UDC 004.056.53
doi:10.31799/1684-8853-2021-1-38-44

Articles

Minimum-storage regenerating codes resistant to special adversary

S. A. Kruglik[a,b], *Junior Researcher, orcid.org/0000-0001-9557-5197, stanislav.kruglik@skoltech.ru*
[a]*Skolkovo Institute of Science and Technology, bld. 1, 30, Bolshoy Boulevard, 121205, Moscow, Russian Federation*
[b]*Moscow Institute of Physics and Technology, 9, Institutskiy Per., 141701, Dolgoprudny, Moscow region, Russian Federation*

Introduction: *To deal with temporally unavailable nodes in distributed storage system engineers apply special classes of erasure correction codes. These codes allow repairing temporally unavailable nodes by downloading a small amount of data from the remaining ones. At the same time, there are safety threats in the presence of an eavesdropper.* ***Purpose:*** *To consider a new mathematical model of eavesdropper that has limited access to all nodes in the system and develop codes resistant to it.* ***Methods:*** *Information-theoretic arguments and mixing information symbols with random ones by systematic Reed — Solomon code.* ***Results:*** *We introduced a new mathematical model of eavesdropper with limited access to all nodes in the distributed storage system. Note that the proposed eavesdropper is passive or, in other words, cannot change accessed data. In this paper, we derived parameters of optimal regenerating codes resistant to such adversary as well as give a technique to ensure the necessary resistance. As a result, we obtained the construction of optimal minimum storage regenerating codes resistant against such adversary.* ***Practical relevance:*** *Proposed constructions can provide resistance against a given adversary while ensuring effective data repair.*

Keywords *— distributed systems, MSR array codes, repair of a temporally unavailable node, mathematical model of system, resistance to adversary.*

For citation: Kruglik S. A. Minimum-storage regenerating codes resistant to special adversary. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 1, pp. 38–44. doi:10.31799/1684-8853-2021-1-38-44

Introduction

Distributed storage systems consisting of thousands of individual nodes that stores a portion of users information become de-facto the standard of modern data storage. High expansion of such systems is leveraged by the constant growth of amount of data stored by humanity. Leading technological companies such as Facebook or Google heavily rely on distributed storage systems [1, 2]. One of the most important problem of current version of such systems is drive failures that occur constantly. To handle it system designers employ erasure-correcting codes for efficient repair of temporally unavailable nodes. Despite several node failures are possible the most common scenario is one node failure and the main goal of research community is to develop codes that optimize the recovery of one node failure in different terms. These terms arose from the distributed nature of systems and the necessity to communicate data between several nodes [3, 4]. One of them, called locality, measures the efficiency of recovery in number of nodes accessed during this procedure [5]. Another one, called repair bandwidth, takes into account the total amount of data transmitted to accomplish the repair [6]. Codes optimized by the second measure are called the regenerating codes and are the main focus of this paper.

In our derivations, we consider a distributed storage system that stores in n nodes B independent random symbols uniformly distributed over the finite field $GF(q)$. Each of these nodes has a storage capacity of l symbols (also termed as sub-packetization level in corresponding literature). We encode B symbols by regenerating code in such a way that in case of one node failure the replacement node can repair its content (or function of it in case of functional repair) by connecting to any set of d helper nodes ($d > k - 1$) and downloading β symbols from each of them. The total amount of downloaded data $d\beta$ is termed as repair bandwidth. Also, regenerating code has such a property that any k nodes can recover all B message symbols. Note that in such a case we have to download all content from them.

In the initial paper on regenerating codes [6] authors utilizing network-flow graph established that parameters of these codes must satisfy the following bound

$$B \le \sum_{i=0}^{k-1} \min(l, (d-i)\beta). \qquad (1)$$

It can be deduced from the form of (1) that achieving equality in it while fixed parameters B, k, and d leads to the tradeoff between the repair bandwidth $d\beta$ and the sub-packetization level l. Two extreme points of this tradeoff determine two classes

of regenerating codes — minimum bandwidth regenerating (MBR) codes and minimum storage regenerating (MSR) codes. In the first case, we initially minimize bandwidth and after it minimize storage on each node. There are a lot of constructions of such codes in the literature, see [7–9] and references therein. Unfortunately, known constructions have code rate no more than 1/2 that restricts their practical applications. Another drawback of MBR codes is that there are no constructions with optimal access property, namely we have to access a large amount of data to accomplish the node repair process while transmitting only the function from them. In case of MSR codes that are the main focus of this paper we first minimize storage on each node and after it the bandwidth. These codes have many advantages over MBR codes, namely there are explicit constructions of high-rate MSR array codes as well as constructions of such codes with optimal-access property. The latter means that in case of node repair we only have to access helper node symbols transmitted to the replacement node. For more details, see papers [9–11] and references in them.

Despite importance of repairing the content of unavailable node, this paper focus on another aspect of distributed storage systems namely safety of stored data. Due to distributed nature of such systems and as a consequence, increasing use of untrusted node providers or communication channels, they are vulnerable to different type of attacks or data leakage [12–14]. In this paper, we focus on threats caused by eavesdropper that gains access to some portion of stored information. The considered eavesdropper also denoted below by E is passive, i. e. E cannot change accessed data. There are two popular approaches to preserve resistance against E. One of them is to use computational cryptography based on difficulty in the computation of some function. Deploying this approach needs to distribute keys as well as provide additional (typically hard) computations that make it irrelevant for distributed storage systems [15]. Another one is an information-theoretic approach in which we mix stored data with random symbols taken uniformly and independent from the same alphabet. In such a case, we ensure that eavesdropper gaining access to the limited number of symbols obtain no information about stored content. In other words, we ensure the zero-mutual information between stored content and information available to E [12]. In this paper we focus on information-theoretic approach only. Note that this problem formulation is highly connected with Wire-Tap Channel II in which eavesdropper has an access to any fixed size subset of symbols transmitted through a noiseless channel [16]. Proposed solution based on coset coding provide resistance against such E while ensuring reconstruction of all information content without the possibility of repair part of it. This fact makes it hard to generalize the given solution to the case of regenerating codes that support single node repair.

Recent papers within safety of regenerating codes focused on resistance against eavesdropper with full access to a limited number of nodes. Some papers also consider a stronger adversary with additional access to data transmitted during the repair. This eavesdropper model corresponds to the case then the adversary can control some subset of nodes. There exist corresponding bounds on the amount of information that can be safely stored in such systems as well as constructions attaining them. For more details, we refer to the papers [12, 13, 17].

In this paper, we continue our research initiated in [18] and consider a new mathematical model of eavesdropper that can access the limited number of symbols from each node in the distributed system. As before we aim to ensure zero mutual information between stored data and data available to E. We consider the minimum storage regenerating codes with optimal access property and derive the technique to make it resistant against given eavesdropper. Note that such consideration is enough natural as these codes ensure node recovery while accessing a small portion of symbols from any node in a given helper set.

The main contribution of this paper is as follows. We consider a new mathematical model of eavesdropper with limited access to all nodes in the distributed storage system, give a bound on parameters of regenerating codes resistant against such adversary as well as propose an explicit construction of MSR-array codes with optimal access property secure against it.

Preliminaries

Within this paper, we use the following notations. By $GF(q)$ we define the finite field with q elements and by $\mathbf{X} = (X_1, \ldots, X_n)^t$ the column vector with n elements over it. We denote the set of n positions as $[n] = \{0, 1, \ldots, n-1\}$ and define the restriction of column vector $\mathbf{X}$ to its subset T as $\mathbf{X}_T$. By superscript t we mean the transpose operations and by superscript s the parameters of safe version of code construction.

By $H(X)$ we define the entropy of discrete random variable X and by $I(X; Y) = H(X) - H(X|Y)$ the mutual information between discrete random variables X and Y. $H(X|Y)$ denote the conditional entropy of random variable X given random variable Y. The same is held for vectors consisting of discrete random variables.

Within this paper, we consider MSR-array codes with optimal access property proposed by Ye and

Barg in paper [11]. Such codes attain the extreme point in bound (1) and have the following parameters: $l = B/k$ and $d\beta = \frac{B}{k(d-k+1)}$. The code construction is explained in Construction 1.

Construction 1. Let us construct array code of length n, sub-packetization level $l = r^{n-1} = (n-k)^{n-1}$ land number of nodes necessary to recover information content k. The code is constructed over $GF(q)$ with size more than n and primitive element γ. We consider the case of $d = n-1$ that corresponds to the most common scenario of one node failure. The code is formed from $l \times n$ matrices over $GF(q)$ each encoding kl information symbols. Encoding procedures are defined using parity-check equations in the following form:

$$(\mathbf{C}_1, \ldots, \mathbf{C}_n) : \sum_{i=1}^{n} \mathbf{A}_{t,i}\mathbf{C}_i = 0, \; t = 1, \ldots, r, \qquad (2)$$

where $\mathbf{C}_i = (c_{i,0}, \ldots, c_{i,l-1})^t$ is a column vector that corresponds to l code symbols over $GF(q)$ stored on node i. $\mathbf{A}_{t,i} = \mathbf{A}_i^{t-1}$, where $t = 1, \ldots, r$ and $i = 1, \ldots, n$ are $l \times l$ matrices over $GF(q)$. Note that by forming the first k column vectors $\mathbf{C}_i$ from $B = kl$ information symbols we can determine the remaining $r = n-k$ column vectors. The specific code families can be obtained by choosing different forms of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ such that and $\mathbf{A}_i - \mathbf{A}_j$ is invertible and multiplication of two matrices has commutative property. In our case to obtain MSR codes with optimal access property we choose $\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}$ to be permutation matrices and $\mathbf{A}_n$ to be an identity matrix. In such a case replacement node has to access $\frac{l}{d-k+1}$ symbols from each of d helper nodes to accomplish the node repair.

In such a case we can determine the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}$ as follows:

$$\mathbf{A}_i = \sum_{a=0}^{l-1} \lambda_{i,a_i} \mathbf{e}_a \mathbf{e}_{a(i,a_i+1 \bmod r)}^t, \; i = 1, \ldots, n-1, \qquad (3)$$

where a_i denotes the i-th element from the right in r-ary representation $(a_{n-1}, \ldots, a_1)$ of a. By $a(i, u)$ we define a decimal element that coincides with a in all positions of r-ary representation except position i that is equal to u. $\mathbf{e}_0, \ldots, \mathbf{e}_{l-1}$ is standard basis of $GF(q^l)$ over $GF(q)$. As elements λ_{i,a_i} let us take $\lambda_{i,0} = \gamma^i$ and $\lambda_{i,u} = 1$ for $u = \{1, 2, \ldots, r-1\}$.

To define node repair procedure let us determine $\beta_{i,u,t}$ as follows:

$$\beta_{i,u,0} = 0;$$

$$\beta_{i,u,t} = \prod_{v=u}^{u+(t-1) \bmod r} \lambda_{i,v}, \; t = \{1, \ldots, r-1\}, \qquad (4)$$

where $u = \{0, \ldots, r-1\}$ and $\lambda_{i,v}$ are defined above. The repair of node $i = \{1, 2, \ldots, n-1\}$ can be done by accessing l/r symbols $\{c_{j,a}: j \neq i, a_i = 0\}$ from the remaining $n-1$ nodes and solving the following equations

$$\beta_{i,a_i,t} c_{i,a(i,a_i+t \bmod r)} = \\ = -c_{n,a} - \sum_{j \neq i,n} \beta_{j,a_j,t} c_{j,a(j,a_j+t \bmod r)}. \qquad (5)$$

The repair of node n can be done by accessing l/r symbols $\{c_{j,a}: j \neq n, a_1 + \ldots + a_{n-1} = 0 \bmod r\}$ and solving the following equations

$$c_{n,a} = -\sum_{i=1}^{n-1} \beta_{i,a_i,t} c_{i,a(i,a_i+t \bmod r)}. \qquad (6)$$

The reconstruction of information content can be accomplished by connecting to the set of any k nodes and downloading all information from them. In such a case from equations (5) we can form the system to define the symbols from the remaining $n-k$ nodes and recover users information as symbols from the first k nodes.

Eavesdropper model

In this paper, we consider a mathematical model of eavesdropper that can download up to t elements from each node in the set-up of the previous section. In other words, it means that $\mathbf{E}$ can accessed elements $\mathbf{C}_{i,E_i}$ where $E_i \subseteq [n]$, $(|E_i| < t+1)$ from each column vector $\mathbf{C}_i$ that represents the content stored on node i. We are focused on resistance against eavesdropper from an information-theoretic point of view that means that $\mathbf{E}$ does not gain any information about stored content $\mathbf{S}$ or, in other words, the mutual information between stored content and elements obtained from all servers by $\mathbf{E}$ is equal to zero. This can be written as

$$I(\mathbf{S}; \mathbf{C}_1, \ldots, \mathbf{C}_n) = 0. \qquad (7)$$

In information-theoretic approach we typically mix stored data with random symbols taken uniformly and independent from the same alphabet. There are two common ways to do it within distributed storage set up. The first of them is directly mixing information and random symbols utilizing storage codes. Note that typically it requires additional properties from code but allows to work within the same field. Another one is precoding information and random symbols by maximum rank distance codes, for example, Gabidulin code. In this paper we modify the last approach for our eavesdropper model, namely we encode information content of each node by Reed — Solomon

based scheme that allows recovering part of information content by accessing a limited number of symbols.

It's important to understand the bound on a message size that can be stored in such a system in presence of a given eavesdropper. In paper [17] by information-theoretic argument, it was proven that the number of information symbols B^s stored by regenerating code can be upper bounded as follows

$$B^s \le \sum_{i=1}^{k} \min(l-t,\ (d-i+1)\beta). \qquad (8)$$

Achieving equality in the bound (8) for a given B^s, k, d and t leads to the tradeoff between the repair bandwidth $d\beta$ and the sub-packetization level l. Let us explicitly find the values of MSR point that correspond to the case of minimizing l first and β after it. The corresponding relaxed optimization problem can be stated as

$$l^{\circ}(d,\ \beta) = \min l,$$
$$subject\ to: \sum_{i=1}^{k} \min\left(l-t,\ \left(1-\frac{i-1}{d}\right)d\beta\right) \ge B^s. \qquad (9)$$

Let us introduce $b_i = \left(1-\frac{k-i}{d}\right)d\beta$ and rewrite (9) as

$$l^{\circ}(d,\ \beta) = \min l,$$
$$subject\ to: \sum_{i=1}^{k} \min(l-t,\ b_i) \ge B^s. \qquad (10)$$

It can be easily seen that $C(l) = \sum_{i=1}^{k} \min(b_i,\ l-t)$ is a piecewise-linear function of l and has the following form:

$$c(1) = \begin{cases} 0 & l \in [0;\ t] \\ k(l-t) & l \in [t;\ b_1+t] \\ \vdots & \\ b_1+\ldots+b_{k-1}+l-t & l \in [b_{k-1}+t;\ b_k+t] \\ b_1+\ldots b_k & l \in (b_k+t,\ \infty) \end{cases}. \qquad (11)$$

This function is strictly monotone increasing on the segment $l \in [0,\ b_k+t]$. To find the extreme point of l such that $C(l) \ge B^s$ we simply take $l = C^{-1}(B^s)$ for the first non-zero value of $C(l)$, where $C^{-1}(*)$ is the inverse function of C. As a result, we receive

$$l = \frac{B^s}{k} + t. \qquad (12)$$

In this case $B^s = kb_1$ that leads to

$$\beta = \frac{B^s}{k(d-k+1)}. \qquad (13)$$

By the similar argument for MBR case we have

$$l = d\beta + t,$$
$$\beta = \frac{2B^s}{k(2d-k+1)}. \qquad (14)$$

Note that this optimization is the main focus of this paper. We shall say that code resistant against eavesdropper is MSR if its parameters coincide with (12) and (13). To construct it we modify Construction 1 of MSR codes without eavesdropper resistance.

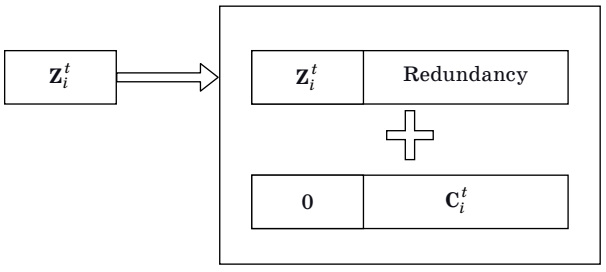
MSR-array codes resistant against eavesdropper

Let us construct MSR-array code resistant against eavesdropper with optimal access property utilizing previously introduced framework. For content $\mathbf{C}_i$ of each node i obtained by Construction 1 let us apply the modified safety scheme based on Reed — Solomon code which was introduced by the first time in paper [19]. Also, we mention paper [20] in which the similar schemes were investigated from another point of view. In that follows to ensure existence of Reed — Solomon codes we assume that we are working in $GF(q)$ with $q > \max(l+t,\ n)$. This scheme is depicted Figure.

In it, we first encode t uniformly and independently distributed random symbols $\mathbf{Z}_i^t = (z_{i,0},\ \ldots,\ z_{i,t-1})$ by systematic Reed — Solomon code of length $l+t$. After it, we add to the last l positions elements $\mathbf{C}_i^t = (c_{i,0},\ \ldots,\ c_{i,l-1})$ of the corresponding node. Defining the obtained row as $\mathbf{Y}_i^t = (y_{i,0},\ \ldots,\ y_{i,t+l-1})$ by the same argument as in [19] we can formally prove that

$$I(\mathbf{Y}_{i,E_i};\ \mathbf{C}_i) = 0 \qquad (15)$$

for any set of $E_i \subseteq [l+t]$ such that ($|E_i| < t+1$).



■ Safety scheme based on Reed — Solomon code

Remark 1. This fact can be understood form the point of view that in Reed — Solomon code any $t - 1$ or less code symbols does not give any information about stored content.

To recover any $0 < r < l + 1$ symbols of $\mathbf{C}_i$ we need to access the first t elements of $\mathbf{Y}_i$ that corresponds to $\mathbf{Z}_i$, encode them by the same Reed — Solomon code and substract necessary redundancy bits from corresponding elements of $\mathbf{Y}_i$. Based on these facts we can formulate the following theorem.

Theorem 1. Let $GF(q)$ be a finite field with $q > \max(l + t, n)$. Then MSR-array code of length n, sub-packetization level $l + t$, number of helper nodes $d = n - 1$ and number of nodes necessary to recover information content k resistant against eavesdropper with access to up to t symbols from any node can be defined by column vectors $\mathbf{Y}_i$. Each $\mathbf{Y}_i$ is formed from vectors $\mathbf{C}_i$ of array-codes from Construction 1 by the modified safety scheme based on Reed — Solomon code with independent and uniformly distributed random symbols $\mathbf{Z}_i$ for each node i.

Proof. From properties of used securing scheme we can write that $I(\mathbf{Y}_{i,E_i}; \mathbf{C}_i) = 0$ for any given $\mathbf{C}_i$ where $E_i \subseteq [l + t]$, ($|E_i| < t + 1$) defines the set of elements from node i available for the eavesdropper. As it holds for any given $\mathbf{C}_i$ and random symbols $\mathbf{Z}_i$ are independent, the elements $\mathbf{Y}_{i,E_i}$ are distributed uniformly and independent over all vectors of length $|E_i|$ over given field $GF(q)$. The last fact leads to $I(\mathbf{C}_j; \mathbf{Y}_{i,E_i}) = 0$. The resistance against eavesdropper means that $I(\mathbf{S}; \mathbf{Y}_{1,E_1}, \ldots, \mathbf{Y}_{n,E_n}) = 0$. As there is a bijection mapping between $\mathbf{C}_1$, ..., $\mathbf{C}_n$ and $\mathbf{S}$ this condition can be reformulated as $I(\mathbf{C}_1, \ldots, \mathbf{C}_n; \mathbf{Y}_{1,E_1}, \ldots, \mathbf{Y}_{n,E_n}) = 0$. Applying the facts above and the chain rule we can easily receive that

$$I(\mathbf{C}_1, \ldots, \mathbf{C}_n; \mathbf{Y}_{1,E_1}, \ldots, \mathbf{Y}_{n,E_n}) = H(\mathbf{Y}_{1,E_1}, \ldots, \mathbf{Y}_{n,E_n}) - \sum_{i=1}^{n} H(\mathbf{Y}_{i,E_i} | \mathbf{C}_i) \le \sum_{i=1}^{n} I(\mathbf{Y}_{i,E_i}; \mathbf{C}_i) = 0. \quad (16)$$

As node repair in Construction 1 is accomplished by downloading l/r symbols from each of $\mathbf{C}_i$ in our construction the replacement node can connect to first t symbols from each $\mathbf{Y}_i$. After it compute the redundancy of Reed — Solomon code, substract it from symbols of $\mathbf{Y}_i$ corresponding to symbols of $\mathbf{C}_i$ necessary for recovery and download only them. In such a case the repair bandwidth as well as sub-packetization level meets the corresponding extreme values (12) and (13). Note that after obtaining the content of failed node the replacement node has to apply to it safety scheme based on Reed — Solomon code.

Reconstruction of information content can be performed in the same way as in Construction 1. The only difference is that after connecting to k servers and downloading all information from them we have to compute the redundancy of Reed — Solomon code that encodes first t symbols and substract it from the last l symbols to obtain corresponding $\mathbf{C}_i$. This ends proof.

Example

To illustrate the proposed framework let us consider the following example. Let us consider $GF(8)$ constructed over primitive polynomial $\varphi(x) = x^3 + x + 1$ with root α. As array-code from Construction 1 let us take code with $n = 3$, $k = 1$, $r = 2$, $l = 4$. The first node stores information symbols while the last two nodes store parity-check symbols. Matrices that form parity-check equations (2) can be written as

$$A_1 = \begin{bmatrix} 0 & \alpha & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha \\ 0 & 0 & 1 & 0 \end{bmatrix}; A_2 = \begin{bmatrix} 0 & 0 & \alpha^2 & 0 \\ 0 & 0 & 0 & \alpha^2 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix};$$

$$A_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (17)$$

If we take as $\mathbf{C}_1 = (1, 1, 1, 1)^t$ when $\mathbf{C}_2 = (\alpha^3, 0, 0, 0)^t$ and $\mathbf{C}_3 = (\alpha, 1, 1, 1)^t$. Let us make obtained code resistant to eavesdropper by the scheme described in the previous section. In such a case, we receive $\mathbf{Y}_1 = (\alpha^6, \alpha^6, \alpha^2, 0, \alpha^6)^t$, $\mathbf{Y}_2 = (\alpha^5, 1, \alpha^5, \alpha^6, \alpha)^t$, $\mathbf{Y}_3 = (\alpha^3, \alpha^4, \alpha, \alpha^5, \alpha^2)^t$. If we have to recover the content of the first node from the remaining one we have to access $\mathbf{Y}_{2,\{0,1,3\}}$ and $\mathbf{Y}_{3,\{0,1,3\}}$. After it, we can find the redundancy of (5, 1) systematic Reed — Solomon code for information symbols $\mathbf{Y}_{2,0}$ and $\mathbf{Y}_{3,0}$. Receiving $(\alpha, \alpha^5, \alpha^6, \alpha)$ and $(\alpha^6, \alpha^3, \alpha^4, \alpha^6)$ as well as corresponding positions from $\mathbf{Y}_{2,\{1,3\}}$ and $\mathbf{Y}_{3,\{1,3\}}$ we obtain $\mathbf{C}_{2,\{0,2\}}$ and $\mathbf{C}_{3,\{0,2\}}$ that form the following parity-check equations

$$\begin{aligned} c_{1,0} + c_{2,0} + c_{3,0} &= 0; \\ \alpha c_{1,1} + \alpha^2 c_{2,2} + c_{3,0} &= 0; \\ c_{1,2} + c_{2,2} + c_{3,2} &= 0; \\ \alpha c_{1,3} + c_{2,0} + c_{3,2} &= 0 \end{aligned} \quad (18)$$

and determine $\mathbf{C}_1 = (1, 1, 1, 1)^t$. After it we have to apply to it introduced safety scheme and obtain $\tilde{\mathbf{Y}}_1 = (\alpha^2, \alpha^4, \alpha^6, \alpha, \alpha^4)^t$.

Conclusion

In this paper, we considered the new mathematical model of passive eavesdropper that has limited access to symbols from each node. We obtained the parameters of regenerating codes reaching extreme points of corresponding bound on the size of the stored message. Also, we proposed the construction of MSR-array codes resistant against the eavesdropper and illustrated the obtained construction by the corresponding example. In further research, we will consider the hybrid eavesdropper model that has a limited access to all nodes together with full access to a small subset of them.

Funding

The reported study was funded by RFBR, projects no. 19-01-00364, 19-37-90022, 20-07-00652 and joint RFBR and JSPS project no. 20-51-50007.

Acknowledgments

Author thanks A. Frolov and G. Kabatiansky for introducing this problem to him and numerous fruitful discussions during work on this paper.

References

1. Aftab U., Siddiqui G. F. Big data augmentation with data warehouse: A survey. *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 2785–2794. doi:10.1109/BigData.2018.8622206
2. Chun B.-G., Dabek F., Haeberlen A., Sit E., Weatherspoon H., Kaashoek M. F., Kubiatowicz J., Morris R. Efficient replica maintenance for distributed storage systems. *3rd Symposium on Networked Systems Design & Implementation*, USENIX Association, 2006, pp. 45–58.
3. Balaji S. B., Krishnan M. N., Vajha M., et al. Erasure coding for distributed storage: an overview. *Science China Information Sciences*, 2018, vol. 61, pp. 1–45. doi:10.1007/s11432-018-9482-6
4. Kruglik S., Frolov A. An information-theoretic approach for reliable distributed storage systems. *Journal of Communications Technology and Electronics*, 2020, vol. 65, no. 12, pp. 1505–1516. doi:10.1134/S1064226920120116
5. Yekhanin S. Locally decodable codes. *Foundations and Trends in Theoretical Computer Science*, 2012, vol. 6, no. 3, pp. 139–255. doi:10.1561/0400000030
6. Dimakis A. G., Godfrey P. B., Wu Y., Wainwright M. J., Ramchandran K. Network coding for distributed storage systems. *IEEE Transactions on Information Theory*, 2010, vol. 56, no. 9, pp. 4539–4551. doi:10.1109/TIT.2010.2054295
7. Han Y. S., Pai H., Zheng R., Varshney P. K. Update-efficient error-correcting product-matrix codes. *IEEE Transactions on Communications*, 2015, vol. 63, no. 6, pp. 1925–1938. doi:10.1109/TCOMM.2015.2424416
8. Lin S., Chung W. Novel repair-by-transfer codes and systematic exact-MBR codes with lower complexities and smaller field sizes. *IEEE Transactions on Parallel and Distributed Systems*, 2014, vol. 25, no. 12, pp. 3232–3241. doi:10.1109/TPDS.2013.2297109
9. Rashmi K. V., Shah N. B., Kumar P. V. Optimal exact-regenerating codes for distributed storage at the MSR and MBR points via a product-matrix construction. *IEEE Transactions on Information Theory*, 2011, vol. 57, no. 8, pp. 5227–5239. doi:10.1109/TIT.2011.2159049
10. Li J., Tang X., Tian C. A generic transformation for optimal repair bandwidth and rebuilding access in MDS codes. *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 1623–1627. doi:10.1109/ISIT.2017.8006804
11. Ye M., Barg A. Explicit constructions of high-rate MDS array codes with optimal repair bandwidth. *IEEE Transactions on Information Theory*, 2017, vol. 63, no. 4, pp. 2001–2014. doi:10.1109/TIT.2017.2661313
12. Kadhe S., Sprintson A. Security for minimum storage regenerating codes and locally repairable codes. *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 1028–1032. doi:10.1109/ISIT.2017.8006684
13. Rawat A. S., Koyluoglu O. O., Silberstein N., Vishwanath S. Secure locally repairable codes for distributed storage systems. *2013 IEEE International Symposium on Information Theory*, IEEE, 2013, pp. 2224–2228. doi:10.1109/ISIT.2013.6620621
14. Agarwal A., Mazumdar A. Security in locally repairable storage. *IEEE Transactions on Information Theory*, 2016, vol. 62, no. 11, pp. 6204–6217. doi:10.1109/TIT.2016.2605118
15. Bian J., Luo S., Li Z., Yang Y. Optimal weakly secure minimum storage regenerating codes scheme. *IEEE Access*, 2019, vol. 7, pp. 151120–151130. doi:10.1109/ACCESS.2019.2947248
16. Ozarow L. H., Wyner A. D. Wire-tap channel II. *AT&T Bell Lab Technical Journal*, 1984, vol. 63, pp. 2135–2157. doi:10.1002/j.1538-7305.1984.tb00072.x
17. Rashmi K. V., Shah N. B., Ramchandran K., Kumar P. V. Information-theoretically secure erasure codes for distributed storage. *IEEE Transactions on Information Theory*, 2018, vol. 64, no. 3, pp. 1621–1646. doi:10.1109/TIT.2017.2769101
18. Kruglik S. Secure MBR array codes in the presence of special type eavesdropper.

19. *Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2020, ruSMART 2020. Lecture Notes in Computer Science*, Springer, 2020, vol. 12526, pp. 1–11. doi:10.1007/978-3-030-65729-1_5
20. Huang W. Coding for security and reliability in distributed system Ph.D. dissertation, California Institute of Technology, 2017. Available at: paradise.caltech.edu/papers/thesis016.pdf (accessed 16 January 2021).
21. Holzbaur L., Kruglik S., Frolov A., Wachter-Zeh A. Secrecy and accessibility in distributed storage. *2020 IEEE Global Communications Conference (GLOBECOM)*, IEEE, 2020, pp. 1–6. doi:10.1109/GLOBECOM42002.2020.9322434

УДК 004.056.53
doi:10.31799/1684-8853-2021-1-38-44

Коды с минимальным хранением, устойчивые к атакам специального типа

С. А. Круглик[а,б], младший научный сотрудник, orcid.org/0000-0001-9557-5197, stanislav.kruglik@skoltech.ru
[а]Сколковский институт науки и технологий, Большой б-р, 30, стр. 1, Москва, 121205, РФ
[б]Московский физико-технический институт, Институтский пер., 9, Долгопрудный, Московская обл., 141701, РФ

Введение: для борьбы с временным или постоянным выходом из строя серверов распределенной системы хранения информации применяются специальные классы кодов, исправляющих стирания. Данные коды позволяют восстановить информацию с временно недоступного узла путем скачивания малого объема информации с других узлов. При этом возникают угрозы защищенности хранимых данных. **Цель:** введение новой математической модели, в которой злоумышленник имеет доступ к небольшому числу символов с каждого узла, и разработка соответствующих кодов, устойчивых к атакам злоумышленника. **Методы:** теоретико-информационный анализ и перемешивание информационных символов со случайными с помощью систематического кода Рида — Соломона. **Результаты:** введена новая математическая модель злоумышленника в распределенной системе хранения информации, имеющего доступ к малому числу символов с каждого узла. Отметим, что рассматривается модель пассивного злоумышленника — «подслушивателя», не способного каким-либо образом видоизменять полученные им данные. Найдены характеристики оптимальных кодов, устойчивых к выходу из строя серверов в распределенной системе хранения информации при наличии злоумышленника, а также построены оптимальные коды-массивы с минимальным хранением, устойчивые к атакам такого рода. **Практическая значимость:** представленная конструкция позволяет сохранить защищенность данных при обеспечении эффективного восстановления пользовательской информации.

Ключевые слова — распределенная система, коды-массивы с минимальным хранением, восстановление недоступного узла, математическая модель системы, устойчивость к действиям злоумышленника.

Финансовая поддержка

Исследование выполнено при поддержке РФФИ в рамках научных проектов № 19-01-00364, 19-37-90022, 20-07-00652, а также РФФИ и ЯОПН в рамках научного проекта № 20-51-50007.

Для цитирования: Kruglik S. A. Minimum-storage regenerating codes resistant to special adversary. *Информационно-управляющие системы*, 2021, № 1, с. 38–44. doi:10.31799/1684-8853-2021-1-38-44
For citation: Kruglik S. A. Minimum-storage regenerating codes resistant to special adversary. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 1, pp. 38–44. doi:10.31799/1684-8853-2021-1-38-44