceil \sum_{b=0}^{b^{\mathsf{th}}-1} \gamma_b + \sum_{i=1}^{F} p_i \lfloor 1 - \mu_i \rfloor.$$

Proof: The proof is similar to the proof of [12, Prop. 2], and is omitted for brevity. ∎

One can obtain the minimum backhaul rate, $\mathsf{R}^*_{\mathsf{PIR}}$ by solving the optimization problem

$$\mathsf{R}^*_{\mathsf{PIR}} = \min_{\substack{\mu_i \in \mathcal{M} \\ n \in \mathcal{A} \\ b^{\mathsf{th}} \in \mathcal{B}}} \frac{\mu_{\mathsf{max}} \mu_{\mathsf{min}}}{\mu_{\mathsf{min}}(n - T + 1) - 1} \sum_{i=1}^{F} p_i \frac{\lceil \mu_i \rceil}{\mu_i} \sum_{b=b^{\mathsf{th}}}^{n} \gamma_b (n - b) + \sum_{i=1}^{F} p_i \lceil \mu_i \rceil \sum_{b=0}^{b^{\mathsf{th}}-1} \gamma_b + \sum_{i=1}^{F} p_i \lfloor 1 - \mu_i \rfloor \tag{5}$$

$$\text{s.t.} \sum_{i=1}^{F} \mu_i \leq M \text{ and } k_{\mathsf{min}} \mid k_i,$$

where $\mathcal{A} = \{1/\mu_{\mathsf{min}} + T, \ldots, N_{\mathsf{SBS}}\}$, $\mathcal{B} = \{b_{\mathsf{l}}^{\mathsf{th}}, \ldots, b_{\mathsf{u}}^{\mathsf{th}}\}$, $b_{\mathsf{u}}^{\mathsf{th}} = \lceil n - ((n - T + 1)\mu_{\mathsf{min}} - 1)/\mu_{\mathsf{max}} \rceil$, and $b_{\mathsf{l}}^{\mathsf{th}} = \lceil n - ((n - T + 1) - 1/\mu_{\mathsf{min}}) \rceil$. The minimum value of n, i.e., $1/\mu_{\mathsf{min}} + T$, comes from the fact that $\mu_{\mathrm{min}}(n - T + 1) - 1$ must be positive, and the expressions for $b_{\mathsf{l}}^{\mathsf{th}}$ and $b_{\mathsf{u}}^{\mathsf{th}}$ come from the inequality $(n - b) k_i L / k_{\mathsf{min}} \leq (n - (k_{\mathsf{max}} + T - 1))L$, which is equivalent to

$$b \geq n - \frac{(n - T + 1)\mu_i - \mu_i/\mu_{\mathsf{min}}}{\mu_{\mathsf{max}}}.$$

Interestingly, it can be shown that the optimal solution of (5) is uniform content allocation.

Theorem 2. *Uniform content allocation, i.e., $\mu_i = \mu$ for all files that are cached, is optimal. Furthermore, the optimal number of files to cache is the maximum possible, i.e., $\mu_i = \mu$ for $i \leq \min(M/\mu, F)$.*

Proof: The argument for the second part of the theorem follows along the lines of [12, Lem. 3]. Thus, we only give a detailed proof for the first part, i.e., that uniform content allocation is optimal. Let $\boldsymbol{\mu}$ denote a feasible solution to (5), and let $\mu = \mu_{\mathsf{min}}$ to simplify notation. Furthermore, assume that the files with indices in $\mathcal{F}_{\mathsf{c}} \triangleq \{i_1, \ldots, i_{|\mathcal{F}_{\mathsf{c}}|}\} \subseteq \{1, \ldots, F\}$ are

cached. Hence, the support of $\boldsymbol{\mu}$ is $\chi(\boldsymbol{\mu}) = \{i_1, \ldots, i_{|\mathcal{F}_\mathsf{c}|}\}$, and the backhaul rate can be lowerbounded as

$$\begin{aligned}
\mathsf{R}_{\mathsf{PIR}}^{\mathsf{feasible}} &= \frac{\mu_{\mathsf{max}}\mu}{\mu\theta - 1}\left[\frac{p_{i_1}}{\mu_{i_1}} + \cdots + \frac{p_{i_{|\mathcal{F}_\mathsf{c}|}}}{\mu_{i_{|\mathcal{F}_\mathsf{c}|}}}\right]\sum_{b=b^{\mathsf{th}}}^{n}\gamma_b(n-b) \\
&\quad + \sum_{i=1}^{F} p_i\lceil\mu_i\rceil \sum_{b=0}^{b^{\mathsf{th}}-1}\gamma_b + \sum_{i=1}^{F} p_i\lfloor 1-\mu_i\rfloor \\
&\geq \frac{\mu^2}{\mu\theta - 1}\left[\frac{p_{i_1}}{\mu} + \cdots + \frac{p_{i_{|\mathcal{F}_\mathsf{c}|}}}{\mu}\right]\sum_{b=b_{\mathsf{uni}}^{\mathsf{th}}}^{n}\gamma_b(n-b) \\
&\quad + \sum_{i=1}^{F} p_i\lceil\mu_i\rceil \sum_{b=0}^{b_{\mathsf{uni}}^{\mathsf{th}}-1}\gamma_b + \sum_{i=1}^{F} p_i\lfloor 1-\mu_i\rfloor, \qquad (6)
\end{aligned}$$

where $\theta \triangleq n - T + 1$ and $b_{\mathsf{uni}}^{\mathsf{th}} \triangleq \left\lceil n - \frac{\theta\mu - 1}{\mu}\right\rceil$, since 1) $\mu_i \leq \mu_{\mathsf{max}}$, $\forall\, i \in \mathcal{F}_\mathsf{c}$, which implies that $\frac{\mu^2}{(\mu\theta-1)\mu} \leq \frac{\mu_{\mathsf{max}}\mu}{(\mu\theta-1)\mu_i}$, $\forall\, i \in \mathcal{F}_\mathsf{c}$, and 2) $b^{\mathsf{th}} \geq b_{\mathsf{l}}^{\mathsf{th}} = b_{\mathsf{uni}}^{\mathsf{th}}$ and $\frac{\mu^2\gamma_b(n-b)}{(\mu\theta-1)\mu} \leq \gamma_b$ for $b \geq b_{\mathsf{uni}}^{\mathsf{th}}$, by definition. Let $\boldsymbol{\mu}'$ denote a vector with $\chi(\boldsymbol{\mu}') = \chi(\boldsymbol{\mu})$ and such that $\mu'_{\mathsf{min}} = \mu'_{\mathsf{max}} = \mu_{\mathsf{min}} = \mu$. Clearly, $\boldsymbol{\mu}'$ is also a feasible solution as the conditions in (5) are satisfied, and it follows that $b_{\mathsf{l}}^{\mathsf{th}} = b_{\mathsf{u}}^{\mathsf{th}} = b_{\mathsf{uni}}^{\mathsf{th}}$. The corresponding backhaul rate is

$$\begin{aligned}
\mathsf{R}_{\mathsf{PIR}}^{\mathsf{uniform}} &= \frac{\mu^2}{\mu\theta - 1}\left[\frac{p_{i_1}}{\mu} + \cdots + \frac{p_{i_{|\mathcal{F}_\mathsf{c}|}}}{\mu}\right]\sum_{b=b_{\mathsf{uni}}^{\mathsf{th}}}^{n}\gamma_b(n-b) \\
&\quad + \sum_{i=1}^{F} p_i\lceil\mu_i\rceil \sum_{b=0}^{b_{\mathsf{uni}}^{\mathsf{th}}-1}\gamma_b + \sum_{i=1}^{F} p_i\lfloor 1-\mu_i\rfloor,
\end{aligned}$$

which is equal to the lower bound on $\mathsf{R}_{\mathsf{PIR}}^{\mathsf{feasible}}$ from (6). Thus, $\mathsf{R}_{\mathsf{PIR}}^{\mathsf{uniform}} \leq \mathsf{R}_{\mathsf{PIR}}^{\mathsf{feasible}}$ and therefore uniform content allocation is optimal. ∎

V. Numerical Results

In Fig. 2, we plot the backhaul rate for a PPP deployment model where SBSs are distributed over the plane according to a PPP and a user at an arbitrary location in the plane can connect to all SBSs that are within radius r_u. Let λ be the density of SBSs per square meter. For this scenario, the probability that a user is in communication range of b SBSs is $\gamma_b = \mathrm{e}^{-\psi}\frac{\psi^b}{b!}$, where $\psi = \lambda\pi r_\mathsf{u}^2$. The popularity of the i-th file is $p_i = \frac{1/i^\alpha}{\sum_\ell 1/\ell^\alpha}$, where $\alpha \in [0.5, 1.5]$ is the skewness factor. In the figure, we plot the optimized backhaul rate from (5) (solid lines) as a function of the density λ for $F = 200$ files, $\alpha = 0.7$, $r_\mathsf{u} = 60$ meters, different cache size constraint M, and $T = 2$. For comparison, we also plot the curves for popular content placement, i.e., with $\mu_{\mathsf{max}} = \mu_{\mathsf{min}} = 1$ in (5) (dashed lines), and the optimal backhaul rate of the protocol from [12] (dotted lines). Note that if the protocols give a backhaul rate above one, we plot one, since a backhaul rate of one can be obtained by downloading the file directly from the MBS. As can be seen, the proposed protocol gives a lower backhaul rate compared to the protocol from [12]. In the figure, we give the optimal values of n and $k = 1/\mu$ for $M = 50$. For convenience, we only give the parameters for the densities where the optimal pair (n, k) changes. The values should be read as follows. Walking the solid curve for $M = 50$ from top-left to bottom-right, $(3, 1)$ is optimal for nonzero densities up to $\lambda = 3.7 \cdot 10^{-4}$. Then, for $\lambda = 3.8 \cdot 10^{-4}$ to $\lambda = 5.5 \cdot 10^{-4}$, $(4, 2)$ is optimal, and so on (the curves are plotted with steps of 10^{-5}).

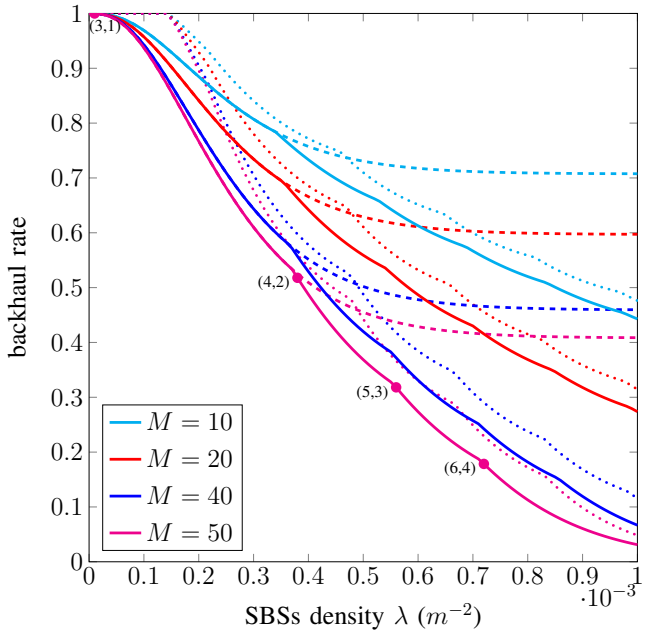

Fig. 2. Backhaul rate as a function of the density of SBSs λ and several values M for $T = 2$, $F = 200$ files, and $\alpha = 0.7$.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[3] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, Dec. 2015.

[4] J. Pedersen, A. Graell i Amat, I. Andriyanova, and F. Brännström, "Optimizing MDS coded caching in wireless networks with device-to-device communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 286–295, Jan. 2019.

[5] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[6] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. 36th IEEE Symp. Found. Comp. Sci. (FOCS)*, Milwaukee, WI, Oct. 1995, pp. 41–50.

[7] 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.

[8] 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.

[9] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[10] 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 Geom.*, vol. 1, no. 1, pp. 647–664, Nov. 2017.

[11] S. Kumar, H.-Y. Lin, E. Rosnes, and A. Graell i Amat, "Achieving maximum distance separable private information retrieval capacity with linear codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, Jul. 2019.

[12] S. Kumar, A. Graell i Amat, E. Rosnes, and L. Senigagliesi, "Private information retrieval from a cellular network with caching at the edge," *IEEE Trans. Commun.*, vol. 67, no. 7, Jul. 2019.

[13] W. C. Huffman and V. Pless, Eds., *Fundamentals of Error-Correcting Codes*. Cambridge, UK: Cambridge University Press, 2010.