Tajeddine, Razane; Gnilke, Oliver W.; Karpuk, David; Freij-Hollanti, Ragnar; Hollanti, Camilla

Robust Private Information Retrieval from Coded Systems with Byzantine and Colluding Servers

Published in:
2018 IEEE International Symposium on Information Theory, ISIT 2018

DOI:
10.1109/ISIT.2018.8437670

Published: 15/08/2018

Document Version
Peer reviewed version

Please cite the original version:
Tajeddine, R., Gnilke, O. W., Karpuk, D., Freij-Hollanti, R., & Hollanti, C. (2018). Robust Private Information Retrieval from Coded Systems with Byzantine and Colluding Servers. In *2018 IEEE International Symposium on Information Theory, ISIT 2018* (Vol. 2018-June, pp. 2451-2455). [8437670] IEEE.
https://doi.org/10.1109/ISIT.2018.8437670

Robust Private Information Retrieval from Coded Systems with Byzantine and Colluding Servers

Razane Tajeddine*, Oliver W. Gnilke*, David Karpuk†, Ragnar Freij-Hollanti‡, Camilla Hollanti*‡
* Department of Mathematics and Systems Analysis, Aalto University School of Science, Espoo, Finland
Emails: {razane.tajeddine, oliver.gnilke, camilla.hollanti}@aalto.fi
† Departamento de Matemáticas, Universidad de los Andes, Bogotá, Colombia
Email: da.karpuk@uniandes.edu.co
‡ Department of Electrical and Computer Engineering, Technical University of Munich, Germany
Email: ragnar.freij@tum.fi

***Abstract*—A private information retrieval (PIR) scheme on coded storage systems with colluding, byzantine, and non-responsive servers is presented. Furthermore, the scheme can also be used for symmetric PIR in the same setting.**

An explicit scheme using an $[n, k]$ generalized Reed-Solomon storage code is designed, protecting against t-collusion and handling up to b byzantine and r non-responsive servers, when $n \geq n' = (\nu+1)k + t + 2b + r - 1$, for some integer $\nu \geq 1$. This scheme achieves a PIR rate of $1 - \frac{k+2b+t+r-1}{n'-r}$. In the case where the capacity is known, namely when $k = 1$, it is asymptotically capacity achieving as the number of files grows.

I. Introduction

Private information retrieval (PIR) is concerned with designing schemes for a user to retrieve a certain file from a storage system without revealing the identity of the file to the servers. This problem was introduced by Chor, Goldreich, Kushilevitz and Sudan in [1], where the database was viewed as an m-bit binary string $x = [x^1 \cdots x^m] \in \{0,1\}^m$ from which the user wants to retrieve one bit x^i while keeping the index i hidden from the server. In this work, we consider files $x = [x^1 \cdots x^m]$ encoded and stored on n servers, and assume that the user wants to retrieve a file x^i from the storage system.

One way to achieve privacy is for the user to download all the files from the system. Of course, this scheme has a very high communication cost, or equivalently a very low PIR rate. The rate of a PIR scheme in this model is measured as the ratio of the amount of the gained information over the amount of the total downloaded data, while upload costs of the requests are usually ignored. In case of a single server storing the database, the trivial solution is the only way to guarantee *information-theoretic privacy* [1].

Related work: Initially, PIR constructions served to reduce the total download cost from a storage system with data replicated on multiple servers [2]–[7].

More recently, PIR schemes were constructed on coded data. The authors in [8] show that downloading one extra bit is enough to achieve privacy, if the number of servers is exponential in the number of files. In [9], the authors derive bounds on the tradeoff between storage cost and download cost for linearly coded data. The optimal upper bounds on PIR rate were derived in [10]. For maximum distance separable (MDS) coded data, PIR schemes were presented in [11], [12] that achieve the asymptotic optimal download cost when the servers are non-colluding. For the case of colluding servers, the authors in [13] constructed a new family of PIR schemes on MDS coded data achieving a lower download cost than the ones in [12]. PIR schemes on arbitrary linear storage codes were constructed in [14]. Another line of work is symmetric PIR, which was studied in [15].

In [7], it is shown that the asymptotic PIR capacity for replicated data, as the number of files $m \to \infty$, for a fixed number of colluding servers t, is $1 - \frac{t}{n}$, where n is the number of nodes. When the data is coded using an $[n, k]$ MDS code, it was shown in [10] that the asymptotic capacity is $1 - \frac{k}{n}$. Codes achieving this PIR rate were first presented in [11].

The problem of constructing PIR schemes on replicated data in which some servers can be byzantine (malicious) was considered in [16]–[18]. The asymptotic capacity of PIR on replicated storage systems with t colluding servers and b byzantine servers was found in [19] to be $1 - \frac{2b+t}{n}$. In [20], the authors investigate the problem of providing symmetric PIR from a replicated system with colluding servers and adversaries in the system. An adaptive robust PIR scheme on MDS coded data with non-responsive servers was devised in [21]. A robust PIR scheme on coded data with colluding and byzantine servers was constructed in [22]. PIR from unsynchronized servers was studied in [23], where the files are stored on multiple servers such that some servers are not updated to the latest version, an adaptive PIR scheme is constructed for the user to retrieve privately the file he/she requires. The setting of unsynchronized servers in is similar to the byzantine servers since in both cases, some servers are responding with false, or not entirely true, information, but the work in [23] is more restrictive and uses an adaptive scheme.

Contributions: In this paper, we present a general construction of robust PIR schemes with byzantine servers storing data coded with an arbitrary linear $[n, k, d]$ code. We significantly improve the star product scheme developed in [13] by designing the queries such that the set of responses is an error-correcting code in its own right, allowing the scheme to tolerate malicious and non-responsive servers.

When the storage code is a generalized Reed Solomon code with t colluding servers, r non-responsive servers, and b

byzantine servers, if $n \geq n' = (\nu+1)k + t + 2b + r - 1$, for some integer $\nu \geq 1$, we achieve a PIR rate of

$$\frac{n' - (k + t + 2b + r - 1)}{n' - r}.$$

II. Preliminaries

A. Basic Definitions

We denote the field with q elements by $\mathbb{F}_q$, where q is a prime power, and the set $\{1, 2, \ldots, n\}$ by $[n]$.

By an $[n, k, d]$ code we refer to a code with length n, dimension k, and minimum (Hamming) distance d. MDS-codes are shortly referred to as $[n, k]$ codes, the minimum distance being implied by the fact that MDS-codes achieve the Singleton bound, *i.e.*, $d = n - k + 1$.

A retrieval scheme is said to be *t-private* if the set of queries sent to any t-tuple of servers has zero mutual information with the identity of the desired file. In other words, a set of t colluding servers cannot draw any conclusions about which file the user is downloading.

B. Generalized Reed-Solomon Codes

We propose a PIR scheme for which generalized Reed-Solomon (GRS) storage codes are naturally well-suited. We recall the basic properties of such codes here.

Definition 1 (GRS Codes)**.** *Let $\alpha = [\alpha_1 \cdots \alpha_n] \in \mathbb{F}_q^n$ such that $\alpha_i \neq \alpha_j$ for all $i \neq j$, and $v = [v_1 \cdots v_n] \in (\mathbb{F}_q^\times)^n$. A generalized Reed-Solomon (GRS) code of dimension $k \leq n$ is*

$$GRS_k(\alpha, v) = \{(v_i f(\alpha_i))_{1 \leq i \leq n} | f \in \mathbb{F}_q[x], \deg(f) < k\}.$$

The canonical generator matrix for an $[n, k]$ GRS code is given by

$$G_k(\alpha, v) := \begin{matrix} z^0 \\ z^1 \\ z^2 \\ \vdots \\ z^{k-1} \end{matrix} \begin{pmatrix} 1 & \cdots & 1 \\ \alpha_1 & \cdots & \alpha_n \\ \alpha_1^2 & \cdots & \alpha_n^2 \\ \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \cdots & \alpha_n^{k-1} \end{pmatrix} \cdot \mathrm{diag}(v), \quad (1)$$

where the rows are indexed (in blue) by the function whose evaluation they represent and $\mathrm{diag}(v)$ is the diagonal matrix with the values v_i on the diagonal.

C. Star Products

The star product of two codes will play an integral role in our PIR scheme, essentially determining its rate.

Definition 2. *Consider sub-vector spaces V, W of $\mathbb{F}_q^n$. The star (or Schur) product*

$$V \star W := \langle \{v \star w \ : \ v \in V, w \in W\} \rangle$$

is defined as the linear span of all elements $v \star w = [v_1 w_1 \cdots v_n w_n]$.

For the star product of GRS codes the following useful equality holds

$$GRS_k(\alpha, v) \star GRS_\ell(\alpha, w) = GRS_{\min\{k+\ell-1, n\}}(\alpha, v \star w)$$

III. A PIR Scheme for Colluding Byzantine Servers

A. General Scheme

We begin by describing a general version of our scheme, when $\nu = 1$, that retrieves the file in a single query. This scheme is then extended to a variant when $\nu \geq 1$, where the file can be subdivided into ν stripes that can be retrieved simultaneously in a single query.

We assume for simplicity that information is arranged in a $\nu m \times k$ matrix X, where every ν rows represents one file. Let $G_C \in \mathbb{F}_q^{k \times n}$ be a generator matrix for the storage code C and $G_{C,j}$ be the j^{th} column. Then server j stores the vector $y_j := X \cdot G_{C,j}$. Among the n servers, we assume t servers collude or share information, b servers are byzantine, and r servers are non-responsive. The server is considered to be non-responsive if the wait time for a response is higher than a certain set cutoff.

As in [13], the queries have a random part (that does not depend on the desired file index) and a deterministic part. When $\nu = 1$, the random parts of the queries are generated by sampling m random vectors from a rank t' retrieval code D. Let E be a row vector or, equivalently, a generator matrix of a one-dimensional code. We define a third code, $C^{\star D}_{\star E} = C \star D + C \star E$, by its generator matrix $\binom{G_{C\star D}}{G_{C\star E}}$, and denote by $d_\star$ its minimum distance. Given that

i) the codes $C \star D$ and $C \star E$ intersect only trivially,
ii) $C \star E$ has rank k, and
iii) $d_\star - 1 \geq 2b + r$,

we have the following theorem.

Theorem 1. *Given codes C, D, and a vector E of dimension n satisfying properties i), ii), and iii) above, there is a PIR scheme that downloads an entire file from one query to each of the n servers, and is correct in the presence of up to b byzantine and r non-responsive servers. Furthermore, this scheme is t-private for $t < d_{D^\perp}$, where $d_{D^\perp}$ is the minimum distance of $D^\perp$, the dual code of D.*

Query Construction: Queries are constructed similarly to the scheme in [13]. Let $G_{D,j}$ be the j^{th} column of the generator matrix of D and $U \in \mathbb{F}_q^{m \times t'}$ a random matrix. Then the queries are given as

$$q_j = UG_{D,j} + E(j)e_i, \quad (2)$$

where i is the index of the desired file, $E(j)$ denotes the j^{th} entry of E, and e_i is the i^{th} standard basis vector.

The servers' responses are $r_j = q_j . y_j$ (the inner product of the query and the server contents) for non-byzantine servers, an arbitrary element in $\mathbb{F}_q$ for byzantine servers, and an erasure symbol for non-responsive servers. When a server is considered non-responsive, *e.g.*, after a given queuing-time threshold, the request will be canceled and nothing is downloaded from the server in question[1]. Therefore, the number of downloaded symbols in total from all the servers will be $n - r$.

[1]"Nothing" here can also be thought of as a zero-entropy response telling the server is busy, hence not contributing to the download cost.

We prove that under the assumptions made, the i^{th} row x^i can be retrieved from the response vector, and any t-set of servers gains no information on the index i.

Proof. Privacy: Resistance against t-collusion follows analogously to the proof in [13].

Correctness: We use any decoding algorithm for the code $C_{\star E}^{\star D}$ to recover the vector $\rho = [q_1.y_1 \cdots q_n.y_n]$ from the possibly error and erasure carrying vector we receive from the servers. The condition iii), $d_\star - 1 \geq 2b + r$, guarantees decodability. The vector ρ is an element of $C_{\star E}^{\star D} = C \star D \oplus C \star E$ and therefore has a unique representation as

$$\rho = (z_1 \dots z_{k'} z'_1 \dots z'_k) \begin{pmatrix} G_{C\star D} \\ G_{C\star E} \end{pmatrix} \tag{3}$$

where k' is the dimension of $C \star D$. On the other hand, the form (2) of the queries gives a decomposition of the response vector as

$$\begin{aligned} \rho &\in C \star D + (E(1)e_i.y_1 \dots E(n)e_i.y_n) \\ &= C \star D + (E(1)e_i.XG_{C,1} \dots E(n)e_i.XG_{C,n}) \\ &= C \star D + X_i G_{C\star E}, \end{aligned}$$

where X_i is the i^{th} row of the data matrix. Thus, we have $X_i = [z'_1 \cdots z'_k]$, and the requested i^{th} row is the last k coordinates in the representation (3) of the response. □

Clearly, the condition that $C \star D \cap C \star E = \emptyset$ implies that that $n \geq k' + k + 2b + r$, where k' is the dimension of $C \star D$. If $n \geq k' + \nu k + 2b + r$ for $\nu > 1$ we can straightforwardly extend the scheme to download more than one row, by extending E to be a matrix with ν rows and requiring that

ii*) $C \star E$ has rank νk.

To apply this in order to download a single file, we must assume that the information is arranged in an $m\nu \times k$ matrix X, so that each file is a $\nu \times k$ matrix.

We now give an explicit description of a scheme based on GRS codes for which $k' = k + t - 1$.

B. Explicit Schemes from GRS Codes

Let $C = GRS_k(\alpha, v)$ be an $[n, k]$ GRS code and assume t-collusion, b byzantine servers and r non-responsive servers. Let ν be the maximal integer such that

$$n \geq n' = (\nu + 1)k + t + 2b + r - 1. \tag{4}$$

We will use only n' of the servers, and download ν rows of the database. Let $D = GRS_t(\alpha, w)$ be an $[n', t]$ GRS code on the same evaluation set. The rows of E will be evaluations of monomials of degrees $\mu k + t - 1$ for $1 \leq \mu \leq \nu$ multiplied by the diagonal matrix $\mathrm{diag}(w)$:

$$E = \begin{matrix} z^{k+t-1} \\ z^{2k+t-1} \\ \vdots \\ z^{\nu k+t-1} \end{matrix} \begin{pmatrix} \alpha_1^{k+t-1} & \cdots & \alpha_{n'}^{k+t-1} \\ \alpha_1^{2k+t-1} & \cdots & \alpha_{n'}^{2k+t-1} \\ \vdots & \ddots & \vdots \\ \alpha_1^{\nu k+t-1} & \cdots & \alpha_{n'}^{\nu k+t-1} \end{pmatrix} \cdot \mathrm{diag}(w).$$

We see now that $C \star D = GRS_{k+t-1}(\alpha, v \star w)$ and

$$C \star E = \begin{matrix} z^{k+t-1} \\ z^{k+t} \\ \vdots \\ z^{(\nu+1)k+t-2} \end{matrix} \begin{pmatrix} \alpha_1^{k+t-1} & \cdots & \alpha_{n'}^{k+t-1} \\ \alpha_1^{k+t} & \cdots & \alpha_{n'}^{k+t} \\ \vdots & \ddots & \vdots \\ \alpha_1^{(\nu+1)k+t-2} & \cdots & \alpha_{n'}^{(\nu+1)k+t-2} \end{pmatrix} \cdot \mathrm{diag}(v \star w).$$

Hence the code $C_{\star E}^{\star D} = GRS_{(\nu+1)k+t-1}(\alpha, v \star w)$ is again a GRS code. It is now easy to see that properties i) and ii*) are fulfilled. Furthermore the minimum distance of $C_{\star E}^{\star D}$ is given as $d_\star = n' - (\nu + 1)k - t + 2$ which by (4) implies $d_\star - 1 \geq 2b + r$. We summarize in the following theorem.

Theorem 2. *Let the database be encoded using an $[n, k]$ GRS code C. Assume t-collusion, b byzantine, and r non-responsive servers, and let $n \geq n' = (\nu + 1)k + t + 2b + r - 1$, where ν is maximal. Then we can achieve a rate of*

$$\frac{\nu k}{n' - r} = \frac{\nu k}{(\nu + 1)k + t + 2b - 1}.$$

This is achieved by puncturing, i.e.*, only using n' servers and choosing an $[n', t]$ GRS code D as the query code, and generating the queries as described above.*

We venture the following conjecture.

Conjecture 1. *The PIR capacity $\mathcal{C}$ for t-colluding, $[n, k, d]$ coded storage system with b byzantine and r non-responsive servers satisfies*

$$\lim_{m\to\infty} \mathcal{C} = \frac{n - (k + t + 2b + r - 1)}{n - r}$$

as the number of files m increases.

Let us conclude this section with an example.

Example 1. *Let $n = 13, k = 2, t = 3, b = 2$, and $r = 1$. We design a scheme with rate $R = \frac{1}{3} = \frac{n-(k+t+2b+r-1)}{n-r}$. We encode using a GRS code of length 13, using the evaluation vector $\alpha = (\alpha_1, \dots, \alpha_{13})$ and $v = \mathbf{1}$ the all ones vector:*

$$G_C = \begin{matrix} z^0 \\ z^1 \end{matrix} \begin{pmatrix} 1 & \cdots & 1 \\ \alpha_1 & \cdots & \alpha_{13} \end{pmatrix}.$$

Every file consists of two rows $\begin{pmatrix} x_1^{s,1} & x_2^{s,1} \\ x_1^{s,2} & x_2^{s,2} \end{pmatrix}$ and is encoded into two codewords $\begin{pmatrix} y_1^{s,1} \cdots y_{13}^{s,1} \\ y_1^{s,2} \cdots y_{13}^{s,2} \end{pmatrix}$. We choose D to be $GRS_3(\alpha, \mathbf{1})$ with generator matrix

$$G_D = \begin{matrix} z^0 \\ z^1 \\ z^2 \end{matrix} \begin{pmatrix} 1 & \cdots & 1 \\ \alpha_1 & \cdots & \alpha_{13} \\ \alpha_1^2 & \cdots & \alpha_{13}^2 \end{pmatrix}.$$

The matrix E has to be chosen such that $C_{\star E}^{\star D}$ is an error correcting code that can tolerate 1 erasure and correct up to 2 errors, hence $d_\star - 1 \geq 5$. We pick

$$E = \begin{matrix} z^4 \\ z^6 \end{matrix} \begin{pmatrix} \alpha_1^4 & \cdots & \alpha_{13}^4 \\ \alpha_1^6 & \cdots & \alpha_{13}^6 \end{pmatrix}.$$

The code $C^{\star D}_{\star E}$ is generated by the matrix

$$G_{C^{\star D}_{\star E}} = \begin{matrix} z^0 \\ z^1 \\ z^2 \\ z^3 \\ z^4 \\ z^5 \\ z^6 \\ z^7 \end{matrix} \begin{pmatrix} 1 & \cdots & 1 \\ \alpha_1 & \cdots & \alpha_{13} \\ \alpha_1^2 & \cdots & \alpha_{13}^2 \\ \alpha_1^3 & \cdots & \alpha_{13}^3 \\ \alpha_1^4 & \cdots & \alpha_{13}^4 \\ \alpha_1^5 & \cdots & \alpha_{13}^5 \\ \alpha_1^6 & \cdots & \alpha_{13}^6 \\ \alpha_1^7 & \cdots & \alpha_{13}^7 \end{pmatrix}$$

and is a $[13, 8, 6]$ GRS code. The queries are generated by choosing $2m$ random codewords $d^{s,t}$, $1 \le s \le m$, $1 \le t \le 2$, from D and adding E to the rows corresponding to the requested file i:

$$(q_1 \dots q_n) = \begin{pmatrix} d^{1,1} \\ d^{1,2} \\ \vdots \\ d^{m,1} \\ d^{m,2} \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ E_1 \\ E_2 \\ \vdots \\ 0 \end{pmatrix}$$

The response vector is a codeword in $C^{\star D}_{\star E}$ with up to 2 errors and 1 erasure. We decode to the vector

$$\rho = \sum_{s=1}^{m} d^{s,1} \star y^{s,1} + d^{s,2} \star y^{s,2} + E_1 \star y^{i,1} + E_2 \star y^{i,2}.$$

Since $C \star D$ and $C \star E$ intersect trivially we can separate the part pertaining to file i from the vector ρ. The code $C \star E$ is generated by the last four rows of $G_{C^{\star D}_{\star E}}$. Condition iii) guarantees that we can recover $x_1^{i,1}, x_2^{i,1}, x_1^{i,2}, x_2^{i,2}$. The achieved rate is $R = \dfrac{2k}{n-r} = \dfrac{4}{12} = \dfrac{1}{3}$.

C. Comparison with Previous PIR Schemes

Recently, Zhang and Ge [22] have constructed a PIR scheme for coded data and colluding servers, which is adaptable for non-responsive and byzantine servers (but not for both simultaneously). In this section we briefly compare the rates obtained in this paper with those of [22] in the asymptotic regime as $m \to \infty$. On the one hand, the current scheme is maximally efficient when $n = n' = (\nu+1)k+t+2b+r-1$. On the other hand, the scheme of [22] only achieves positive rates assuming certain inequalities in the basic system parameters are satisfied, namely the obvious inequalities which guarantee that the expressions below in (5) and (6) are positive. To compare the two schemes at their best, we grant both of these assumptions.

When $b = 0$ and $r > 0$, the asymptotic rate as $m \to \infty$ from [22] can be expressed as

$$\bar{R} = \frac{n}{n-r}\left(\frac{\binom{n-r}{k} + \binom{n-t}{k} - \binom{n}{k}}{\binom{n}{k}}\right). \tag{5}$$

An elementary calculation shows that $\bar{R} < \frac{n-(k+t+r-1)}{n-r}$, the rate obtained for the scheme described in the previous sections.

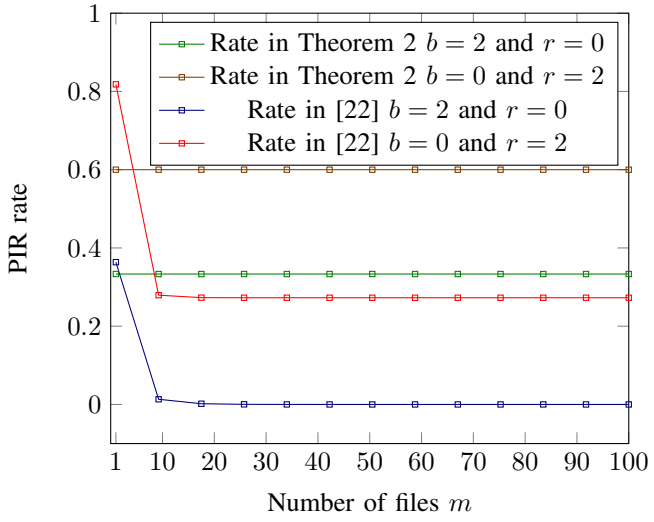

Fig. 1: PIR rate versus number of files m when $n = 12$, $k = 2$, and $t = 3$ following the scheme in [22] and the scheme in this paper.

In the case where $b > 0$ and $r = 0$, the asymptotic rate obtained in [22] is

$$\bar{R} = \frac{2\left(\binom{n-b}{k} - \binom{n}{k}\right) + \binom{n-t}{k}}{\binom{n}{k}} \tag{6}$$

which, again by a simple argument, is less than $\frac{n-(k+t+2b-1)}{n-r}$, the rate obtained by the proposed scheme in this case.

Lastly, we remark that the rates obtained in [22] decrease with an increasing number of files, while the rates we obtain are constant in the number of files. As noted in [22], the rates therein outperform those of [13] for a small number of files. We can see from figure 1 that the same holds here, but we save more precise analysis for an extended version of this paper.

D. A Symmetric Variant

A PIR scheme is *symmetric* if the user, while retrieving the required file x^i, gains no information about any of the other files $x^{i'}$ for $i \neq i'$. To construct a symmetric modification of our scheme, we assume the servers have access to a joint source of randomness. This joint source of randomness outputs a uniform random codeword $s = [s_1 \cdots s_n]$ of $C \star D$, and sends s_j to server j.

All servers then compute $r_j = q_j.y_j + s_j$, which the responsive, non-byzantine servers transmit back to the user. As before, the user receives an erasure symbol from the non-responsive servers, and a random element of $\mathbb{F}_q$ from the Byzantine servers. Also as before, the user decodes in $C^{\star D}_{\star E}$ to obtain the vector ρ as in (3), which is now easily seen to be of the form

$$\rho = s' + X_i G_{C \star E} \tag{7}$$

where s' is uniform on $C \star D$ and independent of any file indices. Clearly ρ is independent of the contents of any file $x^{i'}$ for $i' \neq i$, which guarantees symmetry.

IV. Conclusion and Future Work

A PIR scheme was presented in this paper which can simultaneously handle coded data, colluding servers, non-responsive servers, and byzantine servers. The scheme is an extension of previous work on PIR [13] which is based on the star product of linear codes. In the current work, the response from the servers has additional coding-theoretic properties which allow the user to correct for the erasures and errors produced by the non-responsive and byzantine servers. The scheme has rate $1 - \frac{k+2b+t+r-1}{n'-r}$, where $n' = (\nu+1)k + t + 2b + r - 1$ and ν is the largest integer such that $n' \leq n$. The scheme compares favorably to previous schemes which account for non-responsive and byzantine servers, and additionally is easily symmetrizable.

In the upcoming journal version of this paper, we generalize this scheme so that it does not require puncturing, and achieves a rate $\frac{n-(k+t+2b+r-1)}{n-r}$ for any set of parameters such that $k + t + 2b + r - 1 \leq n$. Additionally, when symmetrizing our scheme, we hope to quantify how much randomness are needed for symmetrization. We loosely conjecture that the current version is doing this with maximal efficiency, namely that $\dim C \star D = k + t - 1$ q-ary units of randomness are necessary. We hope to prove this in the extended journal version of the current work.

Acknowledgments

This work is supported in part by the Academy of Finland, under the grants #276031, #282938, and #303819 to C. Hollanti, and by the Technical University of Munich – Institute for Advanced Study, funded by the German Excellence Initiative and the EU 7th Framework Programme under the grant agreement #291763, via a *Hans Fischer Fellowship* held by C. Hollanti.

O. W. Gnilke and R. Tajeddine were visiting the group of Professor Antonia Wachter-Zeh at the Technical University of Munich while this work was carried out, and wish to thank for the hospitality of the LNT Chair and the COD Group.

O. W. Gnilke is partially supported by the Finnish Cultural Foundation.

References

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

[2] A. Beimel and Y. Ishai, "Information-theoretic private information retrieval: A unified construction," in *Automata, Languages and Programming*, pp. 912–926, Springer, 2001.

[3] A. Beimel, Y. Ishai, E. Kushilevitz, and J.-F. Raymond, "Breaking the $o(n^{1/(2k-1)})$ barrier for information-theoretic private information retrieval," in *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pp. 261–270, IEEE, 2002.

[4] Z. Dvir and S. Gopi, "2 server PIR with sub-polynomial communication," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15, (New York, NY, USA), pp. 577–584, ACM, 2015.

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

[6] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Transactions on Information Theory*, pp. 4075 – 4088, 2017.

[7] H. Sun and S. A. Jafar, "The capacity of private information retrieval with colluding databases," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 941–946, IEEE, 2016.

[8] N. Shah, K. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *2014 IEEE International Symposium on Information Theory (ISIT)*, pp. 856–860, IEEE, 2014.

[9] T. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2842–2846, IEEE, June 2015.

[10] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *arXiv preprint arXiv:1609.08138*, 2016.

[11] R. Tajeddine and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," in *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1411–1415, July 2016.

[12] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Transactions on Information Theory*, 2018.

[13] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM Journal on Applied Algebra and Geometry*, vol. 1, no. 1, pp. 647–664, 2017.

[14] S. Kumar, E. Rosnes, and A. Graell I Amat, "Private information retrieval in distributed storage systems using an arbitrary linear code," in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1421–1425, IEEE, 2017.

[15] Q. Wang and M. Skoglund, "Symmetric private information retrieval for MDS coded distributed storage," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2017.

[16] D. Augot, F. Levy-Dit-Vehel, and A. Shikfa, "A storage-efficient and robust private information retrieval scheme allowing few servers," in *Cryptology and Network Security*, pp. 222–239, Springer, 2014.

[17] A. Beimel and Y. Stahl, "Robust information-theoretic private information retrieval," in *Security in Communication Networks*, pp. 326–341, Springer, 2003.

[18] C. Devet, I. Goldberg, and N. Heninger, "Optimally robust private information retrieval.," in *USENIX Security Symposium*, pp. 269–283, 2012.

[19] K. Banawan and S. Ulukus, "The capacity of private information retrieval from byzantine and colluding databases," *arXiv preprint arXiv:1706.01442*, 2017.

[20] Q. Wang and M. Skoglund, "Secure symmetric private information retrieval from colluding databases with adversaries," *arXiv preprint arXiv:1707.02152*, 2017.

[21] R. Tajeddine and S. El Rouayheb, "Robust private information retrieval on coded data," in *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017.

[22] Y. Zhang and G. Ge, "Private information retrieval from mds coded databases with colluding servers under several variant models," *arXiv preprint arXiv:1705.03186*, 2017.

[23] G. Fanti and K. Ramchandran, "Multi-server private information retrieval over unsynchronized databases," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 437–444, 2014.