Freij-Hollanti, Ragnar; Gnilke, Oliver; Hollanti, Camilla; Horlemann-Trautmann, Anna-Lena; Karpuk, David; Kubjas, Ivo

Reed-Muller Codes for Private Information Retrieval

Published: 18/09/2017

Document Version
Peer reviewed version

Please cite the original version:
Freij-Hollanti, R., Gnilke, O., Hollanti, C., Horlemann-Trautmann, A-L., Karpuk, D., & Kubjas, I. (2017). *Reed-Muller Codes for Private Information Retrieval*. Paper presented at International Workshop on Coding and Cryptography, Saint-Petersburg, Russian Federation.

Reed-Muller Codes for Private Information Retrieval

Ragnar Freij-Hollanti[1], Oliver W. Gnilke[1], Camilla Hollanti[1], Anna-Lena Horlemann-Trautmann[2], David Karpuk[1], and Ivo Kubjas[3]

[1] Aalto University
firstname.lastname@aalto.fi
[2] University of St. Gallen
anna-lena.horlemann@unisg.ch
[3] University of Tartu
ivokub@ut.ee

Abstract. We present private information retrieval protocols for coded storage with colluding servers. While previous schemes require field sizes that grow with the number of servers and files in the system, we restrict the field size and focus especially on the binary case. Reed-Muller codes are shown to be especially useful in this regard and explicit parameters are calculated.

1 Introduction

1.1 Background

Private information retrieval (PIR) seeks to retrieve data from a database without disclosing information about the identity of the data items retrieved, and was introduced by Chor, Goldreich, Kushilevitz, and Sudan in [CGKS95], [CKGS98]. The classic PIR model of [CKGS98] views the database as an n-bit binary string $x = (x^1, \ldots, x^n) \in \{0,1\}^n$, and assumes that the user wants to retrieve a single bit x^i without revealing any information about the index i. The *download rate* of a PIR scheme is measured as the ratio of the gained information over the downloaded information, while upload costs of the requests are usually ignored. The trivial solution of downloading the entire database is the only way to guarantee *information-theoretic privacy* in the case of a single server [CKGS98], but replicating the database onto k servers that do not communicate can significantly increase the rate, as in [CKGS98], [Efr09], [DG15] and the references therein.

Shah, Rashmi, Ramchandran, and Kumar recently introduced a model of PIR for coded data [SRRK12], [SRR14]. Here, all files are distributed over the servers according to a storage code. It is shown in [SRR14] that for a suitably constructed storage code, privacy can be guaranteed by downloading a single bit more than the size of the desired file. However, this requires exponentially many servers in terms of the number of files. Blackburn, Etzion and Paterson achieved the same low download complexity with a linear number of servers [BEP16], but this is still far from

applicable storage systems where the number of files tends to dwarf the number of servers.
Modern distributed storage systems require communication between servers to recover data in the case of node failure. As such, it is natural in a PIR scheme to allow the servers to *collude*, that is, to assume the servers inform each other of their interaction with the user. Explicit PIR schemes for coded storage and colluding servers were previously considered in [TE16], and [FHGHK16].
The maximum possible rate, or *capacity* of a PIR scheme for a replicated storage system was derived in [SJ16a] (without collusion) and [SJ16b] (with collusion). The corresponding PIR capacity of an MDS-coded storage system was given in [BU16], in the case of no colluding servers. The PIR capacity of MDS-coded storage systems with colluding servers is only known for some particular sets of parameters [SJ17].

1.2 Summary of Present Work

While explicit PIR schemes are constructed in [SJ16a], [SJ16b], and [BU16] which achieve capacity, they require the base field to be large. If n is the number of servers and M is the number of files, the capacity-achieving schemes of [SJ16b] require a field size of $q = O(n^M)$, since they rely on existence of MDS codes of high lengths. Realistic storage systems, however, may operate over fields of small size to keep the complexity of the involved operations manageable. One would naturally then like to construct explicit PIR schemes over small base fields.
In this work we construct PIR schemes based on Reed-Muller (RM) codes. The schemes described in [FHGHK16] employed Generalized Reed-Solomon (GRS) codes, and the resulting analysis of the achievable rate relied on the *star product* of two GRS codes again being a GRS code. The class of RM codes is closed under the star product operation as well, and thus naturally lends itself to be employed using the PIR scheme of [FHGHK16]. However, RM codes have the advantage of being defined over the binary field $\mathbb{F}_2$.

2 Basic Definitions

2.1 Private Information Retrieval

Let us describe the distributed storage systems we consider; this setup follows that of [TE16,FHGHK16,BU16]. To provide clear and concise notation, we have consistently used superscripts to refer to files, subscripts to refer to servers, and parenthetical indices for entries of a vector.
Suppose we have files $x^1, \ldots, x^M \in \mathbb{F}_q^k$. The considered data storage scheme proceeds by arranging the files into an $M \times k$ matrix

$$X = \begin{bmatrix} x^1 \\ \vdots \\ x^M \end{bmatrix} = \begin{bmatrix} x^1(1) & \cdots & x^1(k) \\ \vdots & \ddots & \vdots \\ x^M(1) & \cdots & x^M(k) \end{bmatrix}. \tag{1}$$

Each file x^i is encoded using a linear $[n,k,d]_q$-code C having generator matrix G_C, into an encoded file $y^i = x^i G_C$. In matrix form, we encode the matrix X into a matrix Y by right-multiplying by G_C:

$$Y = XG_C = \begin{bmatrix} y^1 \\ \vdots \\ y^M \end{bmatrix} = \begin{bmatrix} y_1 \cdots y_n \end{bmatrix} . \tag{2}$$

The j^{th} column $y_j \in \mathbb{F}_q^M$ of the matrix Y is stored by the j^{th} server. Such a storage system allows any $d-1$ servers to fail while still allowing users to successfully access any of the files x^i. If C is an MDS code, the resulting distributed storage system is maximally robust against server failures.
The following defines precisely what we mean by PIR scheme; for simplicity we have limited ourselves to simple linear schemes, which suffices to describe all schemes we will construct.

Definition 1 *Suppose we have a distributed storage system as above, where M files are stored across n servers. A* PIR scheme *for such a storage system consists of:*

1. *For each index $i \in [M]$, a probability space $(\mathcal{Q}^i, \mu^i)$ of* queries. *When the user wishes to download x^i, a query $q^i \in \mathcal{Q}^i$ is selected randomly according to the probability measure μ^i. Each q^i is itself a tuple $q^i = (q_1^i, \ldots, q_n^i)$, where $q_j^i \in \mathbb{F}_q^M$ is sent to the j^{th} server.*
2. *Responses $r_j^i = \langle q_j^i, y_j \rangle \in \mathbb{F}_q$ which the servers compute and transmit to the user.*
3. *A reconstruction function which takes as input $(r_1^i, \ldots, r_n^i) \in \mathbb{F}_q^n$ and returns c coordinates of x^i.*

The download rate *of a PIR scheme is defined to be c/n.*

Definition 2 *We call a set $T \subseteq [n]$ a* collusion set *if it is possible for the servers indexed by T to share their requests in an attempt to deduce the index of the requested file.*
A PIR scheme protects against the colluding set $T = \{j_1, \ldots, j_t\} \subseteq [n]$ *if we have*

$$I(q_{j_1}^i, \ldots, q_{j_t}^i; i) = 0 \tag{3}$$

where $I(\cdot\,;\cdot)$ denotes the mutual information of two random variables. In other words, there exists a probability distribution $(\mathcal{Q}_T, \mu_T)$ such that, for all $i \in [M]$, the projection of $(\mathcal{Q}^i, \mu^i)$ to the coordinates in T is $(\mathcal{Q}_T, \mu_T)$. If a PIR scheme protects against all colluding sets T of size $\leq t$, we say it protects against t-collusion.

Stated somewhat less formally, if a PIR scheme protects agains the colluding set T, the servers in T will not learn anything about the index i of the file that is being requested, even after sharing their requests with each other.
For the rest of this paper we will exclusively consider linear schemes that use uniform distributions on the query spaces, as in the following fundamental example.

Example 3 *Let $n = 2$ servers each store a copy of a database consisting of M files $x^\ell \in \mathbb{F}_q$. We denote by $x = (x^1, \ldots, x^M)^\top$ the two columns of the storage matrix. To retrieve the i^{th} file the user chooses uniformly at random an element u in $\mathbb{F}_q^M$ and constructs the queries as $q^i = (q_1^i, q_2^i) = (u, u+e_i)$. The space of all queries therefore is given by $\mathcal{Q}^i = \{(u, v) : v - u = e_i\}$ and μ^i is the uniform probability measure on $\mathcal{Q}^i$. The responses $r_j^i := \langle x, q_j^i \rangle$ are calculated as the inner product of the database with the requests and reconstruction is achieved by subtraction of the responses, $x^i = r_2^i - r_1^i$.*
This scheme is secure against either server individually, as both projections of any query space $\mathcal{Q}^i$ onto a coordinate are identical to the complete ambient space $\mathbb{F}_q^n$ with uniform measure. It does not, however, protect against 2-collusion, as the two servers can jointly observe the index i by computing the difference of their query vectors.

2.2 Reed-Muller Codes

In this subsection we define and give some well-known results on Reed-Muller codes. We note that there are various ways to define Reed-Muller codes; for our purposes it is most convenient to view them as evaluation codes from multivariate polynomials.

Definition 4 *Let $0 \leq r \leq m$ be integers and let $P_1, \ldots, P_{2^m}$ be all the points of $\mathbb{F}_2^m$. Then the* r-th order Reed-Muller code *of length $n = 2^m$, denoted by $RM(r, m)$, is defined as*

$$RM(r,m) := \Big\{ (p(P_1), p(P_2), \ldots, p(P_n)) \mid p \in \mathbb{F}_2[x_1, x_2, \ldots, x_m], \deg p \leq r \Big\}.$$

We need the following properties of Reed-Muller codes:

Lemma 5 *[MS77, Ch. 13] Reed-Muller codes satisfy the following properties:*
(i) $RM(r, m)$ is a linear code of dimension $k = \sum_{i=0}^{r} \binom{m}{i}$.
(ii) $RM(r, m)$ has minimum distance 2^{m-r}.
(iii) The dual code of $RM(r, m)$ is $RM(m - r - 1, m)$.

To analyze the parameters of a PIR scheme using Reed-Muller codes we need to know what the star product (also called Schur or Hadamard product) of such codes is. Let us first recall the definition of star product.

Definition 6 *Let C and D be linear codes of length n over $\mathbb{F}_q$. We define their* star product *$C \star D$ to be*

$$C \star D = \operatorname{span}\{[c(1)d(1), \ldots, c(n)d(n)] \in \mathbb{F}_q^n : c \in C, d \in D\},$$

which is again a linear code of length n.

The following result is well-known, but for clarity we will prove the result again in the following.

Lemma 7 *We have* $RM(r,m) \star RM(r',m) = RM(r+r',m)$.

Proof. By the definition of Reed-Muller codes it is easy to see that $RM(r,m) \star RM(r',m)$ *consists of the evaluations of all* $p \in \mathbb{F}_2[x_1, x_2, \ldots, x_m]$ *with degree up to* $r+r'$. *This implies the statement.* ■

It follows from Lemma 5 that $RM(r,m) \star RM(r',m)$ has minimum distance $2^{m-r-r'}$ and dimension $\sum_{i=0}^{r+r'} \binom{m}{i}$.
The number of minimal weight codewords of Reed-Muller codes is also known explicitly, and their structural description will be useful when proving quantitative bounds on the amount of collusion that our PIR schemes tolerate.

Lemma 8 *[MS77, Ch. 13, Thm. 8] Let* $c \in RM(r,m)$ *be a code word of minimal weight. Then* $\mathrm{supp}(c) \subseteq \mathbb{F}_2^m$ *is a flat of dimension* $m-r$.

The following corollary follows by a simple counting argument.

Corollary 9 *[MS77, Ch. 13, Thm. 9] The number of minimum weight codewords in* $RM(r,m)$ *is*

$$2^r \frac{\prod_{i=0}^{m-r-1}(2^{m-i}-1)}{\prod_{i=0}^{m-r-1}(2^{m-r-i}-1)}.$$

3 A General PIR Scheme for Coded Data Stored over Colluding Servers

In this section we briefly summarize the methods of [FHGHK16], which constructed explicit PIR schemes for coded data which protect against t-collusion. The crucial ingredients are the following: the storage code $C \subseteq \mathbb{F}_q^n$, another linear code $D \subseteq \mathbb{F}_q^n$, and the star product $C \star D$.
The main theorem of [FHGHK16] is the following:

Theorem 10 *[FHGHK16] Given a linear storage code* $C \subseteq \mathbb{F}_q^n$ *and a linear code* $D \subseteq \mathbb{F}_q^n$, *there exists a linear PIR scheme for the distributed storage system* $Y = XG_C$ *with rate* $(d_{C\star D}-1)/n$ *which protects against all colluding sets of size* $d_{D^\perp}-1$.

To privately retrieve a file x^i, for every file x^ℓ in the database a codeword d^ℓ is chosen uniformly at random from the code D. A vector $e \notin D$ is then added to d^i. The query $q_j^i \in \mathbb{F}_q^m$ sent to the j^{th} server is then

$$q_j^i = [d^1(j), \ldots, d^i(j)+e(j), \ldots, d^m(j)]$$

and the servers respond with

$$[r_1^i, \ldots, r_n^i] = [\langle q_1^i, y_1 \rangle, \ldots, \langle q_n^i, y_n \rangle] \in C \star D + C \star e.$$

The support of e is then chosen so that decoding the vector $[r_1^i, \ldots, r_n^i]$ to its closest neighbor in $C \star D$ reveals $d_{C \star D} - 1$ coordinates of y^i, coming from the $C \star e$ summand in the above expression. We will refer to the above as a (D, e)-*retrieval scheme*, where D and e are as above.

Example 11 *Suppose that $1 \leq t \leq n - k$. By choosing C and D to both be generalised Reed-Solomon (GRS) codes with the same evaluation vector, the (D, e)-retrieval scheme of [FHGHK16] can achieve a rate of $\frac{n-(k+t-1)}{n}$ while protecting against t-collusion. See [FHGHK16] for more details.*

4 Reed-Muller Codes in the Coded PIR Scheme

We now choose $C = RM(r, m)$ as storage code, such that $n = 2^m$ and $k = \sum_{i=0}^{r} \binom{m}{i}$. The code D is chosen to be $RM(r', m)$, a code of the same length, but possibly different dimension. Then we know, by Lemmas 5 and 7, that

$$C \star D = RM(r + r', m)$$

and hence

$$(C \star D)^{\perp} = RM(m - r - r' - 1, m).$$

We know that $C \star D$ has minimum distance $2^{m-r-r'}$. Thus we can choose any $e \in \mathbb{F}_2^n$ of weight $2^{m-r-r'} - 1$ and erasure decode $r = y \star (d + e) = y \star d + y \star e$ to $y \star d$ in $C \star D$. From this we can recover $y \star e = r - y \star d$. Note that, in contrary to MDS codes, not every set of $\dim(C)$ code symbols is an information set of a Reed-Muller code C[4]. Therefore we have to distinguish between code symbol and information symbol download rate. We get:

Proposition 12 *The scheme described above, using $C = RM(r, m)$ and $D = RM(r', m)$, has a code symbol download rate of*

$$\frac{2^{m-r-r'} - 1}{n} = \frac{2^{m-r-r'} - 1}{2^m} \approx 2^{-r-r'}$$

and protects against arbitrary collusions of size $2^{r'+1} - 1$.

Proof. Follows straight-forwardly from Theorem 10, since the minimum distance of $D^{\perp} = RM(m - r' - 1, m)$ is $2^{r'+1}$. ■

However, we can achieve a better download rate than that with the same codes, if we choose e slightly differently:

[4] An information set corresponds to an invertible maximal submatrix of the generator matrix of the code.

Proposition 13 *In the setting of Proposition 12 we can achieve a code symbol download rate of*

$$\frac{\dim(C \star D)^{\perp}}{n} = \frac{\sum_{i=0}^{m-r-r'-1} \binom{m}{i}}{2^m}.$$

Proof. Choose e to be of weight $\dim(C \star D)^{\perp}$ such that its support corresponds to an invertible submatrix of the parity check matrix of $C \star D$. Then we get $rH_{C\star D}^{\top} = (y \star d)H_{C\star D}^{\top} + (y \star e)H_{C\star D}^{\top} = (y \star e)H_{C\star D}^{\top}$, which has $\sum_{i=0}^{m-r-r'-1} \binom{m}{i}$ symbols of y. ■.

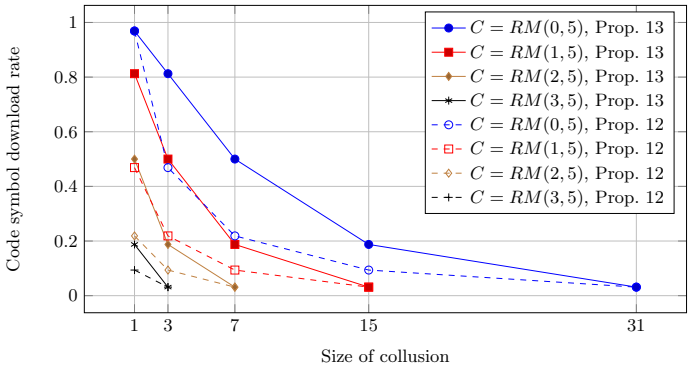

Fig. 1. Code symbol download rate improvement from Prop. 12 to Prop. 13.

Next, we will discuss how to go from a scheme with code symbol download rate c/n to one with information symbol download rate c/n. For this, we need the files that we download to be subdivided into α blocks, for some α that we will choose later. Hence, $X \in \mathbb{F}^{\alpha \times k}$ is stored via the $[n, k]$-code C across n servers.

$$\begin{pmatrix} x_1^1 & \cdots & x_k^1 \\ \vdots & \ddots & \vdots \\ x_1^\alpha & \cdots & x_k^\alpha \end{pmatrix} C = \begin{pmatrix} y_1^1 & \cdots & y_n^1 \\ \vdots & \ddots & \vdots \\ y_1^\alpha & \cdots & y_n^\alpha \end{pmatrix}.$$

We can interpret this as preprocessing the data into a desired format before storing it, or else we can think of our PIR scheme as downloading α different (vector-valued) files.
We assume we have a PIR scheme with rate r, *i.e.*, we can download one symbol each from up to $c = nr$ servers when using the protocol once. We will show how to use this to download all the $k\alpha$ desired symbols when using the protocol $k\alpha/c$ times. First, note that if $c|k$, we apply the scheme $\frac{k}{c}$ times to get k symbols and recover one row of X. We can thus choose $\alpha = 1$. To prove the general theorem, we will need the following lemma:

Lemma 1. *A code C with minimum distance d_C has at least $\lceil \frac{d}{k} \rceil$ disjoint information sets.*

Proof. Start with an arbitrary information set S_1. We will construct disjoint information sets $S_1, \cdots S_{\lceil \frac{d}{k} \rceil}$, each of size k, as follows. Inductively, for $1 \leq i < \frac{d}{k}$, consider the code C projected to the complement of $S_1 \cup \cdots \cup S_i$. This code has full rank, since C can correct $ik \leq d-1$ erasures. Thus, there will be an information set in the remaining columns, which we choose as S_{i+1}.

Theorem 1. *For an arbitrary storage code, the (D, e)-retrieval scheme can be repeated to download an entire file with information symbol download rate $\frac{d_{C \star D}-1}{n}$.*

Proof. Let $c = d_{C\star D} - 1 \leq d_C - 1$, and choose $\ell = \lceil \frac{nr}{k} \rceil$ disjoint information sets of C, as in Lemma 1. Every time the PIR scheme is used it can download the full information content from $\lfloor \frac{c}{k} \rfloor$ rows, using a different information set for each row, and an additional $b = c \mod k < k$ symbols from the last information set S_ℓ. It is therefore enough to see that the information symbol download rate is the same as the code symbol download rate when $c < k$.
To this end, let $\alpha = \frac{\mathrm{lcm}(c,k)}{k}$. We organise our files as being spread over α rows $\{r_0, \ldots, r_{\alpha-1}\}$ of X and download them by applying the (D, e)-PIR scheme $s = \frac{\mathrm{lcm}(c,k)}{c}$ times on the information set S_ℓ. Let $j = \frac{c}{\alpha}$. Now, in the t^{th} iteration of the PIR scheme, from row $0 \leq i < \alpha$ we download the symbols from columns $\{c_{j(i+t-1)}, c_{j(i+t-1)+1}, \ldots, c_{j(i+t-1)+j-1}\}$. Here, all indices are computed modulo k.

$$
\begin{array}{c}
\\
r_0 \\
r_1
\end{array}
\begin{array}{c}
\begin{array}{cccccccccc}
c_0 & \ldots & c_{j-1} & c_j & & c_{2j-1} & c_{2j} & & c_{3j-1}
\end{array} \\
\left(\begin{array}{cccccccccc}
1 & \cdots & 1 & 2 & \cdots & 2 & 3 & \cdots & 3 \\
s & \cdots & s & 1 & \cdots & 1 & 2 & \cdots & 2
\end{array}\right)
\end{array}
$$

Fig. 2. An illustration of the scheme in the case $\alpha = 2$, $k = 6$. An entry t in position i, j in this matrix means that in repetition t of the PIR protocol the symbol in column/server j from the i^{th} row is retrieved.

Example 14 *Suppose that $C = RM(0,4) = \mathrm{Rep}(16)_2$, so that data is stored via a replication system over $n = 16$ servers. Set $D = RM(1,4)$, which is a $[16,5,8]_2$-code. Then $(C \star D)^\perp = D^\perp = RM(2,4)$, which is a $[16,11,4]_2$-code. The corresponding PIR scheme achieves a code symbol download rate of $\frac{\sum_{i=0}^{2}\binom{4}{i}}{16} = \frac{11}{16}$ and protects against all colluding sets of size $2^{1+1} - 1 = 3$.*
Using the scheme of [FHGHK16] which employs GRS codes, one can protect against all colluding sets of size $t = 3$ while achieving a code symbol download rate of $\frac{n-t}{n} = \frac{13}{16}$. However, this improvement in rate

demands a scheme whose underlying field size is $q \geq 16$, compared to the Reed-Muller scheme which uses the binary field.
According to the results of [SJ16c], the PIR capacity for a replication system which stores m files over $n = 16$ servers and protects against all colluding sets of size $t = 3$ is given by

$$\text{Capacity} = \frac{1 - \frac{t}{n}}{1 - \left(\frac{t}{n}\right)^m} = \frac{13}{16} \cdot \frac{1}{1 - \left(\frac{3}{16}\right)^m}. \tag{4}$$

The rate achieved by the Reed-Muller scheme is 81.6% of capacity when the system only has $m = 2$ files, and is 84.1% of capacity when we have $m = 3$ files. When the number of files m increases, the capacity (4) decreases towards $\frac{13}{16}$, while the download rate achieved by the Reed-Muller scheme remains constant.
If we on the other hand fix the rate to be $\frac{11}{16}$ and compare PIR schemes, the scheme of [FHGHK16] achieves this rate by setting D to be a GRS code with parameters $[16, 5, 12]$. Then $D^\perp$ is GRS with parameters $[16, 11, 6]$, and hence this scheme protects against any set of colluding servers of size $t = d_{D^\perp} - 1 = 5$. Again, however, this improvement in privacy requires a field size of $q \geq 16$.

Remark: In the schemes corresponding to Propositions 12 and 13 we would also like the downloaded code symbols to contain as much information about the encoded file as possible. In the first scheme this is not an issue, since we can choose any e of the given weight. In the second scheme we have to be careful how freely we can choose the support of e. Naturally, the best scenario would be to have an information set of C as the set of downloaded symbols.
For certain choices of r' and m, dependent on r, we can guarantee to find such es, as explained in the following.

Theorem 15 *Let $r, r' \geq 1$ and $m = 2r + r' + 1$. Set $C = RM(r, m), D = RM(r', m)$. Then the Reed-Muller PIR scheme from Proposition 12 has a (code and) information symbol download rate of*

$$\frac{\dim C}{n} = \frac{\sum_{i=0}^{r} \binom{m}{i}}{2^m}.$$

This scheme can protect against arbitrary collusions of size $2^{r'+1} - 1$.

Proof. We have $(C \star D)^\perp = RM(m - r - r' - 1, m) = RM(r, m) = C$. Hence any invertible submatrix of $H_{C \star D}$ is also an invertible submatrix of G_C. Therefore, any $e \notin D$ of weight $\dim C$ whose support corresponds to an invertible submatrix of G_C gives us an information set $y \star e$ in C. This implies the statement.

Note that, for retrieving only one file, the number of downloaded symbols in Theorem 15 is optimal, in the sense that those symbols are an information set in C, i.e., we can reconstruct the whole file from them.

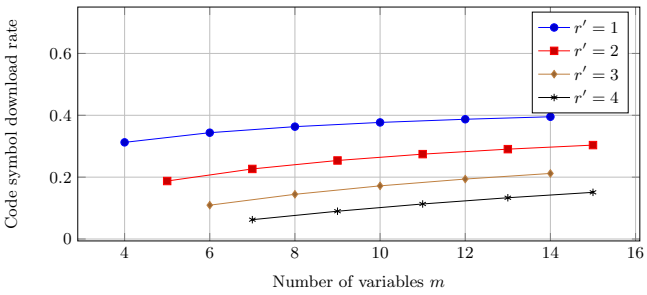

Fig. 3. Symbol download rate of Theorem 15

Example 16 *We consider $C = D = RM(1, 4)$ with generator matrix*

$$G_{C,D} = \left(\begin{array}{cccc|cccc|cccc|cccc} 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&0\\ 1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&0\\ 1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0 \end{array}\right).$$

We have $C \star D = RM(2, 4)$, which has parity check matrix $H_{C\star D} = G_{C,D}$ and minimum distance $2^{4-2} = 4$. Hence, with the classical scheme from Proposition 12 we can download 3 symbols. It protects against arbitrary collusions of size $2^2 - 1 = 3$.
However, if we choose any vector of weight 5 whose support indicates an invertible submatrix of $H_{C\star D} = G_{C,D}$, then we can download the full 5 information symbols. E.g., if we write the stored codeword $y = (a, b, c, d, e)G_{C,D}$ and choose $e = (1110|1000|1000|0000)$ we get

$$(y \star e)H_{C\star D}^{\top} = (a,\ c + d + e,\ b + d + e,\ b + c + e,\ b + c + d),$$

from which we can recover the whole message (a, b, c, d, e).

When $t \geq d_{D^\perp}$, then the Reed-Muller PIR scheme does not protect against all t-colluding sets of servers. However, for $t \approx d_{D^\perp}$, it does protect against "most" t-colluding sets in the following sense.

Proposition 17 *Let $D = RM(r, m)$, and let*

$$d_{D^\perp} = 2^{r+1} \leq t \leq \sum_{i=0}^{r} \binom{m}{i} = \dim(D).$$

Let $T \subseteq \mathbb{F}_2^m$ be a set of $|T| = t$ servers, chosen uniformly at random. Then the probability that the PIR scheme does not protect against collusion in T is bounded from above by

$$\frac{\binom{2^m - 2^{r+1}}{t - 2^{r+1}}}{\binom{2^m}{t}} 2^{m-r-1} \frac{\prod_{i=0}^{r}(2^{m-i} - 1)}{\prod_{i=0}^{r}(2^{r+1-i} - 1)}.$$

If $t < 3 \cdot 2^r$, then this bound is tight.

Proof. We fail to protect against a colluding set T if and only if $\dim(D|_T) < |T|$. This latter condition is equivalent to the existence of a codeword of $D^\perp$ whose support is contained in T.
By Corollary 9, there are $2^{m-r-1}\frac{\prod_{i=0}^{r}(2^{m-i}-1)}{\prod_{i=0}^{r}(2^{r+1-i}-1)}$ minimal length codewords in $D^\perp$. Each of these minimal codewords has its support contained in exactly $\binom{2^m-2^{r+1}}{t-2^{r+1}}$ sets of size t, so there exist at most

$$\binom{2^m-2^{r+1}}{t-2^{r+1}}2^{m-r-1}\frac{\prod_{i=0}^{r}(2^{m-i}-1)}{\prod_{i=0}^{r}(2^{r+1-i}-1)} \tag{5}$$

t-sets that contain the support of some codeword in $D^\perp$.
For the second statement, notice that by Lemma 8, the support of two minimum weight codewords of $D^\perp$ intersect in a flat of dimension at most r in $\mathbb{F}_2^m$. Thus, their union has size at least $2\cdot 2^{r+1}-2^r = 3\cdot 2^r$. As a consequence, if $t < 3\cdot 2^r$, then the collections of non-protected sets corresponding to different minimal codewords in $D^\perp$ are disjoint. Thus, the number of such sets is exactly given by (5).

Example 18 *Continuing Example 14, the* 4*-colluding sets* T *that we fail to protect against are in bijection with minimal weight codewords of* $D^\perp$*. By Corollary 9 there are* 120 *minimal weight codewords of* $D^\perp$*. Hence the Reed-Muller PIR scheme protects against collusion for*

$$\left(\tbinom{16}{4}-120\right)\tbinom{16}{4}^{-1} \approx 93.4\% \tag{6}$$

of subsets of servers of size $t = 4$*.*
Similarly, there are $\binom{16}{5} = 4368$ *subsets* T *of servers of size* 5*, of which* 2688 *satisfy* $\dim(D|_T) = 5$*, according to Proposition 17. It follows that the scheme protects against collusion for* $\frac{2688}{4368} \approx 61.5\%$ *of all subsets of servers of size* 5*.*

As a final remark we note that a binary PIR scheme can also be set up with a GRS code over $\mathbb{F}_{2^h}$ for some integer $h > 1$, where every symbol from $\mathbb{F}_{2^h}$ is represented as an element from $\mathbb{F}_2^h$. However, one can easily check that the performance of these codes in terms of protection against colluding sets and symbol download rate is not good. E.g., we can consider a PIR scheme with C the repetition code of length 16 over $\mathbb{F}_2$ and D the binary expansion of an $[4, 1, 4]$-GRS code over $\mathbb{F}_4$, which is a $[16, 4, 4]$-code over $\mathbb{F}_2$. This scheme protects against colluding sets of size 1 and has a code symbol download rate of $\frac{3}{16}$. The Reed-Muller PIR scheme from Example 14 is hence clearly preferrable over this one.

References

[BEP16] S. Blackburn, T. Etzion, and M. Paterson, *PIR schemes with small download complexity and low storage requirements*, ArXiv: 1609.07027, 2016.

[BU16] K. Banawan and S. Ulukus, *The capacity of private information retrieval from coded databases*, ArXiv: 1609.08138, 2016.

[CGKS95] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, *Private information retrieval*, IEEE Annual Symposium on Foundations of Computer Science, 1995, pp. 41–50.

[CKGS98] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, *Private information retrieval*, Journal of the ACM (JACM) **45** (1998), no. 6, 965–981.

[DG15] Z. Dvir and S. Gopi, *2-Server PIR with Sub-Polynomial Communication*, ACM Symposium on Theory of Computing, 2015, pp. 577–584.

[Efr09] K. Efremenko, *3-query locally repairable codes of subexponential length*, ACM Symposium on the Theory of Computing, 2009, pp. 39–44.

[FHGHK16] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. Karpuk, *Private information retrieval from coded databases with colluding servers*, ArXiv: 1611.02062, 2016.

[MS77] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*, North Holland, Amsterdam, 1977.

[SJ16a] H. Sun and S.A. Jafar, *The capacity of private information retrieval*, ArXiv: 1602.09134, 2016.

[SJ16b] ———, *The capacity of robust private information retrieval with colluding databases*, ArXiv: 1605.00635, 2016.

[SJ16c] ———, *The capacity of robust private information retrieval with colluding databases*, ArXiv: 1605.00635, 2016.

[SJ17] ———, *Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti et al*, ArXiv: 1701.07807, 2017.

[SRR14] N. B. Shah, K. V. Rashmi, and K. Ramchandran, *One extra bit of download ensures perfectly private information retrieval*, 2014 IEEE International Symposium on Information Theory, June 2014, pp. 856–890.

[SRRK12] N. B. Shah, K. V. Rashmi, K. Ramchandran, and P. V. Kumar, *Privacy-preserving and secure distributed storage codes*, preprint available at `http://people.eecs.berkeley.edu/~nihar/publications/privacy_security.pdf`, 2012.

[TE16] R. Tajeddine and S. El Rouayheb, *Private information retrieval from MDS coded data in distributed storage systems*, 2016 IEEE International Symposium on Information Theory (ISIT), July 2016, See `http://www.ece.iit.edu/ salim/` for an extended version, pp. 1411–1415.