Distributed Rateless Codes Based on Random Matrices

Yilong Xiao, Haibo Jiang
School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu, China
E-mail: yilongxiao@foxmail.com

***Abstract*—When multiple source node nodes need to transmit their data packets through a single relay to a common destination, distributed rateless codes can be employed. In this paper, a new kind of completely random rateless codes and its distributed encoding process is proposed based on random matrices theory. The proposed method is very easy to implement. The decoding performance is determined by the rank property of random matrices. Both theoretical analysis and simulation results show that the proposed method is effective.**

Keywords-rateless codes; random matrices; distributed encoding process

I. INTRODUCTION

Rateless codes are new class of erasure correcting codes. These codes have the property that, for a fixed number of data packets, any number of code packets can be generated, and the data packets can be recovered with high probability from any subset of the code packets that is only slightly large than the data packets itself.

Up to now, rateless codes such as LT codes [1], Raptor codes [2] etc. have been used in many applications. But most of these application patterns are centralized. It means that all the data packets are first centralized in one location and then encoded into code packets and transmitted to one or more receivers.

In this paper, we focus on the distributed encoding scheme of rateless codes. In this scheme, there are k, $k > 1$ source nodes and each has m data packets. All the km data packets are need to be transmitted to a same destination T through a common relay node N.

To solve this problem, several distributed schemes [3-7] based on LT codes have been proposed. But due to the asymptotic performance of LT codes, all proposed schemes require the value of km must be large enough. Specially, the schemes in [3-5] are only suitable for $k \in \{2, 4\}$.

Different from previous work, in this paper, we first propose a new kind of rateless codes based on random matrices and then the distributed encoding scheme is given. Our method is easy to implement than all other proposed method and does not have parameter limitation.

II. RATELESS CODES BASED ON RANDOM MATIRCES

We describe the encoding process of our new rateless system codes as algorithm 1 (k, $a\ln k/k$), where k and $a\ln k/k$ are parameters.

```
Algorithm 1 (k, alnk/k)
Input: k data packets $D_1, D_2, \ldots, D_k$; constant $a > 1$;
Output: $m > k$ code packets $C_1, C_2, \ldots, C_m$
Begin
  For $i = 1$ to $m$
     1. Generate a random 0-1sequence $p_1, p_2, \ldots, p_k$
        where $\Pr\{p_j = 1\} = a\ln k/k$, $\Pr\{p_j = 0\} = 1\text{-}\Pr\{p_j = 1\}$;
     2. Generate code packet $C_i$
        $C_i = p_1D_1 \oplus p_2D_2 \oplus \ldots \oplus p_kD_k$. ⊕ means XOR operation
End
```

A. ML Decoding Peformance Analysis

Assuming that the decoder receives arbitrary $k + \varepsilon$, $\varepsilon > 0$ code packets from erasure channel. For convenience, we express these $k + \varepsilon$ code packets as C_1, C_2, … , $C_{k+\varepsilon}$.

Since each code packet is a linear combination of some data packets, so each C_i, $i = 1, 2, \ldots, k + \varepsilon$ can be represented as follows:

$$C_i = \mathbf{g}_i \cdot [D_1, D_2, \cdots, D_k]^T \qquad (1)$$

Wherein, $D_1, \ldots, D_k$ represent k unknown data packets, $\mathbf{g}_i$ is the generator row vector of C_i over F_2^k. According to algorithm 1, each element g_{ij}, $j = 1, \ldots, k$ of $\mathbf{g}_i$ is valued independently and identically as

$$\Pr(g_{ij} = r) = \begin{cases} \dfrac{a\ln k}{k}, & r = 1 \\ 1 - \dfrac{a\ln k}{k}, & r = 0 \end{cases} \qquad (2)$$

The $k + \varepsilon$ generator row vectors of $k + \varepsilon$ code packets constitute the $(k + \varepsilon) \times k$ generator matrix $\mathbf{G}_{(k+\varepsilon)\times k}$ of such $k + \varepsilon$ code packets, i.e. $\mathbf{G}_{(k+\varepsilon)\times k} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{k+\varepsilon}]^T$. Thus, such $k + \varepsilon$ code packets can be represented as follows:

$$\begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_{k+\varepsilon} \end{bmatrix} = \mathbf{G}_{(k+\varepsilon)\times k} \cdot \begin{bmatrix} D_1 \\ \vdots \\ D_k \end{bmatrix} \qquad (3)$$

According to ML decoding theory, if generator matrix $\mathbf{G}_{(k+\varepsilon)\times k}$ is full column rank, the system of equations defined in (3) ($D_1, \ldots, D_k$ are unknown) has unique solution. By adopting ML decoding algorithm, $D_1, \ldots, D_k$ can be recovered. Thus, the probability of recover k unknown data

Published by Atlantis Press, Paris, France.

packets from any $k + \varepsilon$ code packets is equal to the probability that the generator matrix $\mathbf{G}_{(k+\varepsilon)\times k}$ of such $k + \varepsilon$ code packets is full column rank.

Let $p^{failure}$ denote the probability that $\mathbf{G}_{(k+\varepsilon)\times k}$ is not full column rank. When equation (3) stands, we have following Theorem l.

Theorem 1: If every element of $\mathbf{G}_{(k+\varepsilon)\times k}$ is valued independently and identically according to equation (2), the upper bound of $p^{failure}$ is:

$$P^{failur} \le \sum_{w=1}^{k} \binom{k}{w} \cdot \left(\sum_{s=0,2,\ldots,2\left\lfloor \frac{w}{2} \right\rfloor} \binom{w}{s} \cdot \left(\frac{a\ln k}{k}\right)^{s} \cdot \left(1-\frac{a\ln k}{k}\right)^{w-s} \right)^{k+\varepsilon} \tag{4}$$

Proof: If the column vectors of $\mathbf{G}_{(k+\varepsilon)\times k}$ is linearly dependent, the matrix $\mathbf{G}_{(k+\varepsilon)\times k}$ will not be full column rank, so:

$$\begin{aligned} P^{failure} &= \Pr\left\{\exists \mathbf{x} \in F_2^k, \mathbf{x} \ne \mathbf{0} : \mathbf{G}_{(k+\varepsilon)\times k} \cdot \mathbf{x}^T = \mathbf{0}\right\} \\ &\le \sum\nolimits_{\mathbf{x} \in F_2^k, \mathbf{x} \ne \mathbf{0}} \Pr\left\{\mathbf{G}_{(k+\varepsilon)\times k} \cdot \mathbf{x}^T = \mathbf{0}\right\} \end{aligned} \tag{5}$$

Let $\mathbf{R}$ be an arbitrary row vector of $\mathbf{G}_{(k+\varepsilon)\times k}$ and w be the number of 1's of vector $\mathbf{x}$. Since every element of $\mathbf{R}$ is independently and is valued 1 according to equation (2), so:

$$\Pr\left\{\mathbf{R}\cdot\mathbf{x}^T = 0\right\} = \sum_{s=0,2,\ldots,2\left\lfloor \frac{w}{2} \right\rfloor} \binom{w}{s} \cdot \left(\frac{a\ln k}{k}\right)^{s} \cdot \left(1-\frac{a\ln k}{k}\right)^{w-s} \tag{6}$$

Because every row vector of $\mathbf{G}_{(k+\varepsilon)\times k}$ is independent, so the probability of $\mathbf{G}_{(k+\varepsilon)\times k}\cdot\mathbf{x}^T = \mathbf{0}$ is

$$\begin{aligned} &\Pr\left\{\mathbf{G}_{(k+\varepsilon)\times k} \cdot \mathbf{x}^T = \mathbf{0}\right\} \\ &= \left(\sum_{s=0,2,\ldots,2\left\lfloor \frac{w}{2} \right\rfloor} \binom{w}{s} \cdot \left(\frac{a\ln k}{k}\right)^{s} \cdot \left(1-\frac{a\ln k}{k}\right)^{w-s} \right)^{k+\varepsilon} \end{aligned} \tag{7}$$

There are $\binom{k}{w}$ possible different $\mathbf{x}$'s for each w. This completes the proof. ■

So, the probability $p^{success}$ of recover k unknown data packets from any $k + \varepsilon$ code packets have following lower bound

$$P^{success} \ge 1 - \sum_{w=1}^{k} \binom{k}{w} \cdot \left(\sum_{s=0,2,\ldots,2\left\lfloor \frac{w}{2} \right\rfloor} \binom{w}{s} \cdot \left(\frac{a\ln k}{k}\right)^{s} \cdot \left(1-\frac{a\ln k}{k}\right)^{w-s} \right)^{k+\varepsilon} \tag{8}$$

III. DISTRIBUTED ENCODING SCHEME

Figure 1 shows the communication network considered in this paper. There are $k \ge 2$ source nodes-$\{S_i : 1 \le i \le k\}$, each has m data packets with fixed size. The total number of data packets at all source nodes is $n = km$.

The k source nodes want to transmit their data packets to the destination T via the relay N. The relay node N is assumed to have limited capability for processing and storage. Also, the communication between the source nodes is not allowed. Further, the link between N and T is lossy, and hence erasure correction capability built into the sequence of packets transmitted to T is needed.

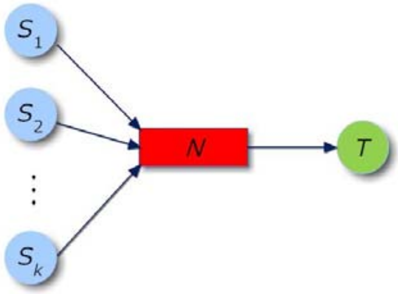

Figure 1. A k-source single-sink network.

In following section, we will describe how to realize above communication requirement by our scheme.

In our scheme, the value of k can be arbitrary odd or even. Else, no matter the value of km is large or small, our scheme has similar performance. Because that our scheme relies on the rank property of random matrix. So we call this scheme Distributed Encoding based on Random Matrices (DERM).

A. DERM Scheme

We conclude the distributed encoding process at each source nodes and the transmission process of DERM scheme as following steps.

```
/* Step 1: */
    For $i$ = 1 to $k$
        Each source node generates one code packet $C_i$
        according to algorithm 1 ($m$, $a\ln k/k$) and transmits it to
        relay $N$.
    End
/* Step 2: */
    Relay $N$ accepts $k$ different coding packets $C_1$ ,..., $C_k$ from
    $k$ source nodes and then computes the sum (bitwise XOR)
    of $C_1$,..., $C_k$. The computing result is denoted by $X_j$.
/* Step 3: */
    Relay $N$ transmit $X_j$ to destination $T$ over erasure channel.
/* Step 4: */
    Destination $T$ try to recover $km$ data packets from $n > km$
    received packets $X_1$, $X_2$, ..., $X_n$ based on ML decoding
    algorithm.
    If Decoding process is succeed
        Relay $N$ inform every source node of stopping
        encoding process.
    Else
        Go to step1.
    End
```

B. *Performance Analysis*

For convenience, we assume that destination T receive $n > km$ packets from relay N which are denoted as $X_1, X_2, \dots, X_n$.

We also assume that the packets set in source node S_i is $D_i = \{d_{i,1}, d_{i,2}, \dots, d_{i,m}\}$, $i = 1, 2, \dots, k$.

Because packet C_i is generated by source node S_i based on algorithm 1 (m, $a\ln k/k$). So

$$C_i = \mathbf{g}_i \cdot \left[d_{i,1},\ d_{i,2}, \cdots,\ d_{i,m}\right]^T \tag{9}$$

In (9), $\mathbf{g}_i$ is the generator row vector of C_i over $F_2{}^m$. Each element of $\mathbf{g}_i$ is valued independently and identically according to (2).

Further, we assume that packet X_1 received by destination T is the sum (bitwise XOR) of $C_1, \dots, C_k$, so we have

$$\begin{aligned} X_1 &= \mathbf{g}_1 \cdot [d_{1,1},\ \dots,\ d_{1,m}]^T \oplus \cdots \oplus \mathbf{g}_k \cdot [d_{m,1},\ \dots,\ d_{m,m}]^T \\ &= [\mathbf{g}_1\ \mathbf{g}_2\ \dots\ \mathbf{g}_k] \cdot \left[d_{1,1}\ \dots\ d_{1,m}\ d_{2,1}\ \dots\ d_{2,m}\ \dots\ d_{m,1}\ \dots\ d_{m,m}\right]^T \\ &= \mathbf{g}1 \cdot \left[d_{1,1}\ \dots\ d_{1,m}\ d_{2,1}\ \dots\ d_{2,m}\ \dots\ d_{m,1}\ \dots\ d_{m,m}\right]^T \end{aligned} \tag{10}$$

In (10), **g1** is the generator row vector of X_1 over $F_2{}^k$. Each element of **g1** is valued independently and identically according to (2).

The n generator row vector **g1**, **g2**, … , ***gn*** of $X_1, X_2, \dots, X_n$ form a generator matrix $\mathbf{G}_{n\times k} = [\mathbf{g1}, \mathbf{g2}, \dots, \mathbf{gn}]^T$ which has same property to the one in (3).

Thus, the probability that destination T can recover km data packets at k source nodes from packets $X_1, X_2, \dots, X_n$ can be described by inequality (8) when $n = k + \varepsilon$.

IV. Simulation Results

In this section, we investigate the performance of our DERM scheme by simulation for different parameters k, m, and a. The main performance metric is how many packets must be received for destination T to successfully recover the km data packets in k source nodes.

We define the recovery overhead $h = n - km$.

The simulation results are listed in figure 2 to 5.

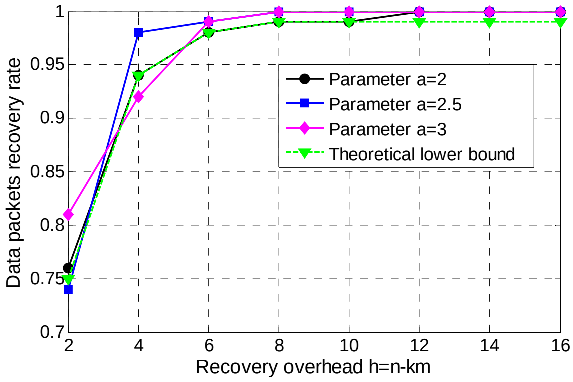

Figure 2. Decoding performance of DERM scheme for $k = 5$, $m = 20$.

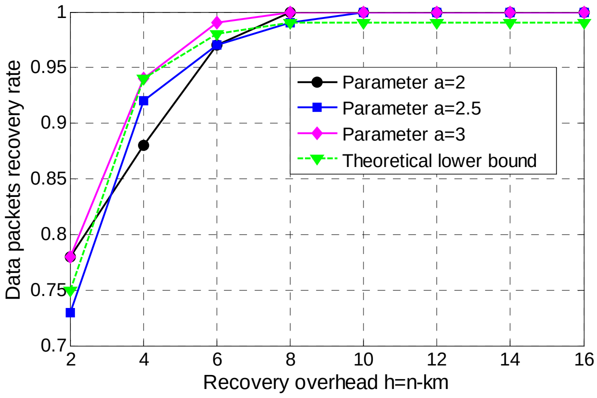

Figure 3. Decoding performance of DERM scheme for $k = 5$, $m = 40$.

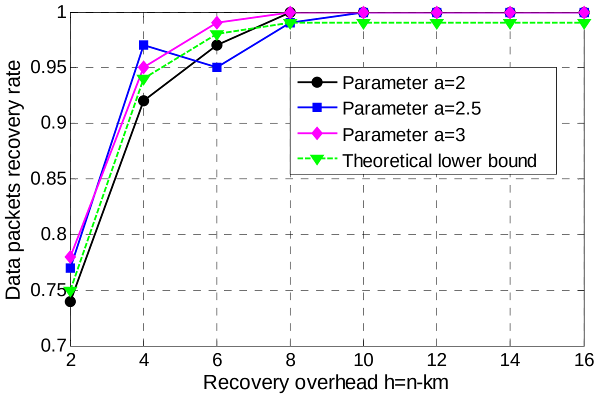

Figure 4. Decoding performance of DERM scheme for $k = 5$, $m = 100$.

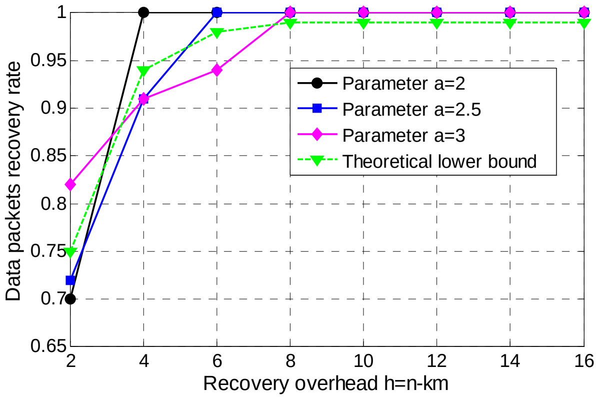

Figure 5. Decoding performance of DERM scheme for $k = 10$, $m = 100$.

From figure 2 to figure 5 we know that when the value of parameter a in algorithm 1 is bigger than 2.5, no matter what the value of km is, $km + 10$ received packets are enough for destination T to recover km data packets in k source nodes. The simulation results approximately equal to the theoretical lower bound defined by inequality (8). This demonstrates that our distributed encoding method of rateless code is effective.

V. Future Work

In this section, we want to introduce another distributed encoding scheme of rateless codes will be explored in future.

Just like the network model illustrated in figure 6, we assume that source node S want to transmit its n data packets $\{d_1, d_2, \dots, d_n\}$ to receiver $T_1, T_2, \dots, T_k$. In order to reduce the work load of encoding and to improve the generating rate of coding packets, source nodes S first transmit its data packets to node N_1 and N_2, then the encoding process is performed parallelly in N_1 and N_2. We further assume that node N_1 and N_2 have limited storage space, each of them can only store about $n/2$ data packets transmitted from nodes S.

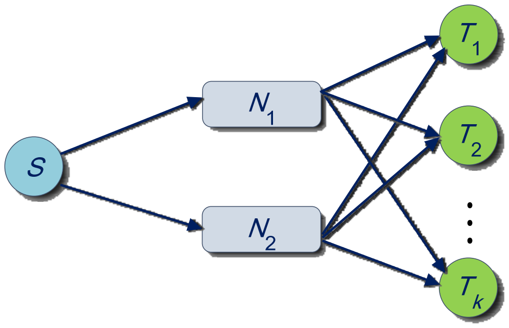

Figure 6. A single-source two-sink network.

We expect that every receiver T_i, $i = 1, 2, \dots, k$ can recover the n data packets $\{d_1, d_2, \dots, d_n\}$ from any $n + \varepsilon$ code packets generated in N_1 and N_2.

One possible way is that source node S transmits $n/2$ data packets $\{d_1, d_2, \dots, d_{n/2}\}$ to N_1 and another $n/2$ data packets $\{d_{n/2+1}, d_2, \dots, d_n\}$ to N_2, then N_1 and N_2 employ encoding algorithm 1 for its data packets respectively. In this way, every receiver T_i, $i = 1, 2, \dots, k$ need to totally receive $(n/2 + \varepsilon) + (n/2 + \varepsilon) = n + 2\varepsilon$ code packets to recover the original n data packets. But we hope that $n + \varepsilon$ received code packets are enough, so this way can't attain our expectation. In future work, we want to solve this problem.

VI. Conclusion

In this paper, design of distributed encoding scheme of rateless nodes is considered. We first provide a new kind of rateless codes based on random matrices. Then we give and analyze the distributed encoding scheme of new rateless code. Our scheme has no limitation to the number of source nodes and the number of total data packets needed to transmit compared to previous relative work. Simulation results show that our scheme fulfill the expected performance.

Acknowledgment

This research is financially supported by Major State Basic Research Development Program of China (2011CB302402), National High Technology Research and Development Program of China (2008AA10Z402) and Sichuan Natural Science Foundation (2008GZ0001). We wish to thank all the people for their contributions to this paper.

References

[1] M. Luby, “LT codes,” Proc. The 43rd Annual IEEE Symp. on Foundations of Computer Science (FCS 02), IEEE Press, November 2002, pp. 271–280.

[2] A. Shokrollahi, “Raptor codes,” IEEE Transaction on Information Theory, vol. 52, no 6, pp. 2551–2567, 2006.

[3] S. Puducheri, J. Kliewer, and T. E. Fuja, “The design and performance of distributed LT codes,” IEEE Transaction on Information Theory, vol. 53, no. 10, pp. 3740-3754, 2007.

[4] A. Talari and N. Rahnavard, “Distributed unequal error protection rateless codes over erasure channels: A two-source scenario,” IEEE Transaction on Communications, vol. 60, no. 8, pp. 2084-2090, 2012.

[5] A. Talari and N. Rahnavard, “Distributed rateless codes with UEP property,” Proc. IEEE International Symp. Information Theory (ISIT 2009), June 2010, pp. 2453-2457.

[6] D. Sejdinovic, R. J. Piechocki, A. Doufexi, and M. Ismail, "Decentralised distributed fountain coding: asymptotic analysis and design," IEEE Communications Letters, vol. 14, no. 1, pp. 42-44, 2010.

[7] D. Sejdinovic, R. J. Piechocki, and A. Doufexi, “AND-OR tree analysis of distributed LT codes,” Proc. IEEE Information Theory Workshop on Networking and Information Theory (ITW 2009), IEEE Press, June 2009, pp. 261–265.