

# Optimal $k$ -Deletion Correcting Codes

Jin Sima and Jehoshua Bruck

Department of Electrical Engineering, California Institute of Technology, Pasadena 91125, CA, USA

## Abstract

Levenshtein introduced the problem of constructing  $k$ -deletion correcting codes in 1966, proved that the optimal redundancy of those codes is  $O(k \log N)$ , and proposed an optimal redundancy single-deletion correcting code (using the so-called VT construction). However, the problem of constructing optimal redundancy  $k$ -deletion correcting codes remained open. Our key contribution is a solution to this longstanding open problem. We present a  $k$ -deletion correcting code that has redundancy  $8k \log n + o(\log n)$  and encoding/decoding algorithms of complexity  $O(n^{2k+1})$  for constant  $k$ .

## I. INTRODUCTION

A set of binary vectors of length  $N$  is a  $k$ -deletion code (denoted by  $\mathcal{C}$ ) iff any two vectors in  $\mathcal{C}$  do not share a subsequence of length  $N - k$ . The problem of constructing a  $k$ -deletion code was introduced by Levenshtein [1]. He proved that the optimal redundancy (defined as  $N - \log |\mathcal{C}|$ ) is  $O(k \log N)$ . Specifically, it is in the range  $k \log N + o(\log N)$  to  $2k \log N + o(\log N)$ . In addition, he proposed the following optimal construction (the well-known Varshamov-Tenengolts (VT) code [2]):

$$\left\{ (c_1, \dots, c_N) : \sum_{i=1}^N i c_i \equiv 0 \pmod{N+1} \right\}, \quad (1)$$

that is capable of correcting a single deletion with redundancy not more than  $\log(N+1)$  [1]. The encoding/decoding complexity of VT codes is linear in  $N$ . Generalizing the VT construction to correct more than a single deletion was elusive for more than 50 years. In particular, the past approaches [4], [5], [6] result in asymptotic code rates that are bounded away from 1.

A recent breakthrough paper [7] proposed a  $k$ -deletion code construction with  $O(k^2 \log k \log N)$  redundancy and  $O_k(N \log^4 N)$  encoding/decoding complexity. For the case  $k = 2$  deletions, the redundancy was improved in [12], [13]. Specifically, the code in [13] has redundancy of  $7 \log N$  and linear encoding/decoding complexity. The work in [14] considered correction with high probability and proposed a  $k$ -deletion code construction with redundancy  $(k+1)(2k+1) \log N + o(\log N)$  and encoding/decoding complexity  $O(N^{k+1}/\log^{k-1} N)$ . The result for this randomized coding setting was improved in [8], where redundancy  $O(k \log(n/k))$  and complexity  $\text{poly}(n, k)$  were achieved. However, finding a deterministic  $k$ -deletion code construction that achieves the optimal order redundancy  $O(k \log N)$  remained elusive.

Our key contribution is a solution to this longstanding open problem: We present a code construction that achieves  $O(k \log N)$  redundancy and  $O(N^{2k+1})$  encoding/decoding computational complexity (note that the complexity is polynomial in  $N$ ). The following theorem summarizes our main result. We note that in this paper, the optimality of the code is redundancy-wise rather than cardinality-wise, i.e., the result focus on asymptotic rather than exact size of the code. The problem of finding optimal cardinality  $k$  deletion code appears highly nontrivial even for  $k = 1$ .

**Theorem 1.** *For any integer  $n > k$  and  $N = n + 8k \log n + o(\log n)$ , there exists an encoding function  $\mathcal{E} : \{0, 1\}^n \rightarrow \{0, 1\}^N$ , computed in  $O(n^{2k+1})$  time, and a decoding function  $\mathcal{D} : \{0, 1\}^{N-k} \rightarrow \{0, 1\}^n$ , computed in  $O(n^{k+1})$  time, such that for any  $\mathbf{c} \in \{0, 1\}^n$  and subsequence  $\mathbf{d} \in \{0, 1\}^{N-k}$  of  $\mathcal{E}(\mathbf{c})$ , we have that  $\mathcal{D}(\mathbf{d}) = \mathbf{c}$ .*

Recently, an independent work [9] proposed a  $k$  deletion code with  $O(k \log n)$  redundancy and better complexity of  $\text{poly}(n, k)$ . Compare to the constant  $8k \log n$  in this paper, the constant in [9] is not explicitly given and is at least  $200k \log n$ . Moreover, the approaches in [9] and this paper are different.

Next we identify and describe our key ideas. The key building blocks in our code construction are: (i) *generalizing the VT construction* to  $k$  deletions by considering constrained sequences, (ii) separating the encoded vector to blocks and using *concatenated codes* and (iii) a novel strategy to *separate the vector to blocks by a single pattern*.

In our previous work for 2-deletions codes [13], we *generalized the VT construction*. In particular, we proved that while the higher order parity checks  $\sum_{i=1}^n i^j c_i \pmod{n^j + 1}$ ,  $j = 0, 1, \dots, t$  might not work in general, those parity checks work in the two deletions case when the sequences are constrained to have no adjacent 1's. In this paper we generalize this idea, specifically, the higher order parity checks work for  $k = t/2$  deletions when the sequences we need to protect satisfy the *following constraint*: The distance between any two adjacent 1's is at least  $k$ .

The fact that we can correct  $k$  deletions using the generalization of the VT construction on constrained sequences, enables a *concatenated code construction*, which separates the sequence  $\mathbf{c}$  into small blocks. Each block is protected by an inner

This work was presented in part at the IEEE International Symposium on Information Theory, Paris, France, July 2019.

<sup>1</sup>The notion  $O_k$  denotes *parameterized complexity*, i.e.,  $O_k(N \log^4 N) = f(k)O(N \log^4 N)$  for some function  $f$ .

code, usually a  $k$ -deletion code. All the blocks together are protected by an outer code, for example, a Reed-Solomon code. Separating and identifying the boundaries between blocks is one of the main challenges in the concatenated code construction. The work in [10], [11] resolved this issue by inserting markers between blocks. In [7], an approach that uses occurrences of short subsequences, called patterns, as markers was proposed. The success of decoding in existing approaches requires that the patterns can not be destroyed or generated by  $k$  deletions / insertions.

Here, we improve the redundancy in [7] by using *a single pattern to separate the blocks* and allowing it to be destroyed or generated by deletions / insertions. The pattern, which we call *synchronization pattern*, is a length  $3k + \lceil \log k \rceil + 4$  sequence  $\mathbf{a} = (a_1, \dots, a_{3k + \lceil \log k \rceil + 4})$  satisfying

- $a_{3k+i} = 1$  for  $i \in [0, \lceil \log k \rceil + 4]$ , where  $[0, \lceil \log k \rceil + 4] = \{0, \dots, \lceil \log k \rceil + 4\}$ .
- There does not exist a  $j \in [1, 3k - 1]$ , such that  $a_{j+i} = 1$  for  $i \in [0, \lceil \log k \rceil + 4]$ .

Namely, a *synchronization pattern* is a sequence that ends with  $\lceil \log k \rceil + 5$  consecutive 1's and no other 1 run with length  $\lceil \log k \rceil + 5$  exists. For a sequence  $\mathbf{c} = (c_1, \dots, c_n)$ , define a *synchronization vector*  $\mathbb{1}_{sync}(\mathbf{c}) \in \{0, 1\}^n$  by

$$\mathbb{1}_{sync}(\mathbf{c})_i = \begin{cases} 1, & \text{if } (c_{i-3k+1}, c_{i-3k+2}, \dots, c_{i+\lceil \log k \rceil + 4}) \text{ is a synchronization pattern,} \\ 0, & \text{else.} \end{cases}$$

Note that  $\mathbb{1}_{sync}(\mathbf{c})_i = 0$  for  $i \in [1, 3k - 1]$  and for  $i \in [n - \lceil \log k \rceil - 3, n]$ . It can be seen from the definition that any two consecutive 1 entries in  $\mathbb{1}_{sync}(\mathbf{c})$  have distance at least  $3k$ .

Now we are ready to describe our construction that is a generalization of the VT code. Define the integer vectors

$$\mathbf{m}^{(\ell)} \triangleq (1^\ell, 1^\ell + 2^\ell, \dots, \sum_{j=1}^n j^\ell)$$

for  $\ell \in [0, \dots, 6k]$ , where the  $i$ -th entry of  $\mathbf{m}^{(\ell)}$  is the sum of the  $\ell$ -th powers of the first  $i$  entries. Given a sequence  $\mathbf{c} \in \{0, 1\}^n$  we compute a (VT like) redundancy of dimension  $6k + 1$  as follows:

$$f(\mathbf{c})_\ell \triangleq \mathbf{c} \cdot \mathbf{m}^{(\ell)} \bmod 3kn^{\ell+1}, \quad (2)$$

for  $\ell \in [0, 6k]$ . It will be shown that the vector  $f(\mathbb{1}_{sync}(\mathbf{c}))$  helps protect the *synchronization vector*  $\mathbb{1}_{sync}(\mathbf{c})$  from  $k$  deletions in  $\mathbf{c}$ .

The rest of the paper is organized as follows. Section II provides an outline of our construction and some of the basic lemmas. Section III presents our VT generalization for recovering the *synchronization vector*. Section IV explains how to correct  $k$  deletions based on the *synchronization vector*, when the *synchronization patterns* appear frequently. Section V describes an algorithm to transform a sequence into one with dense *synchronization patterns*. Section VI presents the encoding and decoding of the code. Section VII concludes the paper.

## II. OUTLINE AND PRELIMINARIES

In this section we give an overview of the ingredients that constitute our code construction and the existing results that will be used. For a sequence  $\mathbf{c} \in \{0, 1\}^n$ , define its deletion ball  $B_k(\mathbf{c})$  as the collection of sequences that share a length  $n - k$  subsequence with  $\mathbf{c}$ . We first present a lemma showing that the *synchronization vector*  $\mathbb{1}_{sync}(\mathbf{c})$  can be recovered from  $k$  deletions with the help of  $f(\mathbb{1}_{sync}(\mathbf{c}))$ . Its proof will be given in Section III-A.

**Lemma 1.** *For integers  $n$  and  $k$  and sequences  $\mathbf{c}, \mathbf{c}'$ , if  $\mathbf{c}' \in B_k(\mathbf{c})$  and  $f(\mathbb{1}_{sync}(\mathbf{c})) = f(\mathbb{1}_{sync}(\mathbf{c}'))$ , then  $\mathbb{1}_{sync}(\mathbf{c}) = \mathbb{1}_{sync}(\mathbf{c}')$ .*

Lemma 1 implies that  $\mathbb{1}_{sync}(\mathbf{c})$  can be protected using  $O(k^2 \log n)$  bit redundancy  $f(\mathbb{1}_{sync}(\mathbf{c}))$ . To further reduce the redundancy and get it down to  $O(k \log n)$ , we apply modulo operations on  $f(\mathbb{1}_{sync}(\mathbf{c}))$ . For an integer vector  $\mathbf{v} = (v_0, \dots, v_{6k})$  that satisfies  $0 \leq v_e < 3kn^{e+1}$ ,  $e \in [0, 6k]$ , let

$$M(\mathbf{v}) = \sum_{e=0}^{6k} v_e \prod_{i=0}^{e-1} 3kn^{i+1} \quad (3)$$

be a one-to-one mapping that maps the vector  $\mathbf{v}$  into an integer  $M(\mathbf{v}) \in [0, (3k)^{6k+1} n^{(3k+1)(6k+1)} - 1]$ . Then we have the following lemma, which will be proved in Section III-B.

**Lemma 2.** *For integers  $n$  and  $k$ , there exists a function  $p : \{0, 1\}^n \rightarrow [1, 2^{2k \log n + o(\log n)}]$ , such that if  $M(f(\mathbb{1}_{sync}(\mathbf{c}))) \equiv M(f(\mathbb{1}_{sync}(\mathbf{c}'))) \bmod p(\mathbf{c})$  for two sequences  $\mathbf{c} \in \{0, 1\}^n$  and  $\mathbf{c}' \in B_k(\mathbf{c})$ , then  $\mathbb{1}_{sync}(\mathbf{c}) = \mathbb{1}_{sync}(\mathbf{c}')$ . Hence if*

$$(M(f(\mathbb{1}_{sync}(\mathbf{c}))) \bmod p(\mathbf{c}), p(\mathbf{c})) = (M(f(\mathbb{1}_{sync}(\mathbf{c}'))) \bmod p(\mathbf{c}'), p(\mathbf{c}'))$$

and  $\mathbf{c}' \in B_k(\mathbf{c})$ , we have that  $\mathbb{1}_{sync}(\mathbf{c}) = \mathbb{1}_{sync}(\mathbf{c}')$ .

Lemma 2 presents a  $4k \log n + o(\log n)$  bit hash for correcting  $\mathbb{1}_{sync}(\mathbf{c})$ . With the knowledge of the *synchronization vector*  $\mathbb{1}_{sync}(\mathbf{c})$ , the next lemma shows that the sequence  $\mathbf{c}$  can be further recovered using another  $4k \log n + o(\log n)$  bit hash, when  $\mathbf{c}$  satisfies a "*k dense*" property. The proof of Lemma 3 will be given in Section IV.

A sequence  $\mathbf{c} \in \{0, 1\}^n$  is said to be *k dense* if the lengths of the 0 runs in  $\mathbb{1}_{sync}(\mathbf{c})$  is at most

$$L \triangleq (\lceil \log k \rceil + 5)2^{\lceil \log k \rceil + 9} \lceil \log n \rceil + (3k + \lceil \log k \rceil + 4)(\lceil \log n \rceil + 9 + \lceil \log k \rceil).$$

For *k dense*  $\mathbf{c}$ , the distance between two consecutive 1 entries in  $\mathbb{1}_{sync}(\mathbf{c})$  is at most  $L + 1$ .

**Lemma 3.** *For integers  $k$  and  $n > k$ , there exists a function  $Hash_k : \{0, 1\}^n \rightarrow \{0, 1\}^{4k \log n + o(\log n)}$ , such that every *k dense* sequence  $\mathbf{c} \in \{0, 1\}^n$  can be recovered, given its synchronization vector  $\mathbb{1}_{sync}(\mathbf{c})$ , its length  $n - k$  subsequence  $\mathbf{d}$ , and  $Hash_k(\mathbf{c})$ .*

Finally, to encode for arbitrary sequence  $\mathbf{c} \in \{0, 1\}^n$ , a mapping that transforms any sequence to a *k dense* sequence is given in the following lemma. The details will be given in Section V-A.

**Lemma 4.** *For integers  $k$  and  $n > k$ , there exists a map  $T : \{0, 1\}^n \rightarrow \{0, 1\}^{n+3k+3\lceil \log k \rceil+15}$ , computable in  $poly(n, k)$  time, such that  $T(\mathbf{c})$  is a *k dense* sequence for  $\mathbf{c} \in \{0, 1\}^n$ . Moreover, the sequence  $\mathbf{c}$  can be recovered from  $T(\mathbf{c})$ .*

The next lemmas are from existing results. Lemma 5 gives a *k deletion* correcting hash function for short sequences. It is an extension of the result in [7]. Lemma 6 is a slight variation of the result in [1]. It shows the equivalence between correcting deletions and correcting deletions and insertions. Lemma 7 (See [15]) gives an upper bound on the number of divisors of a positive integer  $n$ . Lemma 6 and Lemma 7 will be used in proving Lemma 2.

**Lemma 5.** *Let  $k$  be a fixed integer. For integers  $w$  and  $n$ , there exists a hash function  $H : \{0, 1\}^w \rightarrow \{0, 1\}^{\lceil (w/\lceil \log n \rceil) \rceil (2k \log \log n + O(1))}$ , computable in  $O_k((w/\log n)n \log^{2k} n)$  time, such that any sequence  $\mathbf{c} \in \{0, 1\}^w$  can be recovered from its length  $w - k$  subsequence  $\mathbf{d}$  and the hash  $H(\mathbf{c})$ .*

*Proof.* We first show that there exists a hash function  $H' : \{0, 1\}^{\lceil \log n \rceil} \rightarrow \{0, 1\}^{2k \log \log n + O(1)}$ , computable in  $O_k(n \log^{2k} n)$  time, that protects a sequence  $\mathbf{c}' \in \{0, 1\}^{\lceil \log n \rceil}$  from *k deletions*. Note that  $|B_k(\mathbf{c}')| \leq \binom{\lceil \log n \rceil}{k} 2^k \leq 2 \lceil \log n \rceil^{2k}$ . Hence it suffices to use brute force and assign hash values for each possible  $\mathbf{c}'$ , and compare with all sequences in  $B_k(\mathbf{c}')$ . The total complexity is  $O_k(n \log^{2k} n)$  and the size of the hash value  $H(\mathbf{c}')$  is  $\log(|B_k(\mathbf{c}')| + 1) \leq 2k \log \log n + O(1)$ .

Now split  $\mathbf{c}$  into  $\lceil (w/\lceil \log n \rceil) \rceil$  blocks  $c_{(i-1)\lceil \log n \rceil+1}, \dots, c_{i\lceil \log n \rceil}$ ,  $i \in [1, \lceil (w/\lceil \log n \rceil) \rceil]$  of length  $\lceil \log n \rceil$ . If the length of the last block is less than  $\lceil \log n \rceil$ , add zeros to the end of the last block such that its length is  $\lceil \log n \rceil$ . Assign a hash value  $\mathbf{h}_i = H'((c_{(i-1)\lceil \log n \rceil+1}, \dots, c_{i\lceil \log n \rceil}))$ ,  $i \in [1, \lceil (w/\lceil \log n \rceil) \rceil]$  for each block. Let  $H(\mathbf{c}) = (\mathbf{h}_1, \dots, \mathbf{h}_{\lceil (w/\lceil \log n \rceil) \rceil})$  be the concatenation of  $\mathbf{h}_i$  for  $i \in [1, \lceil (w/\lceil \log n \rceil) \rceil]$ . The length of  $H(\mathbf{c})$  is  $\lceil (w/\lceil \log n \rceil) \rceil (2k \log \log n + O(1))$  and the complexity of  $H(\mathbf{c})$  is  $O_k((w/\log n)n \log^{2k} n)$ .

Let  $\mathbf{d}$  be a length  $n - k$  subsequence of  $\mathbf{c}$ . Then  $d_{(i-1)\lceil \log n \rceil+1}, \dots, d_{i\lceil \log n \rceil-k}$  is a length  $\lceil \log n \rceil - k$  subsequence of the block  $c_{(i-1)\lceil \log n \rceil+1}, \dots, c_{i\lceil \log n \rceil}$ . Hence the  $i$ -th block  $c_{(i-1)\lceil \log n \rceil+1}, \dots, c_{i\lceil \log n \rceil}$  can be recovered from  $\mathbf{h}_i$  and  $d_{(i-1)\lceil \log n \rceil+1}, \dots, d_{i\lceil \log n \rceil-k}$ . Therefore,  $\mathbf{c}$  can be recovered given  $\mathbf{d}$  and  $H(\mathbf{c})$  and the proof is done.  $\square$

**Lemma 6.** *Let  $r, s$ , and  $k$  be integers satisfying  $r + s \leq k$ . For sequences  $\mathbf{c}, \mathbf{c}' \in \{0, 1\}^n$ , if  $\mathbf{c}'$  and  $\mathbf{c}$  share a common resulting sequence after  $r$  deletions and  $s$  insertions in both, then  $\mathbf{c}' \in B_k(\mathbf{c})$ .*

**Lemma 7.** *For a positive integer  $n \geq 3$ , the number of divisors of  $n$  is upper bounded by  $2^{1.6 \ln n / (\ln \ln n)}$ .*

The proofs of Lemma 1, Lemma 2, Lemma 3, and Lemma 4 rely on several propositions that will be presented in the next sections. For convenience, a dependency graph for the theorem, lemmas and propositions is given in Fig. 1.

### III. PROTECTING THE SYNCHRONIZATION VECTORS

In this section we present a hash function to protect the synchronization vector  $\mathbb{1}_{sync}(\mathbf{c})$  from *k deletions* and prove Lemma 2. We first prove Lemma 1, based on Proposition 1 and Proposition 2. In Proposition 1 we present a bound on the radius of the  $B_k$  ball for the synchronization vector. In Proposition 2, we prove that the higher order parity check helps correct multiple deletions for sequences in which the distance between two consecutive 1's is at least  $3k$ . Since  $\mathbb{1}_{sync}(\mathbf{c})$  is such a sequence, we conclude that the higher order parity check helps recover  $\mathbb{1}_{sync}(\mathbf{c})$ . After proving Lemma 1, we finally use Lemma 7 to further compress the size of the hash function that protects  $\mathbb{1}_{sync}(\mathbf{c})$  and thus prove Lemma 2.

**Proposition 1.** *For  $\mathbf{c}, \mathbf{c}' \in \{0, 1\}^n$ , if  $\mathbf{c}' \in B_k(\mathbf{c})$ , then  $\mathbb{1}_{sync}(\mathbf{c}') \in B_{3k}(\mathbb{1}_{sync}(\mathbf{c}))$ .*

*Proof.* Since  $\mathbf{c}' \in B_k(\mathbf{c})$ , the sequences  $\mathbf{c}'$  and  $\mathbf{c}$  share a common subsequence after *k deletions* in both. We now show that a single deletion in  $\mathbf{c}$  causes at most two deletions and one insertion in its *synchronization vector*  $\mathbb{1}_{sync}(\mathbf{c})$ . We first show that a deletion in  $\mathbf{c}$  can destroy or generate at most 1 *synchronization pattern*. This is because the deletion occurs within the *synchronization pattern* destroyed or generated. Otherwise the *synchronization pattern* is not destroyed or generated by the deletion. Therefore, we need to consider four cases in total. Let  $\mathbf{d}'$  be the subsequence of  $\mathbf{c}$  after a single deletion.

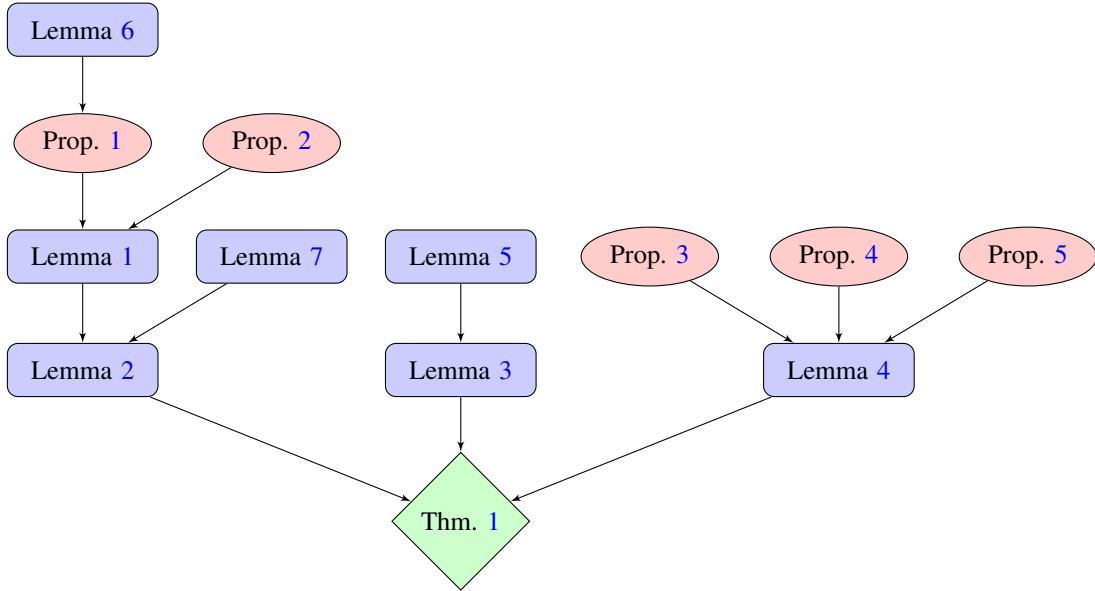


Fig. 1. Dependencies of the claims in the paper.

- 1) The deletion destroys a *synchronization pattern*  $(c_{i+1}, \dots, c_{i+3k+\lceil \log k \rceil + 4})$  for some  $i$  and no *synchronization pattern* is generated. Then the sequence  $\mathbb{1}_{\text{sync}}(\mathbf{d}')$  can be obtained by deleting the 1 entry  $\mathbb{1}_{\text{sync}}(\mathbf{c})_{i+3k}$  in  $\mathbb{1}_{\text{sync}}(\mathbf{c})$ .
- 2) The deletion generates a new *synchronization pattern*  $(c'_{i'+1}, \dots, c'_{i'+3k+\lceil \log k \rceil + 4})$  for some  $i'$  and destroys a *synchronization pattern*  $(c_{i+1}, \dots, c_{i+3k+\lceil \log k \rceil + 4})$ . The sequence  $\mathbb{1}_{\text{sync}}(\mathbf{d}')$  can be obtained by deleting the 1 entry  $\mathbb{1}_{\text{sync}}(\mathbf{c})_{i+3k}$  and the 0 entry  $\mathbb{1}_{\text{sync}}(\mathbf{c})_{i+3k-1}$  in  $\mathbb{1}_{\text{sync}}(\mathbf{c})$  and inserting a 1 entry at  $\mathbb{1}_{\text{sync}}(\mathbf{c})_{i'+3k}$ .
- 3) The deletion generates a new *synchronization pattern*  $(c'_{i'+1}, \dots, c'_{i'+3k+\lceil \log k \rceil + 4})$  for some  $i'$  and no *synchronization pattern* is destroyed. Then the  $\mathbb{1}_{\text{sync}}(\mathbf{d}')$  can be obtained by deleting two 0 entries  $\mathbb{1}_{\text{sync}}(\mathbf{c})_{i'+3k}$  and  $\mathbb{1}_{\text{sync}}(\mathbf{c})_{i'+3k+1}$  in  $\mathbb{1}_{\text{sync}}(\mathbf{c})$  and inserting a 1 entry at  $\mathbb{1}_{\text{sync}}(\mathbf{c})_{i'+3k}$ .
- 4) No *synchronization pattern* is generated or destroyed. Then  $\mathbb{1}_{\text{sync}}(\mathbf{d}')$  can be obtained by deleting a 0 entry  $\mathbb{1}_{\text{sync}}(\mathbf{c})_j$ , where  $j$  is the location of the deletion.

In a summary, in each of the above cases, a single deletion in  $\mathbf{c}$  causes at most two deletions and one insertion in  $\mathbb{1}_{\text{sync}}(\mathbf{c})$ . Hence  $k$  deletions in  $\mathbf{c}$  and  $\mathbf{c}'$  cause at most  $2k$  deletions and  $k$  insertions in  $\mathbb{1}_{\text{sync}}(\mathbf{c})$  and  $\mathbb{1}_{\text{sync}}(\mathbf{c}')$  respectively. According to Lemma 6, we have that  $\mathbb{1}_{\text{sync}}(\mathbf{c}') \in B_{3k}(\mathbb{1}_{\text{sync}}(\mathbf{c}))$  when  $\mathbf{c}' \in B_k(\mathbf{c})$ . Hence Proposition 1 is proved.  $\square$

Let  $\mathcal{R}_m$  be the set of length  $n$  sequences in which the 0 runs have length at least  $m - 1$ , meaning that any two consecutive 1 entries in a sequence  $\mathbf{c} \in \mathcal{R}_m$  have distance at least  $m$ .

**Proposition 2.** For  $\mathbf{c}, \mathbf{c}' \in \mathcal{R}_{3k}$ , if  $\mathbf{c}' \in B_{3k}(\mathbf{c})$  and  $\mathbf{c} \cdot \mathbf{m}^{(e)} = \mathbf{c}' \cdot \mathbf{m}^{(e)}$  (recall that  $f(\mathbf{c})_e = \mathbf{c} \cdot \mathbf{m}^{(e)} \bmod 3kn^{e+1}$ ) for  $e \in [0, 6k]$ , then  $\mathbf{c} = \mathbf{c}'$ .

*Proof.* We first compute the difference  $\mathbf{c} \cdot \mathbf{m}^{(e)} - \mathbf{c}' \cdot \mathbf{m}^{(e)}$ . Since  $\mathbf{c}' \in B_{3k}(\mathbf{c})$ , there exist two subsets  $\delta = \{\delta_1, \dots, \delta_{3k}\} \subset \{1, \dots, n\}$  and  $\delta' = \{\delta'_1, \dots, \delta'_{3k}\} \subset \{1, \dots, n\}$  such that deleting bits with indices  $\delta$  and  $\delta'$  respectively from  $\mathbf{c}$  and  $\mathbf{c}'$  result in the same length  $n - 3k$  subsequence, i.e.,  $(c_i : i \notin \delta) = (c'_i : i \notin \delta')$ . Let  $\Delta = \{i : c_i = 1\}$  and  $\Delta' = \{i : c'_i = 1\}$  be the indices of 1 entries in  $\mathbf{c}$  and  $\mathbf{c}'$  respectively. Let  $S_1 = \Delta \cap \delta$  be the indices of 1 entries that are deleted in  $\mathbf{c}$ . Then  $S_1^c = \Delta \cap ([1, n] \setminus \delta)$  denotes the indices of 1 entries that are not deleted. Similarly let  $S_2 = \Delta' \cap \delta'$  and  $S_2^c = \Delta' \cap ([1, n] \setminus \delta')$  be the indices of 1 entries that are deleted and not in  $\mathbf{c}'$  respectively. Let the elements in  $\delta \cup \delta'$  be ordered by  $1 \leq p_1 \leq p_2 \leq \dots \leq p_{6k} \leq n$ .

Denoting  $p_0 = 0$  and  $p_{6k+1} = n$ , we have that

$$\begin{aligned}
\mathbf{c} \cdot \mathbf{m}^{(e)} - \mathbf{c}' \cdot \mathbf{m}^{(e)} &= \sum_{\ell \in \Delta} \mathbf{m}_\ell^{(e)} - \sum_{\ell \in \Delta'} \mathbf{m}_\ell^{(e)} \\
&= \sum_{\ell \in \Delta} \left( \sum_{i=1}^{\ell} i^e \right) - \sum_{\ell \in \Delta'} \left( \sum_{i=1}^{\ell} i^e \right) \\
&= \sum_{i=1}^n \left( \sum_{\ell \in \Delta \cap [i, n]} i^e \right) - \sum_{i=1}^n \left( \sum_{\ell \in \Delta' \cap [i, n]} i^e \right) \\
&= \sum_{i=1}^n (|\Delta \cap [i, n]| - |\Delta' \cap [i, n]|) i^e \\
&= \sum_{i=1}^n (|S_1 \cap [i, n]| + |S_1^c \cap [i, n]| - |S_2 \cap [i, n]| - |S_2^c \cap [i, n]|) i^e \\
&= \sum_{j=0}^{6k} \sum_{i=p_j+1}^{p_{j+1}} (|S_1 \cap [i, n]| - |S_2 \cap [i, n]| + |S_1^c \cap [i, n]| - |S_2^c \cap [i, n]|) i^e \\
&\stackrel{(1)}{=} \sum_{j=0}^{6k} \sum_{i=p_j+1}^{p_{j+1}} (|S_1 \cap [p_{j+1}, n]| - |S_2 \cap [p_{j+1}, n]| + |S_1^c \cap [i, n]| - |S_2^c \cap [i, n]|) i^e, \tag{4}
\end{aligned}$$

where (1) holds since there is no 1 entry in interval  $(p_j, p_{j+1})$ . In the following we show

1)  $-1 \leq |S_1^c \cap [i, n]| - |S_2^c \cap [i, n]| \leq 1$  for  $i \in [1, n]$ .

2) For each interval  $(p_j, p_{j+1})$ ,  $j = 0, \dots, 6k$ , we have either  $|S_1^c \cap [i, n]| - |S_2^c \cap [i, n]| \leq 0$  for all  $i \in (p_j, p_{j+1})$  or  $|S_1^c \cap [i, n]| - |S_2^c \cap [i, n]| \geq 0$  for all  $i \in (p_j, p_{j+1})$ .

*Proof of (1):* Since deleting bits with indices  $\delta$  in  $\mathbf{c}$  and deleting bits with indices  $\delta'$  in  $\mathbf{c}'$  result in the same subsequence. Hence for every  $i \in S_1^c$ , there is a unique corresponding index  $i' \in S_2^c$  such that the two 1 entries  $c_i$  and  $c'_{i'}$  end in the same location after deletions, i.e.,  $i - |\delta \cap [1, i-1]| = i' - |\delta' \cap [1, i'-1]|$ . This implies that  $|i' - i| \leq 3k$ . Fix integers  $i$  and  $i'$ . Then by definition of  $i$  and  $i'$ , for every  $x \in S_1^c \cap [i+1, n]$  there is a unique corresponding  $y \in S_2^c \cap [i'+1, n]$  such that the two 1 entries  $c_x$  and  $c'_y$  end in the same location after deletions. Hence we have that  $|S_1^c \cap [i+1, n]| = |S_2^c \cap [i'+1, n]|$  and that  $|S_1^c \cap [i, n]| = |S_2^c \cap [i', n]|$ . Without loss of generality assume that  $i' \geq i$ . Then,

$$\begin{aligned}
|S_1^c \cap [i, n]| - |S_2^c \cap [i, n]| &= |S_1^c \cap [i, n]| - |S_2^c \cap [i', n]| - |S_2^c \cap [i+1, i']| \\
&= -|S_2^c \cap [i+1, i']| \\
&\stackrel{(a)}{\geq} -|S_2^c \cap [i+1, i+3k]| \\
&\stackrel{(b)}{\geq} -1,
\end{aligned}$$

where (a) follows from the fact that  $|i' - i| \leq 3k$  and (b) follows from the fact that  $\mathbf{c}, \mathbf{c}' \in \mathcal{R}_{3k}$ . Similarly, when  $i' \leq i$ , we have that  $|S_1^c \cap [i, n]| - |S_2^c \cap [i, n]| \leq 1$ . This completes the proof of (1)

We now prove (2) by contradiction. Supposed on the contrary, there exist  $i_1, i_2 \in (p_j, p_{j+1})$  such that  $i_1 < i_2$  and

$$(|S_1^c \cap [i_1, n]| - |S_2^c \cap [i_1, n]|)(|S_1^c \cap [i_2, n]| - |S_2^c \cap [i_2, n]|) < 0$$

By symmetry it can be assumed that  $|S_1^c \cap [i_1, n]| - |S_2^c \cap [i_1, n]| = -1$  and  $|S_1^c \cap [i_2, n]| - |S_2^c \cap [i_2, n]| = 1$ . Note that  $|S_1^c| = |S_2^c|$  and there exists a one-to-one correspondence between the elements in  $S_1^c$  and the elements in  $S_2^c$ . Hence from  $|S_1^c \cap [i_1, n]| - |S_2^c \cap [i_1, n]| = -1$ , there exist two integers  $x \in S_1^c \cap [i_1-1, n]$  and  $y \in S_2^c \cap [i_1, n]$  such that the two 1 entries  $c_x$  and  $c'_y$  are in the same location after deletions, i.e.,  $x - |\delta \cap [1, x-1]| = y - |\delta' \cap [1, y-1]|$ . Therefore, we have that

$$\begin{aligned}
i_1 - |\delta \cap [1, i_1-1]| &> i_1 - 1 - |\delta \cap [1, i_1-1]| \\
&\geq x - |\delta \cap [1, x-1]| \\
&= y - |\delta' \cap [1, y-1]| \\
&\geq i_1 - |\delta' \cap [1, i_1-1]|,
\end{aligned}$$

which implies that

$$|\delta \cap [1, i_1-1]| < |\delta' \cap [1, i_1-1]|. \tag{5}$$

Similarly, from  $|S_1^c \cap [i_2, n]| - |S_2^c \cap [i_2, n]| = 1$  we have that

$$|\delta \cap [1, i_2 - 1]| > |\delta' \cap [1, i_2 - 1]|. \quad (6)$$

Eq. (5) and Eq. (6) implies that

$$\begin{aligned} |\delta \cap [1, i_2 - 1]| - |\delta \cap [1, i_1 - 1]| &\geq |\delta' \cap [1, i_2 - 1]| + 1 - |\delta' \cap [1, i_1 - 1]| + 1 \\ &\geq 2. \end{aligned} \quad (7)$$

However, since  $i_1, i_2 \in (p_j, p_{j+1}]$ , we have that  $|\delta \cap [1, i_1]| = |\delta \cap [1, i_2 - 1]|$  and  $|\delta' \cap [1, i_1]| = |\delta' \cap [1, i_2 - 1]|$ , which implies that

$$\begin{aligned} |\delta \cap [1, i_2 - 1]| - |\delta \cap [1, i_1 - 1]| &\leq |\delta \cap [1, i_2 - 1]| - |\delta \cap [1, i_1]| + 1 \\ &= 1. \end{aligned}$$

contradicting to Eq. (7). Hence there do not exist different integers  $i_1, i_2 \in (p_j, p_{j+1}]$  such that

$$(|S_1^c \cap [i_1, n]| - |S_2^c \cap [i_1, n]|)(|S_1^c \cap [i_2, n]| - |S_2^c \cap [i_2, n]|) < 0.$$

Hence (2) is proved.

Denote

$$s_i \triangleq |S_1 \cap [i, n]| - |S_2 \cap [i, n]| + |S_1^c \cap [i, n]| - |S_2^c \cap [i, n]|$$

Note that  $|S_1 \cap [i, n]| - |S_2 \cap [i, n]| = |S_1 \cap [p_{j+1}, n]| - |S_2 \cap [p_{j+1}, n]|$  for  $i \in (p_j, p_{j+1}]$ . Hence from (1) and (2) it follows that for each interval  $(p_j, p_{j+1}]$ ,  $j \in \{0, \dots, 6k\}$ , either  $s_i \geq 0$  for all  $i \in (p_j, p_{j+1}]$  or  $s_i \leq 0$  for all  $i \in (p_j, p_{j+1}]$ . Let  $\mathbf{x} = (x_0, \dots, x_{6k}) \in \{-1, 1\}^{6k+1}$  be a vector defined by

$$x_i = \begin{cases} -1, & \text{if } s_j < 0 \text{ for some } j \in (p_i, p_{i+1}] \\ 1, & \text{else.} \end{cases}$$

Then from Eq. (4) we have that

$$\mathbf{c} \cdot \mathbf{m}^{(e)} - \mathbf{c}' \cdot \mathbf{m}^{(e)} = \sum_{j=0}^{6k} \left( \sum_{i=p_j+1}^{p_{j+1}} |s_i| i^e \right) x_j \quad (8)$$

Let  $A$  be a  $(6k+1) \times (6k+1)$  matrix with entries defined by  $A_{e,j} = \sum_{i=p_{j-1}+1}^{p_j} |s_i| i^{e-1}$  for  $e, j \in \{1, \dots, 6k+1\}$ . If  $\mathbf{c} \cdot \mathbf{m}^{(e)} = \mathbf{c}' \cdot \mathbf{m}^{(e)}$  for  $e \in [0, 6k]$ , we have the following linear equation

$$A\mathbf{x} = \begin{bmatrix} \sum_{i=p_0+1}^{p_1} |s_i| i^0 & \cdots & \sum_{i=p_{6k}+1}^{p_{6k+1}} |s_i| i^0 \\ \vdots & \ddots & \vdots \\ \sum_{i=p_0+1}^{p_1} |s_i| i^{6k} & \cdots & \sum_{i=p_{6k}+1}^{p_{6k+1}} |s_i| i^{6k} \end{bmatrix} \begin{bmatrix} x_0 \\ \vdots \\ x_{6k} \end{bmatrix} = 0 \quad (9)$$

We show that this is impossible unless  $A$  is a zero matrix. Suppose on the contrary that  $A$  is nonzero, let  $j_1 < \dots < j_Q$  be the indices of all nonzero columns of  $A$ . Let  $B$  be a submatrix of  $A$ , obtained by choosing the intersection of the first  $Q$  rows and columns with indices  $j_1, \dots, j_Q$ . Then we have that

$$B\mathbf{x}' = \begin{bmatrix} \sum_{i=p_{j_1-1}+1}^{p_{j_1}} |s_i| i^0 & \cdots & \sum_{i=p_{j_Q-1}+1}^{p_{j_Q}} |s_i| i^0 \\ \vdots & \ddots & \vdots \\ \sum_{i=p_{j_1-1}+1}^{p_{j_1}} |s_i| i^{Q-1} & \cdots & \sum_{i=p_{j_Q-1}+1}^{p_{j_Q}} |s_i| i^{Q-1} \end{bmatrix} \begin{bmatrix} x_0 \\ \vdots \\ x_{Q-1} \end{bmatrix} = 0 \quad (10)$$

By the multi-linearity of the determinant,

$$\begin{aligned} \det(B) &= \sum_{i_1 \in (p_{j_1-1}, p_{j_1}], \dots, i_Q \in (p_{j_Q-1}, p_{j_Q})} \det \begin{pmatrix} |s_{i_1}| i_1^0 & \cdots & |s_{i_Q}| i_Q^0 \\ \vdots & \ddots & \vdots \\ |s_{i_1}| i_1^{Q-1} & \cdots & |s_{i_Q}| i_Q^{Q-1} \end{pmatrix} \\ &= \sum_{i_1 \in (p_{j_1-1}, p_{j_1}], \dots, i_Q \in (p_{j_Q-1}, p_{j_Q})} \left[ \prod_{q=1}^Q |s_{i_q}| \prod_{1 \leq m < \ell \leq Q} (i_\ell - i_m) \right] \end{aligned} \quad (11)$$

is positive since  $i_\ell > i_m$  for  $m > \ell$  and for  $i_1 \in (p_{j_1-1}, p_{j_1}], \dots, i_Q \in (p_{j_Q-1}, p_{j_Q}]$ . Note that all the columns of  $B$  are nonzero. Therefore, the linear equation  $B\mathbf{x}' = 0$  does not have nonzero solutions, contradicting to the fact that  $\mathbf{x}' \in \{-1, 1\}^Q$ . Hence  $A$  is a zero matrix, meaning that

$$\begin{aligned} & |S_1 \cap [i, n]| - |S_2 \cap [i, n]| + |S_1^c \cap [i, n]| - |S_2^c \cap [i, n]| \\ & = |\Delta \cap [i, n]| - |\Delta' \cap [i, n]| = 0 \end{aligned}$$

for  $i \in \{1, \dots, n\}$ . This implies  $\Delta = \Delta'$  and thus  $\mathbf{c} = \mathbf{c}'$ . Hence Proposition 2 is proved.  $\square$

#### A. Proof of Lemma 1

From Proposition 1 we have that  $\mathbb{1}_{\text{sync}}(\mathbf{c}') \in B_{3k}(\mathbb{1}_{\text{sync}}(\mathbf{c}))$ . Hence  $(\mathbb{1}_{\text{sync}}(\mathbf{c}')_i, \dots, \mathbb{1}_{\text{sync}}(\mathbf{c}')_n) \in B_{3k}((\mathbb{1}_{\text{sync}}(\mathbf{c})_i, \dots, \mathbb{1}_{\text{sync}}(\mathbf{c})_n))$ . This implies that  $||\Delta \cap [i, n]| - |\Delta' \cap [i, n]|| \leq 3k$ , where  $\Delta = \{i : \mathbb{1}_{\text{sync}}(\mathbf{c})_i = 1\}$  and  $\Delta' = \{i : \mathbb{1}_{\text{sync}}(\mathbf{c}')_i = 1\}$ . Hence According to Eq. (4), we have that

$$\begin{aligned} |\mathbb{1}_{\text{sync}}(\mathbf{c}) \cdot \mathbf{m}^{(e)} - \mathbb{1}_{\text{sync}}(\mathbf{c}') \cdot \mathbf{m}^{(e)}| &= \left| \sum_{i=1}^n (|\Delta \cap [i, n]| - |\Delta' \cap [i, n]|) i^e \right|, \\ &\leq \sum_{i=1}^n 3ki^e \\ &< 3kn^{e+1}. \end{aligned} \quad (12)$$

If  $f(\mathbb{1}_{\text{sync}}(\mathbf{c})) = f(\mathbb{1}_{\text{sync}}(\mathbf{c}'))$ , then  $\mathbb{1}_{\text{sync}}(\mathbf{c}) \cdot \mathbf{m}^{(e)} \equiv \mathbb{1}_{\text{sync}}(\mathbf{c}') \cdot \mathbf{m}^{(e)} \pmod{3kn^{e+1}}$ , which from Eq. (12) implies that  $\mathbb{1}_{\text{sync}}(\mathbf{c}) \cdot \mathbf{m}^{(e)} = \mathbb{1}_{\text{sync}}(\mathbf{c}') \cdot \mathbf{m}^{(e)}$ . Since  $\mathbb{1}_{\text{sync}}(\mathbf{c}') \in B_{3k}(\mathbb{1}_{\text{sync}}(\mathbf{c}))$  and  $\mathbb{1}_{\text{sync}}(\mathbf{c}), \mathbb{1}_{\text{sync}}(\mathbf{c}') \in \mathcal{R}_{3k}$ , from Proposition 2 we conclude that  $\mathbb{1}_{\text{sync}}(\mathbf{c}) = \mathbb{1}_{\text{sync}}(\mathbf{c}')$ . Hence Lemma 1 is proved.

#### B. Proof of Lemma 2

We are now ready to prove Lemma 2. We have shown in Lemma 1 that  $f(\mathbb{1}_{\text{sync}}(\mathbf{c})) \neq f(\mathbb{1}_{\text{sync}}(\mathbf{c}'))$  for  $\mathbf{c}' \in B_k(\mathbf{c}) \setminus \{\mathbf{c}\}$ . Hence

$|M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}))) - M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}')))| \neq 0$  (recall Eq. (3) for definition of function  $M(\mathbf{v})$  for  $\mathbf{c}' \in B_k(\mathbf{c}) \setminus \{\mathbf{c}\}$ ). According to Lemma 7, the number of divisors of  $|M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}))) - M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}')))|$  is upper bounded by

$$2^{2[(3k+1)(6k+1) \ln n + (6k+1) \ln 3k] / \ln((3k+1)(6k+1) \ln n + (6k+1) \ln 3k)} = 2^{o(\log n)}.$$

For any sequence  $\mathbf{c} \in \{0, 1\}^n$ , let

$$\mathcal{P}(\mathbf{c}) = \{p : p \mid |M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}')) - M(f(\mathbb{1}_{\text{sync}}(\mathbf{c})))| \text{ for some } \mathbf{c}' \in B_k(\mathbf{c}) \setminus \{\mathbf{c}\}\}$$

be the set of all divisors of the numbers  $\{|M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}')) - M(f(\mathbb{1}_{\text{sync}}(\mathbf{c})))| : \mathbf{c}' \in B_k(\mathbf{c}) \setminus \{\mathbf{c}\}\}$ . Since  $|B_k(\mathbf{c})| \leq \binom{n}{k} 2^k \leq 2n^{2k}$ , we have that

$$\begin{aligned} |\mathcal{P}(\mathbf{c})| &\leq 2n^{2k} 2^{o(\log n)} \\ &= 2^{2k \log n + o(\log n)}. \end{aligned}$$

Therefore, there exists a number  $p(\mathbf{c}) \in [1, 2^{2k \log n + o(\log n)}]$  such that  $p(\mathbf{c}) \nmid |M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}')) - M(f(\mathbb{1}_{\text{sync}}(\mathbf{c})))|$  for all  $\mathbf{c}' \in B_k(\mathbf{c}) \setminus \{\mathbf{c}\}$ . Hence, if  $M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}')) \equiv M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}))) \pmod{p(\mathbf{c})}$  and  $\mathbf{c}' \in B_k(\mathbf{c})$ , we have that  $p(\mathbf{c}) \mid |M(f(\mathbb{1}_{\text{sync}}(\mathbf{c}')) - M(f(\mathbb{1}_{\text{sync}}(\mathbf{c})))|$  and thus  $\mathbf{c}' = \mathbf{c}$ . This completes the proof of Lemma 2.

### IV. HASH FOR $k$ DENSE SEQUENCES

In this section, we present a hash function for correcting  $k$  deletions in a  $k$ -dense sequence  $\mathbf{c}$ . The encoding/decoding assumes the knowledge of the *synchronization vector*  $\mathbb{1}_{\text{sync}}(\mathbf{c})$  and proves Lemma 3.

Let the positions of the 1 entries in  $\mathbb{1}_{\text{sync}}(\mathbf{c})$  be  $t_1 < t_2 < \dots < t_J$ , where  $J = \sum_{i=1}^n \mathbb{1}_{\text{sync}}(\mathbf{c})_i$  is the number of 1 entries in  $\mathbb{1}_{\text{sync}}(\mathbf{c})$ . Furthermore, let  $t_0 = 0$  and  $t_{J+1} = n + 1$ . Split  $\mathbf{c}$  into blocks  $\mathbf{a}_0, \dots, \mathbf{a}_J$ , where

$$\mathbf{a}_j = (c_{t_j+1}, c_{t_j+2}, \dots, c_{t_{j+1}-1}) \quad (13)$$

for  $j \in [0, J]$ . Since  $\mathbf{c}$  is  $k$ -dense, we have that the length  $|\mathbf{a}_j|$  of  $\mathbf{a}_j$  is not greater than  $L$ . Define the hash function  $\text{Hash}_k$  as follows.

$$\text{Hash}_k(\mathbf{c}) = \text{RS}_{2k}((H(\mathbf{a}_0), \dots, H(\mathbf{a}_J))), \quad (14)$$

where  $\text{RS}_{2k}(\mathbf{c})$  is the redundancy of a systematic Reed-Solomon code (see [16] for details) protecting the length  $J + 1$  sequence  $(H(\mathbf{a}_0), \dots, H(\mathbf{a}_J))$  from  $2k$  symbol substitution errors. Here the symbols are  $H(\mathbf{a}_j)$  (see Lemma 5 for definition



of  $H(\mathbf{c})$ ,  $j \in [0, J]$ , each having alphabet size not greater than  $2^{\lceil(L/\lceil\log n\rceil)\rceil(2k \log \log n + O(1))}$  and can be represented using  $\lceil(L/\lceil\log n\rceil)\rceil(2k \log \log n + O(1))$  bits. The length of  $Hash_k(\mathbf{c})$  is  $\max\{4k \log(J+1), 4k \lceil(L/\lceil\log n\rceil)\rceil(2k \log \log n + O(1))\} = 4k \log n + o(\log n)$ . We now present the following procedure that recovers  $\mathbf{c}$  from its length  $n - k$  subsequence  $\mathbf{d}$ , given the hash function  $Hash_k(\mathbf{c})$  and the synchronization vector  $\mathbb{1}_{sync}(\mathbf{c})$  recovered in Section III.

- 1) **Step 1:** Let  $\mathbb{1}_{sync}(\mathbf{d}) \in \{0, 1\}^{n-k}$  be the *synchronization vector* of  $\mathbf{d}$ . Recall that the locations of 1 entries in  $\mathbb{1}_{sync}(\mathbf{c})$  are  $1 \leq t_1 < \dots < t_J \leq n$ . Let  $t_0 = 0$  and  $t_{J+1} = n + 1$ .
- 2) **Step 2:** Let  $\mathbb{1}_{sync}(\mathbf{d})_0 = \mathbb{1}_{sync}(\mathbf{d})_{n+1-k} = 1$ . For each  $j \in [0, J]$ , if there exist two numbers  $i_j \in [t_j - k, t_j]$  and  $i_{j+1} \in [t_{j+1} - k, t_{j+1}]$  such that  $\mathbb{1}_{sync}(\mathbf{d})_{i_j} = \mathbb{1}_{sync}(\mathbf{d})_{i_{j+1}} = 1$ , let  $\mathbf{a}'_j = (d_{i_j+1}, d_{i_j+2}, \dots, d_{i_{j+1}-1})$ . Else let  $\mathbf{a}'_j = 0$ .
- 3) **Step 3:** Apply the Reed-Solomon decoder to recover  $H(\mathbf{a}_j)$  ( $\mathbf{a}_j$  defined in (13)),  $j \in [0, J]$ , from  $(H(\mathbf{a}'_0), \dots, H(\mathbf{a}'_J), Hash_k(\mathbf{c}))$ .
- 4) **Step 4:** Let  $\mathbf{b}_j = (d_{t_j+1}, d_{t_j+2}, \dots, d_{t_{j+1}-k-1})$ , recover  $\mathbf{a}_j$  by using  $\mathbf{b}_j$  and  $H(\mathbf{a}_j)$ . Then recover  $\mathbf{c}$  from  $\mathbf{a}_j$ ,  $j \in [0, J]$ .

Since  $\mathbf{c}_{t_j} = \mathbb{1}_{sync}(\mathbf{c})_{t_j} = 1$  for  $j \in [1, J]$ , it suffices to show that  $\mathbf{a}_j$ ,  $j \in [0, J]$  can be recovered correctly. Furthermore, note that  $(d_{t_j+1}, \dots, d_{t_{j+1}-k-1})$  is a length  $|\mathbf{a}_j| - k$  subsequence of  $\mathbf{a}_j$ , hence  $\mathbf{a}_j$  can be correctly found given  $\mathbf{b}_j$  and  $H(\mathbf{a}_j)$ ,  $j \in [0, J]$ . It is then left to recover  $H(\mathbf{a}_j)$  for  $j \in [0, J]$ . To this end, we show that there are at most  $2k$  indices  $j$ , such that  $\mathbf{a}'_j \neq \mathbf{a}_j$ . Then there are at most  $2k$  symbol errors in the sequence  $(H(\mathbf{a}_0), \dots, H(\mathbf{a}_J))$ , which can be corrected given the Reed-Solomon code redundancy  $Hash_k(\mathbf{c})$ .

Let  $i_j$ ,  $j \in [1, J]$  be the index of  $\mathbf{c}_{t_j}$  in  $\mathbf{d}$  after deletions, where  $i_j = -1$  if  $\mathbf{c}_{t_j}$  is deleted. Let  $i_0 = 0$  and  $i_{J+1} = n + 1 - k$ . The interval  $[i_j, i_{j+1}]$ ,  $j \in [0, J]$  is called *good* if  $\mathbb{1}_{sync}(\mathbf{d})_{i_j} = \mathbb{1}_{sync}(\mathbf{d})_{i_{j+1}} = 1$  and  $i_{j+1} - i_j = t_{j+1} - t_j$ . Note that by definition,  $\mathbb{1}_{sync}(\mathbf{d})_{i_0} = \mathbb{1}_{sync}(\mathbf{d})_{i_{J+1}} = 1$ . Since  $\mathbf{d}_{i_j} = \mathbf{c}_{t_j}$  and  $i_j \in [t_j - k, t_j]$  for  $j \in [0, J]$ , we have that  $\mathbf{a}'_j = \mathbf{a}_j$  if the interval  $[i_j, i_{j+1}]$  is *good*. Now we show that a deletion can destroy at most 2 *good* intervals. Suppose that the deletion occurs in interval  $[t_j, t_{j+1}]$  in  $\mathbf{c}$  for some  $j$ . If the deletion turns the 1 entry  $\mathbb{1}_{sync}(\mathbf{c})_{t_j}$  to 0, then the 1 entry  $\mathbb{1}_{sync}(\mathbf{c})_{t_{j+1}}$  stays. As a result, at most two *good* intervals  $[i_{j-1}, i_j]$  and  $[i_j, i_{j+1}]$  are destroyed. Similarly, if the deletion turns  $\mathbb{1}_{sync}(\mathbf{c})_{t_{j+1}}$  to 0, then at most two *good* intervals  $[i_j, i_{j+1}]$  and  $[i_{j+1}, i_{j+2}]$  are destroyed. If both 1 entries  $\mathbb{1}_{sync}(\mathbf{c})_{t_j}$  and  $\mathbb{1}_{sync}(\mathbf{c})_{t_{j+1}}$  are not affected, then the deletion destroys only one *good* interval  $[i_j, i_{j+1}]$ . In conclusion, a deletion affects at most two *good* intervals and thus at most  $2k$  block errors  $\mathbf{a}'_j \neq \mathbf{a}_j$  occur. Therefore, the sequence  $\mathbf{c}$  can be recovered and this proves Lemma 3.

## V. TRANSFORMATION TO $k$ DENSE SEQUENCES

In this section we present an algorithm to compute the map  $T(\mathbf{c})$  (see Lemma 4), which transforms any sequence  $\mathbf{c} \in \{0, 1\}^n$  into a  $k$  dense sequence, thus proving Lemma 4. Let  $\mathbf{1}^x$  and  $\mathbf{0}^y$  denote sequences of consecutive  $x$  1's and consecutive  $y$  0's respectively. We will show in Proposition 3 that any sequence  $\mathbf{c}$  satisfying the following two properties is a  $k$  dense sequence. Then, algorithms will be presented in Subsection V-A to generate sequences that satisfy the two properties and thus is  $k$  dense.

**Property 1:** Every length  $B \triangleq (\lceil\log k\rceil + 5)2^{\lceil\log k\rceil+9} \lceil\log n\rceil$  interval of  $\mathbf{c}$  contains the pattern  $\mathbf{1}^{\lceil\log k\rceil+5}$ , i.e., for any integer  $i \in [1, n - B + 1]$ , there exists an integer  $j \in [i, i + B - \lceil\log k\rceil - 5]$  such that  $(c_j, c_{j+1}, \dots, c_{j+\lceil\log k\rceil+4}) = \mathbf{1}^{\lceil\log k\rceil+5}$ .

**property 2:** Every length  $R \triangleq (3k + \lceil\log k\rceil + 4)(\lceil\log n\rceil + 9 + \lceil\log k\rceil)$  interval of  $\mathbf{c}$  contains a length  $3k + \lceil\log k\rceil + 4$  subinterval that does not contain the pattern  $\mathbf{1}^{\lceil\log k\rceil+5}$ , i.e., for any integer  $i \in [1, n - R + 1]$ , there exists an integer  $j \in [i, i + R - 3k - \lceil\log k\rceil - 4]$ , such that  $(c_m, c_{m+1}, \dots, c_{m+\lceil\log k\rceil+4}) \neq \mathbf{1}^{\lceil\log k\rceil+5}$  for every  $m \in [j, j + 3k - 1]$ .

**Proposition 3.** *If a sequence  $\mathbf{c}$  satisfies Property 1 and Property 2, then it is a  $k$  dense sequence.*

*Proof.* Let the locations of the 1 entries in  $\mathbb{1}_{sync}(\mathbf{c})$  be  $t_1 < \dots < t_J$ . Let  $t_0 = 0$  and  $t_{J+1} = n + 1$ . It suffices to show that  $t_{i+1} - t_i \leq B + R + 1 = L + 1$  for any  $i \in [0, J]$ .

According to Property 2, there exists an index  $j^* \in [t_i, t_i + R - 3k - \lceil\log k\rceil - 4]$ , such that  $(c_m, c_{m+1}, \dots, c_{m+\lceil\log k\rceil+4}) \neq \mathbf{1}^{\lceil\log k\rceil+5}$  for every  $m \in [j^*, j^* + 3k - 1]$ . According to Property 1, there exists an integer  $x \in [j^* + 1, j^* + B]$  such that  $(c_x, c_{x+1}, \dots, c_{x+\lceil\log k\rceil+4}) = \mathbf{1}^{\lceil\log k\rceil+5}$ . Let  $\ell = \min_{x \geq j^*, (c_x, c_{x+1}, \dots, c_{x+\lceil\log k\rceil+4}) = \mathbf{1}^{\lceil\log k\rceil+5}} x$ . Then we have that  $\ell \in [j^* + 1, j^* + B]$  and that  $(c_m, c_{m+1}, \dots, c_{m+\lceil\log k\rceil+4}) \neq \mathbf{1}^{\lceil\log k\rceil+5}$  for every  $m \in [j^*, \ell]$ . By definition of  $j^*$ , we have that  $\ell - j^* \geq 3k$ . Since  $(c_\ell, c_{\ell+1}, \dots, c_{\ell+\lceil\log k\rceil+4}) = \mathbf{1}^{\lceil\log k\rceil+5}$ , we have that  $\mathbb{1}_{sync}(\mathbf{c})_\ell = 1$ . Therefore, we conclude that

$$\begin{aligned} t_{i+1} - t_i &\leq \ell - t_i \\ &\leq j^* + B - t_i \\ &\leq R + B + 1 \\ &= L + 1 \end{aligned}$$

□

The following lemma presents a function that outputs a sequence satisfying Property 1.



**Proposition 4.** For integers  $k$  and  $n > k$ , there exists a map  $T_1 : \{0, 1\}^n \rightarrow \{0, 1\}^{n+2\lceil \log k \rceil + 10}$ , computable in  $O(n^2 k \log n \log^2 k)$  time, such that  $T_1(\mathbf{c})$  satisfies Property 1. Moreover, either  $(T_1(\mathbf{c})_{n+\lceil \log k \rceil + 6}, \dots, T_1(\mathbf{c})_{n+2\lceil \log k \rceil + 10}) = \mathbf{1}^{\lceil \log k \rceil + 5}$  or  $(T_1(\mathbf{c})_{n+\lceil \log k \rceil + 5}, \dots, T_1(\mathbf{c})_{n+2\lceil \log k \rceil + 9}) = \mathbf{1}^{\lceil \log k \rceil + 5}$ . The sequence  $\mathbf{c}$  can be recovered from  $T_1(\mathbf{c})$ .

*Proof.* We first show that every sequence  $\mathbf{b} \in \{0, 1\}^B$  containing no consecutive  $\lceil \log k \rceil + 5$  1's can be uniquely represented by a sequence  $\phi(\mathbf{b})$  of length  $B - \lceil \log n \rceil - 2\lceil \log k \rceil - 12$ . Split  $\mathbf{b}$  into  $2^{\lceil \log k \rceil + 9} \lceil \log n \rceil$  blocks of length  $(\lceil \log k \rceil + 5)$ . Since each block is not  $\mathbf{1}^{\lceil \log k \rceil + 5}$ , it can be represented by a symbol of alphabet size  $2^{\lceil \log k \rceil + 5} - 1$ . Therefore, the sequence  $\mathbf{b}$  can be uniquely represented by a sequence  $\mathbf{v}$  of  $2^{\lceil \log k \rceil + 9} \lceil \log n \rceil$  symbols, each having alphabet size  $2^{\lceil \log k \rceil + 5} - 1$ . Convert  $\mathbf{v}$  into a binary sequence  $\phi(\mathbf{b})$ . Then  $\phi(\mathbf{b})$  can be represented by bits with length

$$\begin{aligned} D &\triangleq \log_2(2^{\lceil \log k \rceil + 5} - 1)^{2^{\lceil \log k \rceil + 9} \lceil \log n \rceil} \\ &= \log_2(1 - 1/2^{\lceil \log k \rceil + 5})^{2^{\lceil \log k \rceil + 9} \lceil \log n \rceil} + (\lceil \log k \rceil + 5)2^{\lceil \log k \rceil + 9} \lceil \log n \rceil \\ &= 16 \log_2(1 - 1/2^{\lceil \log k \rceil + 5})^{2^{\lceil \log k \rceil + 5} \lceil \log n \rceil} + B \\ &\stackrel{(a)}{\leq} -16 \log_2 e \lceil \log n \rceil + B \\ &\leq B - 16 \lceil \log n \rceil \\ &\leq B - \lceil \log n \rceil - 2\lceil \log k \rceil - 12, \end{aligned} \tag{15}$$

where (a) follows from the fact that the function  $(1 - 1/x)^x$  is increasing in  $x$  for  $x > 1$  and that  $\lim_{x \rightarrow \infty} (1 - 1/x)^x = 1/e$ . Therefore,  $\phi(\mathbf{b})$  can be represented by  $B - \lceil \log n \rceil - 2\lceil \log k \rceil - 12$  bits.

For a sequence  $\mathbf{c} \in \{0, 1\}^n$ , the encoding procedure for computing  $T_1(\mathbf{c})$  is as follows.

- 1) **Initialization:** Let  $T_1(\mathbf{c}) = \mathbf{c}$ . Append  $\mathbf{1}^{2\lceil \log k \rceil + 10}$  to the end of the sequence  $T_1(\mathbf{c})$ . Let  $i = 1$  and  $n' = n$ . Go to Step 1.
- 2) **Step 1:** If  $(c_j, c_{j+1}, \dots, c_{j+\lceil \log k \rceil + 4}) \neq \mathbf{1}^{\lceil \log k \rceil + 5}$  for every  $j \in [i, i + B - \lceil \log k \rceil - 5]$ , go to Step 2. Else go to Step 4.
- 3) **Step 2:** If  $i \leq n' - B + 1$ , delete  $(T_1(\mathbf{c})_i, \dots, T_1(\mathbf{c})_{i+B-1})$  from  $T_1(\mathbf{c})$  and append  $(i, \phi(T_1(\mathbf{c})_i, \dots, T_1(\mathbf{c})_{i+B-1}), 0, \mathbf{1}^{2\lceil \log k \rceil + 10}, 0)$  to the end of  $T_1(\mathbf{c})$ . Let  $n' = n' - B$  and  $i = 1$ . Go to Step 1. Else go to Step 3.
- 4) **Step 3:** Delete  $(T_1(\mathbf{c})_i, \dots, T_1(\mathbf{c})_{n'})$  from  $T_1(\mathbf{c})$  and append  $(i, \phi(T_1(\mathbf{c})_i, \dots, T_1(\mathbf{c})_{n'}), \mathbf{0}^{i+B-n'-1}, 0, \mathbf{1}^{2\lceil \log k \rceil + 10 - (i+B-n'-1)}, 0)$  to the end of  $T_1(\mathbf{c})$ . Let  $n' = i - 1$  and  $i = 1$ . Go to Step 1.
- 5) **Step 4:** If  $i \leq n'$ , let  $i = i + 1$  and go to Step 1. Else output  $T_1(\mathbf{c})$ .

It can be seen that the length of  $T_1(\mathbf{c})$  keeps constant and is  $n + 2\lceil \log k \rceil + 10$ . The integer  $n'$  is defined such that  $(T_1(\mathbf{c})_{n'+1}, \dots, T_1(\mathbf{c})_{n+2\lceil \log k \rceil + 10})$  are the appended bits and  $(T_1(\mathbf{c})_1, \dots, T_1(\mathbf{c})_{n'})$  are the remaining bits in  $\mathbf{c}$  after the deleting operations in Step 2 and Step 3. The appended bits are not deleted in the procedure. Note that  $(T_1(\mathbf{c})_{n'+1}, \dots, T_1(\mathbf{c})_{n'+\lceil \log k \rceil + 5}) = \mathbf{1}^{\lceil \log k \rceil + 5}$ . Hence in Step 3, the integer  $i$  satisfies  $1 \leq i + B - n' - 1 \leq \lceil \log k \rceil + 4$  and the appended bits  $\mathbf{1}^{2\lceil \log k \rceil + 10 - (i+B-n'-1)}$  has length at least  $\lceil \log k \rceil + 6$ .

Since either  $i$  increases from 1 to  $n'$  or  $n'$  decreases by  $B$  in each step. The algorithm terminates within  $O(n^2)$  times of Step 1, Step 2 and Step 3. Since it takes  $O(B \log k)$  to check if a subsequence is an all 1's vector. The total complexity is  $O(n^2 k \log n \log^2 k)$ .

We now show that the output sequence  $T_1(\mathbf{c})$  satisfies Property 1. Note that for any  $i \in [1, n']$ , there exists some  $j \in [i, i + B - \lceil \log k \rceil - 5]$  such that  $T_1(\mathbf{c})_j = T_1(\mathbf{c})_{j+1} = \dots = T_1(\mathbf{c})_{j+\lceil \log k \rceil + 4} = 1$ . Otherwise  $T_1(\mathbf{c})_i$  is deleted in Step 2 or Step 3. Moreover, the distance between two consecutive patterns  $\mathbf{1}^{\lceil \log k \rceil + 5}$  after position  $n'$  is at most  $B - \lceil \log k \rceil - 5$ . Hence the sequence  $T_1(\mathbf{c})$  satisfies Property 1.

It can be seen that when the inserting operation in Step 2 or Step 3 does not occur, then we have that  $(T_1(\mathbf{c})_{n+\lceil \log k \rceil + 6}, \dots, T_1(\mathbf{c})_{n+2\lceil \log k \rceil + 10}) = \mathbf{1}^{\lceil \log k \rceil + 5}$ . If the insertion operation in Step 2 or Step 3 occurs, then we have that  $(T_1(\mathbf{c})_{n+\lceil \log k \rceil + 5}, \dots, T_1(\mathbf{c})_{n+2\lceil \log k \rceil + 9}) = \mathbf{1}^{\lceil \log k \rceil + 5}$ .

We now give the following decoding procedure that recovers  $\mathbf{c}$  from  $T_1(\mathbf{c})$ .

- 1) **Initialization:** Let  $\mathbf{c} = T_1(\mathbf{c})$  and go to Step 1.
- 2) **Step 1:** If  $c_{n+2\lceil \log k \rceil + 10} = 0$ , find the length  $\ell$  of the 1 run that ends with  $c_{n+2\lceil \log k \rceil + 9}$ . let  $i$  be the decimal representation of  $(c_{n+4\lceil \log k \rceil + 21 - B - \ell}, c_{n+4\lceil \log k \rceil + 22 - B - \ell}, \dots, c_{n+4\lceil \log k \rceil + 20 - B - \ell + \lceil \log n \rceil})$ . Let  $\mathbf{b}$  be the sequence obtained by computing  $\phi^{-1}(c_{n+4\lceil \log k \rceil + 21 - B - \ell + \lceil \log n \rceil}, c_{n+4\lceil \log k \rceil + 22 - B - \ell + \lceil \log n \rceil}, \dots, c_{n+2\lceil \log k \rceil + 8 - \ell})$ , where the function  $\phi$  is defined in the paragraph before Eq. (15) and is invertible. Note that the length of  $\mathbf{b}$  is  $B$ . Delete  $(c_{n+4\lceil \log k \rceil + 21 - B - \ell}, c_{n+4\lceil \log k \rceil + 22 - B - \ell}, \dots, c_{n+2\lceil \log k \rceil + 10})$  from  $\mathbf{c}$  and insert  $(\mathbf{b}_1, \dots, \mathbf{b}_{B-2\lceil \log k \rceil - 10 + \ell})$  at location  $i$  of  $\mathbf{c}$ . Repeat. Else delete  $c_{n+1}, \dots, c_{n+2\lceil \log k \rceil + 10}$  and output  $\mathbf{c}$ .

Note that in the encoding procedure, every appended subsequence has length  $B - 2\lceil \log k \rceil - 10 + \ell$  and ends with a 0. Hence the algorithm stops when all the subsequences appended in Step 2 or Step 3 are deleted. Moreover, the encoding procedure consists of a series of deleting and appending operations. The decoding procedure exactly reverses the series of operations in the encoding procedure. Specifically, let  $T_{1,i}(\mathbf{c})$ ,  $i \in [0, I]$  be the sequence  $T_1(\mathbf{c})$  obtained after the  $i$ -th deleting and appending operation in the encoding procedure, where  $I$  is the number of deleting and appending operations in total in

the encoding procedure. We have that  $T_{1,0}(\mathbf{c}) = \mathbf{c}$  and that  $T_{1,I}(\mathbf{c})$  is the final output  $T_1(\mathbf{c})$ . Then the decoding procedure obtains  $T_{1,I-i}(\mathbf{c})$ ,  $i \in [0, I]$  after the  $i$ -th deleting and inserting operation. Hence we get the output  $T_{1,I-I}(\mathbf{c}) = \mathbf{c}$  in the decoding procedure.  $\square$

The following lemma shows that a small sequence containing  $\mathbf{1}^{\lceil \log k \rceil + 5}$  patterns can be encoded into a sequence without the  $\mathbf{1}^{\lceil \log k \rceil + 5}$  pattern. The lemma will be used to generate sequence satisfying Property 2.

**Proposition 5.** *For an integer  $k$ , let  $\mathbf{c} \in \{0, 1\}^{3k + \lceil \log k \rceil + 4}$  be a sequence such that  $c_i = c_{i+1} = \dots = c_{i + \lceil \log k \rceil + 4} = 1$  for some  $i \in [1, 3k]$ . There exists a mapping  $T_2 : \{0, 1\}^{3k + \lceil \log k \rceil + 4} \rightarrow \{0, 1\}^{3k + \lceil \log k \rceil + 3}$ , computable in  $O(k^2 \log k)$  time, such that  $T_2(\mathbf{c})$  contains no  $\lceil \log k \rceil + 5$  consecutive 1 bits. In addition, the sequence  $\mathbf{c}$  can be recovered from  $T_2(\mathbf{c})$ .*

*Proof.* Given  $\mathbf{c} \in \{0, 1\}^{3k + \lceil \log k \rceil + 4}$ , the encoding procedure for computing  $T_2(\mathbf{c})$  is given as follows.

- 1) **Initialization:** Let  $T_2(\mathbf{c}) = \mathbf{c}$ . Append 0 to the end of the sequence  $T_2(\mathbf{c})$ . Find the smallest  $i \in [1, 3k]$  such that  $T_2(\mathbf{c})_i = T_2(\mathbf{c})_{i+1} = \dots = T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4} = 1$ . Delete  $(T_2(\mathbf{c})_i, \dots, T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4})$  from  $T_2(\mathbf{c})$  and append  $(i, \mathbf{0}^{\lceil \log k \rceil + 3 - \lceil \log(3k) \rceil})$  to the end of  $T_2(\mathbf{c})$ . Let  $n' = 3k - 1$  and  $i = 1$ . Go to Step 1.
- 2) **Step 1:** If  $i \leq n' - \lceil \log k \rceil - 4$  and  $T_2(\mathbf{c})_i = T_2(\mathbf{c})_{i+1} = \dots = T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4} = 1$ , delete  $(T_2(\mathbf{c})_i, \dots, T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4})$  from  $T_2(\mathbf{c})$  and append  $(i, \mathbf{0}^{\lceil \log k \rceil + 5 - \lceil \log(3k) \rceil}, 1)$  to the end of  $T_2(\mathbf{c})$ . Let  $n' = n' - \lceil \log k \rceil - 5$  and  $i = 1$ . Repeat. Else go to Step 2.
- 3) **Step 2:** If  $i \leq n'$ , let  $i = i + 1$  and go to Step 1. Else output  $T_2(\mathbf{c})$ .

The length of the sequence  $T_2(\mathbf{c})$  keeps constant and is

$$3k + \lceil \log k \rceil + 4 + 1 - \lceil \log k \rceil - 5 + \lceil \log k \rceil + 3 = 3k + \lceil \log k \rceil + 3.$$

The integer  $n'$  is defined such that  $(T_2(\mathbf{c})_{n'+1}, \dots, T_2(\mathbf{c})_{3k + \lceil \log k \rceil + 3})$  are appended bits and  $(T_2(\mathbf{c})_1, \dots, T_2(\mathbf{c})_{n'})$  are the remaining bits in  $T_2(\mathbf{c})$  after deleting  $(T_2(\mathbf{c})_i, \dots, T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4})$ . Note that  $T_2(\mathbf{c})_{n'+1} = 0$ . Hence if  $(T_2(\mathbf{c})_i, \dots, T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4}) = \mathbf{1}^{\lceil \log k \rceil + 5}$ , then  $i + \lceil \log k \rceil + 4 \leq n'$  and thus  $T_2(\mathbf{c})_{n'+1}$  is not deleted.

Since either  $i$  increases from 1 to  $n'$  or  $n'$  decreases in each step. The algorithm terminates within  $O(k^2)$  times of Step 1 and Step 2. Since it takes  $O(\log k)$  to check if a subsequence is an all 1's vector. The total complexity is  $O(k^2 \log k)$ .

We now show that  $T_2(\mathbf{c})$  contains no  $\mathbf{1}^{\lceil \log k \rceil + 5}$  patterns. According to the encoding procedure, for  $i \in [1, n']$ , we have that  $(T_2(\mathbf{c})_i, \dots, T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4}) \neq \mathbf{1}^{\lceil \log k \rceil + 5}$ . Furthermore, the distance between two consecutive 0 entries after position  $n'$  is at most  $\lceil \log 3k \rceil + 1$ . Hence for any  $i \in [n' + 1, 3k - 1]$ , we have that  $(T_2(\mathbf{c})_i, \dots, T_2(\mathbf{c})_{i + \lceil \log k \rceil + 4}) \neq \mathbf{1}^{\lceil \log k \rceil + 5}$ . Therefore,  $T_2(\mathbf{c})$  does not contain  $\mathbf{1}^{\lceil \log k \rceil + 5}$  patterns.

To show that  $T_2(\mathbf{c})$  is decodable, we present the following procedure that recovers  $\mathbf{c}$  from  $T_2(\mathbf{c})$ .

- 1) **Initialization:** Let  $\mathbf{c} = T_2(\mathbf{c})$  and go to Step 1.
- 2) **Step 1:** If  $c_{3k + \lceil \log k \rceil + 3} = 1$ , let  $i$  be the decimal representation of  $(c_{3k-1}, c_{3k}, \dots, c_{3k + \lceil \log 3k \rceil - 2})$ . Delete  $(c_{3k-1}, c_{3k}, \dots, c_{3k + \lceil \log k \rceil + 3})$  from  $\mathbf{c}$  and insert  $\mathbf{1}^{\lceil \log k \rceil + 5}$  at location  $i$  of  $\mathbf{c}$ . Repeat. Else go to Step 2.
- 3) **Step 2:** Let  $i$  be the decimal representation of  $(c_{3k+1}, c_{3k+2}, \dots, c_{3k + \lceil \log(3k) \rceil})$ . Delete  $(c_{3k}, c_{3k+2}, \dots, c_{3k + \lceil \log k \rceil + 3})$  from  $\mathbf{c}$  and insert  $\mathbf{1}^{\lceil \log k \rceil + 5}$  at location  $i$  of  $\mathbf{c}$ . Output  $\mathbf{c}$ .

Note that in the encoding procedure, the appended subsequence in the Initialization Step ends with a 0. The appended subsequence in Step 1 ends with a 1. Hence the algorithm stops when  $c_{3k + \lceil \log k \rceil + 3} = 0$  and the subsequence  $(c_{3k}, c_{3k+2}, \dots, c_{3k + \lceil \log k \rceil + 3})$  is deleted.

Similarly to the proof of correctness of decoding in Proposition 4, the decoding procedure exactly reverses the series of operations in the encoding procedure. Let  $T_{2,i}(\mathbf{c})$ ,  $i \in [0, I]$  be the sequence obtained after the  $i$ -th deleting and appending operation in the encoding procedure, where  $I$  is the number of deleting and appending operations in total in the encoding procedure. Then  $T_{2,i}(\mathbf{c})$  is the sequence obtained after the  $I - i$ -th deleting and inserting operation in the decoding procedure. Therefore, the decoding procedure recovers the sequence  $\mathbf{c}$  after the  $I$ -th operation.  $\square$

#### A. Proof of Lemma 4

We are now ready to present the encoding and decoding procedure for computing  $T(\mathbf{c})$ . The encoding procedure is as follows.

- 1) **Initialization:** Let  $T(\mathbf{c}) = T_1(\mathbf{c})$ . Append  $(\mathbf{0}^{3k}, \mathbf{1}^{\lceil \log k \rceil + 5})$  to the end of the sequence  $T(\mathbf{c})$ . Let  $n' = n + 2\lceil \log k \rceil + 10$  (the length of  $T_1(\mathbf{c})$ ) and  $i = 1$ . Go to Step 1.
- 2) **Step 1:** If  $i \leq \min\{n', n + 3k + 3\lceil \log k \rceil + 16 - R\}$  and for every  $j \in [i, i + R - 3k - \lceil \log k \rceil - 4]$ , there exists  $m \in [j, j + 3k - 1]$  such that  $(T(\mathbf{c})_m, T(\mathbf{c})_{m+1}, \dots, T(\mathbf{c})_{m + \lceil \log k \rceil + 4}) = \mathbf{1}^{\lceil \log k \rceil + 5}$ , split  $(T(\mathbf{c})_i, T(\mathbf{c})_{i+1}, \dots, T(\mathbf{c})_{i+R-1})$  into  $(\lceil \log n \rceil + 9 + \lceil \log k \rceil)$  blocks  $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_{\lceil \log n \rceil + 9 + \lceil \log k \rceil}$  of length  $3k + \lceil \log k \rceil + 4$ . Delete  $(\mathbf{b}_2, \dots, \mathbf{b}_{\lceil \log n \rceil + 8 + \lceil \log k \rceil})$  from  $T(\mathbf{c})$  and append  $(0, T_2(\mathbf{b}_2), T_2(\mathbf{b}_3), \dots, T_2(\mathbf{b}_{\lceil \log n \rceil + 8 + \lceil \log k \rceil}), i + 3k + \lceil \log k \rceil + 4, \mathbf{1}^{\lceil \log k \rceil + 5}, 0)$  to the end of  $T(\mathbf{c})$ . Let  $n' = n' - R + 6k + 2\lceil \log k \rceil + 8$  and  $i = 1$ . Repeat. Else go to Step 2.
- 3) **Step 2:** If  $i \leq n'$ , let  $i = i + 1$  and go to Step 1. Else output  $T(\mathbf{c})$ .

The length of  $T(\mathbf{c})$  keeps constant and is  $n + 3k + 3\lceil \log k \rceil + 15$ . The subsequence  $T(\mathbf{c})_{n'+1}, \dots, T(\mathbf{c})_{n+3k+3\lceil \log k \rceil+15}$  are appended bits. The subsequence  $T(\mathbf{c})_1, \dots, T(\mathbf{c})_{n'}$  consists of the remaining bits in  $T(\mathbf{c})$  after deletions in Step 1. Note that  $(T(\mathbf{c})_{n'+1}, \dots, T(\mathbf{c})_{n+3k+3\lceil \log k \rceil+15})$  is not deleted.

We show that  $T(\mathbf{c})_{n'+1}, \dots, T(\mathbf{c})_{n'+3k} = \mathbf{0}^{3k}$  and that either  $T(\mathbf{c})_{n'-\lceil \log k \rceil-4}, \dots, T(\mathbf{c})_{n'} = \mathbf{1}^{\lceil \log k \rceil+5}$  or  $T(\mathbf{c})_{n'-\lceil \log k \rceil-5}, \dots, T(\mathbf{c})_{n'-1} = \mathbf{1}^{\lceil \log k \rceil+5}$ . The proof is based on induction on the number  $r$  of deleting operations in Step 1 that have been done. According to Proposition 4, the claim holds after the Initialization Step and hence after  $r = 0$  deleting operations. Suppose the claim holds after  $r$  deleting operations and we have that  $(T(\mathbf{c})_{n'+1}, \dots, T(\mathbf{c})_{n'+3k}) = \mathbf{0}^{3k}$ . Then in the  $r+1$ -deleting operation in Step 1, the "if" condition only holds when  $i \leq n' - R$ . Otherwise, for  $j = n' \in [i, i + R - 1]$ , we have that  $(c_m, c_{m+1}, \dots, c_{m+\lceil \log k \rceil+4}) \neq \mathbf{1}^{\lceil \log k \rceil+5}$  for every  $m \in [j, j + 3k - 1]$ , contradicting to the "if" condition. Furthermore, note that the bits  $(T(\mathbf{c})_{i+R-3k-\lceil \log k \rceil-4}, \dots, T(\mathbf{c})_{n+3k+3\lceil \log k \rceil+15})$  after block  $\mathbf{b}_{\lceil \log n \rceil+9+\lceil \log k \rceil}$  are not deleted in the deleting operation. Hence the bits  $T(\mathbf{c})_{n'-\lceil \log k \rceil-5}, \dots, T(\mathbf{c})_{n+3k+3\lceil \log k \rceil+15}$  stay in  $T(\mathbf{c})$ . Therefore, the claim holds after  $r+1$ -th deleting operation and holds after any number of deleting operations.

We now show that  $T(\mathbf{c})$  satisfies Property 1. From Proposition 4, the initial sequence  $T(\mathbf{c}) = (T_1(\mathbf{c}), \mathbf{0}^{3k}, \mathbf{1}^{\lceil \log k \rceil+5})$  satisfies Property 1. We prove that  $T(\mathbf{c})$  keeps Property 1 after the deleting and inserting operations. Note that the deleting operation does not delete  $\mathbf{b}_1$  and  $\mathbf{b}_{\lceil \log n \rceil+9+\lceil \log k \rceil}$ , which both contain  $\mathbf{1}^{\lceil \log k \rceil+5}$  as a subsequence. Hence the deleting operation do not change the distance of consecutive  $\mathbf{1}^{\lceil \log k \rceil+5}$  patterns before block  $\mathbf{b}_1$  and after  $\mathbf{b}_{\lceil \log n \rceil+9+\lceil \log k \rceil}$ . Since the distance of consecutive  $\mathbf{1}^{\lceil \log k \rceil+5}$  patterns in  $(\mathbf{b}_1, \mathbf{b}_{\lceil \log n \rceil+9+\lceil \log k \rceil})$  is at most  $6k + 2\lceil \log k \rceil + 8 \leq B - \lceil \log k \rceil - 5$ , the sequence  $T(\mathbf{c})$  keeps Property 1 after a deleting operation. Moreover, the distance of consecutive  $\mathbf{1}^{\lceil \log k \rceil+5}$  patterns in the appended bits  $(T(\mathbf{c})_{n'+1}, \dots, T(\mathbf{c})_{n+3k+3\lceil \log k \rceil+15})$  is at most  $R - 6k - 2\lceil \log k \rceil - 8 \leq B - \lceil \log k \rceil - 5$ . Note that the appended bits are not deleted in the deleting operations. Hence  $T(\mathbf{c})$  keeps Property 1 after an inserting operation and we conclude that the output  $T(\mathbf{c})$  satisfies Property 1.

Next, we prove that  $T(\mathbf{c})$  satisfies Property 2. According to the encoding procedure, for any  $i \in [1, \min\{n', n+3k+3\lceil \log k \rceil+16-R\}]$ , there exists some  $j \in [i, i + R - 3k - \lceil \log k \rceil - 4]$ , such that  $(T(\mathbf{c})_m, T(\mathbf{c})_{m+1}, \dots, T(\mathbf{c})_{m+\lceil \log k \rceil+4}) \neq \mathbf{1}^{\lceil \log k \rceil+5}$  for every  $m \in [j, j + 3k - 1]$ . Note that the appended bits  $(T(\mathbf{c})_{n'+1}, \dots, T(\mathbf{c})_{n+3k+3\lceil \log k \rceil+15})$  are not deleted. Hence for  $i \in [n' + 1, n + 3k + 3\lceil \log k \rceil + 16 - R]$ , the interval  $[i, i + R - 1]$  contains a subinterval  $[j, j + 3k + \lceil \log k \rceil + 3]$  such that  $T(\mathbf{c})_j = 0$  and  $(T(\mathbf{c})_{j+1}, \dots, T(\mathbf{c})_{j+3k+\lceil \log k \rceil+3}) = T_2(\mathbf{b}_2)$ , where  $\mathbf{b}_2 \in \{0, 1\}^{3k+\lceil \log k \rceil+4}$  contains the  $\mathbf{1}^{\lceil \log k \rceil+5}$  pattern. Then  $(T(\mathbf{c})_m, T(\mathbf{c})_{m+1}, \dots, T(\mathbf{c})_{m+\lceil \log k \rceil+4}) \neq \mathbf{1}^{\lceil \log k \rceil+5}$  for every  $m \in [j, j+3k-1]$ . Therefore, the sequence  $T(\mathbf{c})$  satisfies Property 2. According to Proposition 3, we conclude that  $T(\mathbf{c})$  satisfies Property 1 and Property 2 and is  $k$  dense.

Since either  $i$  increases from 1 to  $n'$  or  $n'$  decreases in the encoding procedure, the procedure terminates within  $O(n^2)$  iterations. Each iteration takes  $O(k \log k R)$  time to check a violation of Property 2. Hence the complexity is at most  $O(n^2 k^2 \log k (\log n))$ . Therefore, the total complexity is  $\text{poly}(n, k)$ .

Finally we present the following decoding procedure that recovers  $\mathbf{c}$  from  $T(\mathbf{c})$ .

1) **Initialization:** Let  $\mathbf{c} = T(\mathbf{c})$  and go to Step 1.

2) **Step 1:** If  $c_{n+3k+3\lceil \log k \rceil+15} = 0$ , let  $i$  be the decimal representation of  $(c_{n+3k+2\lceil \log k \rceil+10-\lceil \log n \rceil}, c_{n+3k+2\lceil \log k \rceil+11-\lceil \log n \rceil}, \dots, c_{n+3k+2\lceil \log k \rceil+9})$ . Split  $(c_{n+9k+5\lceil \log k \rceil+25-R}, \dots, c_{n+3k+2\lceil \log k \rceil+9-\lceil \log n \rceil})$  into  $\lceil \log n \rceil + 7 + \lceil \log k \rceil$  blocks  $(\mathbf{b}'_1, \dots, \mathbf{b}'_{\lceil \log n \rceil+7+\lceil \log k \rceil})$  of length  $3k + \lceil \log k \rceil + 3$ . Compute  $\mathbf{b}_j = T_2^{-1}(\mathbf{b}'_j)$  for  $j \in [1, \lceil \log n \rceil + 7 + \lceil \log k \rceil]$ , where  $T_2^{-1}(\mathbf{b}'_j)$  is obtained by applying  $T_2$  decoder (Proposition 5) on  $\mathbf{b}'_j$ . Delete  $(c_{n+9k+5\lceil \log k \rceil+24-R}, \dots, c_{n+3k+3\lceil \log k \rceil+15})$  from  $\mathbf{c}$  and insert  $\mathbf{b}_1, \dots, \mathbf{b}_{\lceil \log n \rceil+7+\lceil \log k \rceil}$  at location  $i$  of  $\mathbf{c}$ . Repeat. Else delete  $(c_{n+2\lceil \log k \rceil+11}, \dots, c_{n+3k+3\lceil \log k \rceil+15})$  and output  $T_1^{-1}(\mathbf{c})$ .

According to the encoding procedure, the inserted bits end with a 1 entry in the Initialization Step and with a 0 entry in Step 1. Note that the inserted bits are not deleted in the encoding procedure. Hence the decoding algorithm stops when an ending 1 entry is detected.

Similarly to the to the proof of correctness of decoding in Proposition 4 and Proposition 5, the decoding procedure exactly reverses the series of operations in the encoding procedure. Therefore, the decoding procedure decodes the sequence  $\mathbf{c}$  correctly. The proof is done.

## VI. ENCODING

In this section we present the encoding function  $\mathcal{E}$  and prove Theorem 1. The function  $\mathcal{E}$  is given by

$$\mathcal{E}(\mathbf{c}) = (T(\mathbf{c}), R'(\mathbf{c}), R''(\mathbf{c})).$$

where

$$R'(\mathbf{c}) = (M(f(\mathbb{1}_{\text{sync}}(T(\mathbf{c})))) \bmod p(T(\mathbf{c})), p(T(\mathbf{c})), \text{Hash}_k(T(\mathbf{c}))), \text{ and}$$

$$R''(\mathbf{c}) = \text{Rep}_{k+1}(H(R'(\mathbf{c}))).$$

Here  $M(\mathbf{v})$  is the function defined in Eq. (3) and  $\text{Rep}_{k+1}(H(R'(\mathbf{c})))$  is the  $k+1$ -fold repetition of the bits in  $H(R'(\mathbf{c}))$  (See Lemma 5 for definition of function  $H(\mathbf{c})$ ). The function  $\text{Hash}_k(\mathbf{c})$  is defined in Eq. (14) and the function  $T(\mathbf{c})$  is defined in Lemma 4.

The length of  $R'(\mathbf{c})$  is  $N_1 = 8k \log(n + 3k + 3\lceil \log k \rceil + 15) + o(\log(n + 3k + 3\lceil \log k \rceil + 15)) = 8k \log n + o(\log n)$ . The length of  $R''(\mathbf{c})$  is  $N_2 = 2k(k + 1)(N_1/\lceil \log n \rceil) \log \log n = o(\log n)$ . Therefore, the redundancy of  $\mathcal{E}(\mathbf{c})$  is  $N_1 + N_2 = 8k \log n + o(\log n)$ . Let  $N = n + 8k \log n + o(\log n)$  be the length of  $\mathcal{E}(\mathbf{c})$ . To show that  $\mathbf{c}$  can be recovered from a length  $N - k$  subsequence  $\mathbf{d}$  of  $\mathcal{E}(\mathbf{c})$ , we prove that

- 1) The redundancy  $R'(\mathbf{c})$  can be recovered from  $k$  deletions with the help of  $R''(\mathbf{c})$ .
- 2) The sequence  $T(\mathbf{c})$  can be recovered from  $k$  deletions with the help of  $R'(\mathbf{c})$ .

Note that  $(d_{n+N_1+1}, \dots, d_{n+N_1+N_2-k})$  is a length  $N_2 - k$  subsequence of  $R''(\mathbf{c})$ . Since  $R''(\mathbf{c})$  is a  $k$  deletion code protecting  $H(R'(\mathbf{c}))$ , the hash function  $H(R'(\mathbf{c}))$  can be recovered from  $(d_{n+N_1+1}, \dots, d_{n+N_1+N_2-k})$ . Moreover,  $(d_{n+1}, \dots, d_{n+N_1-k})$  is a length  $N_1 - k$  subsequence of  $R'(\mathbf{c})$ . From Lemma 5 the hash function  $H(R'(\mathbf{c}))$  protects  $R'(\mathbf{c})$  from  $k$  deletions. Hence (1) holds.

We now prove (2). According to Lemma 2, the *synchronization vector*  $\mathbb{1}_{\text{sync}}(T(\mathbf{c}))$  can be recovered from  $M(f(\mathbb{1}_{\text{sync}}(T(\mathbf{c})))) \bmod p(T(\mathbf{c}))$  and  $p(T(\mathbf{c}))$ . Since by Lemma 4,  $T(\mathbf{c})$  is a  $k$  dense sequence. Hence from Lemma 3, it can be recovered based on  $\mathbb{1}_{\text{sync}}(T(\mathbf{c}))$  and  $\text{Hash}_k(T(\mathbf{c}))$ . Finally, the sequence  $\mathbf{c}$  can be recovered from  $T(\mathbf{c})$  by Lemma 4. Hence (2) holds and  $\mathbf{c}$  can be recovered.

The encoding complexity of  $\mathcal{E}(\mathbf{c})$  is  $O(n^{2k+1})$ , which comes from brute force searching for  $p(T(\mathbf{c}))$ . The decoding complexity is  $O(n^k + 1)$ , which comes from brute force searching for the correct  $\mathbb{1}_{\text{sync}}(T(\mathbf{c}))$ , given  $M(f(\mathbb{1}_{\text{sync}}(T(\mathbf{c})))) \bmod p(T(\mathbf{c}))$  and  $p(T(\mathbf{c}))$ .

## VII. CONCLUSION AND FUTURE WORK

We construct a  $k$ -deletion correcting code with optimal order redundancy. Interesting open problems include finding complexity  $O(N^{O(1)})$  encoding/decoding algorithms for our code, as well as constructing a systematic  $k$ -deletion code with optimal redundancy.

## REFERENCES

- [1] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.
- [2] R. R. Varshamov and G. M. Tenengolts, "Codes which correct single asymmetric errors," *Autom. Remote Control*, vol. 26, no. 2, pp. 286–290, 1965.
- [3] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, vol. 6, pp. 1–33, 2009.
- [4] A. S. Helberg and H. C. Ferreira, "On multiple insertion/deletion correcting codes," *IEEE Trans. on Inf. Th.*, vol. 48, no. 1, pp. 305–308, 2002.
- [5] K. A. Abdel-Ghaffar, F. Paluncic, H. C. Ferreira, and W. A. Clarke, "On Helberg's generalization of the Levenshtein code for multiple deletion/insertion error correction," *IEEE Trans. on Inf. Th.*, vol. 58, no. 3, pp. 1804–1808, 2012.
- [6] F. Paluncic, K. A. Abdel-Ghaffar, H. C. Ferreira, and W. A. Clarke, "A multiple insertion/deletion correcting code for run-length limited sequences," *IEEE Trans. on Inf. Th.*, vol. 58, no. 3, pp. 1809–1824, 2012.
- [7] J. Brakensiek, V. Guruswami, and S. Zbarsky, "Efficient low-redundancy codes for correcting multiple deletions," *IEEE Trans. on Inf. Th.*, vol. 64, no. 5, pp. 3403–3410, 2018.
- [8] B. Haeupler, "Optimal document exchange and new codes for small number of insertions and deletions." *arXiv:1804.03604* [cs.DS], 2018.
- [9] K. Cheng, Z. Jin, X. Li and K. Wu, "Deterministic document exchange protocols, and almost optimal binary codes for edit errors," *IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 200–211, 2018.
- [10] L. J. Schulman and D. Zuckerman, "Asymptotically good codes correcting insertions, deletions, and transpositions," *IEEE Trans. on Inf. Th.*, vol. 45, no. 7, pp. 2552–2557, 1999.
- [11] V. Guruswami and C. Wang, "Deletion codes in the high-noise and high-rate regimes," *IEEE Trans. on Inf. Th.*, vol. 63, no. 4, pp. 1961–1970, 2017.
- [12] R. Gabrys and F. Sala, "Codes correcting two deletions," *IEEE Trans. on Inf. Th.*, vol. 65, no. 2, pp. 965–974, Feb. 2019.
- [13] J. Sima, N. Raviv, and J. Bruck, "Two Deletion Correcting Codes from Indicator Vectors," *IEEE Int. Symp. on Inform. Theory.*, Vail, USA, pp. 421–425, 2018.
- [14] S. K. Hanna and S. El Rouayheb, "Guess & check codes for deletions, insertions, and synchronization," *IEEE Trans. on Inf. Th.*, vol. 65, no. 1, pp. 3–15, Jan. 2019.
- [15] J. L. Nicolas, "On highly composite numbers," in *Ramanujan revisited*, Urbana-Champaign, Ill., pp. 215–244, 1987.
- [16] S. Gao, "A new algorithm for decoding Reed-Solomon codes," *Communications, Information and Network Security*, Springer, Boston, MA, pp. 55–68, 2003.