On the Varshamov-Tenengolts Construction on Binary Strings

Ankur A. Kulkarni[*] Negar Kiyavash R. Sreenivas

Abstract

This paper is motivated by the problem of finding the largest single-deletion-correcting code for binary strings. The Varshamov-Tenengolts construction classifies binary strings into non-overlapping sets, the largest set of these is asymptotically the largest single-deletion-correcting code. However despite the asymptotic optimality little is known about the quality of the construction as a function of the string length. We show that these sets are also responsible for the (near) solution of several combinatorial problems on a certain hypergraph. Furthermore our results are valid for any string length. We show that the sets collectively solve strong vertex coloring and edge coloring on the hypergraph exactly. For any string length n we show that the largest of these sets is within $\frac{n+1}{n-1}$ of optimal matching on the hypergraph, which also corresponds to the largest single-deletion-correcting code. Moreover, we show for any n the smallest of these sets is within $\frac{n^2-n}{n^2-3n+4-\frac{4}{2^n}}$ of the smallest cover of this hypergraph and that each of these sets is a perfect matching. We then obtain similar results on the dual of this hypergraph.

1 Introduction and background

The standard model for a communication system comprises of a string of symbols that is sent over an imperfect medium, called a channel, which reproduces the input symbols with errors. This note is about a particular kind of error, called the *deletion error*. In a deletion error, the output of the channel comprises of a *substring* of the input, where some of the input symbols are deleted and the undeleted symbols are aligned while maintaining their original order (the position of the missing symbols is not known at the output). Several problems pertaining to this channel remain unsolved. One such problem is ascertaining the size of, and providing a construction for, the largest *code*, i.e. the largest set of input strings so that no two of them have a common substring.

Although this problem on the deletion channel is hard in general, the case of this channel where only binary symbols are sent and only a single symbol is deleted appears to hold promise for a complete solution. A code construction devised by Varshamov and Tenengolts is known to be *asymptotically* the largest [10]. Moreover, it has been conjectured to be the largest for any input length, and verified to be largest for input lengths up to 10 [18].

In this note we are motivated by the observation that the problem of the largest code is not the only problem this construction appears to be solving. We show that the same construction is responsible for the solution for several other combinatorial problems on binary strings. In a sense, this paper may be thought of as an attempt to characterize the *inverse optimality* – finding problems for which a candidate solution is optimal – of the Varshamov-Tenengolts construction.

Let $\mathbb{F}_2^n$ denote the set of all binary sequences or *strings* of length $n \in \mathbb{N}$, i.e.,

$$\mathbb{F}_2^n = \{x | x = x_1 x_2 \dots x_n, \text{ such that } x_i \in \{0,1\}, i = 1, \dots, n\}.$$

Each x_i is called a *bit*. Let $v : \mathbb{F}_2^n \to \{0, 1, \dots, n\}$ be the function

$$v(x) = \sum_{i=1}^{n} i x_i \mod n+1, \qquad \forall x \in \mathbb{F}_2^n, \tag{1}$$

[*] Ankur is with the Systems and Control Engineering group at the Indian Institute of Technology Bombay, Mumbai, India, 400076. The other authors are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Illinois, 61801, U.S.A. This work was done while Ankur was at this laboratory. Kiyavash and Sreenivas are also at the Department of Industrial and Enterprise Systems Engineering at the same university. The authors can be contacted at `kulkarni.ankur@iitb.ac.in`, `kiyavash@illinois.edu` and `rsree@illinois.edu`, respectively. The work of Kulkarni and Kiyavash was supported in part by AFOSR under Grants FA9550-11-1-0016, FA9550-10-1-0573, and NSF grants CCF 10-54937 CAR and CCF 10-65022 Kiyavash.

where the sum is interpreted as a decimal sum. This function was introduced by Varshamov and Tenengolts [21]. It classifies binary strings in $\mathbb{F}_2^n$ into $n+1$ non-overlapping sets, denoted $\mathrm{VT}_a(n)$ for $a = 0, \ldots, n$, where $\mathrm{VT}_a(n)$ comprises of strings x for which the "VT weight" $v(x) = a$:

$$\mathrm{VT}_a(n) = \{x \in \mathbb{F}_2^n \mid v(x) = a\}.$$

We refer to these sets collectively as the *Varshamov-Tenengolts construction.*

If $x = x_1 x_2 \ldots x_n$ is a string in $\mathbb{F}_2^n$, a string $y = x_{i_1} x_{i_2} \ldots x_{i_{n-1}} \in \mathbb{F}_2^{n-1}$ is called a *subsequence* of x if the indices satisfy $1 \leq i_1 < \ldots < i_{n-1} \leq n$. Thus y is obtained from x by the deletion of a single bit. We call x a *supersequence* of y, and x is obtained from y by the insertion of a single bit. Let $D(x) \subseteq \mathbb{F}_2^{n-1}$ be the set of all subsequences of x (called the *deletion set* of x) and let $I(x) \subseteq \mathbb{F}_2^{n+1}$ denote the set of supersequences of x (called the *insertion set* of x), obtained from x by a single deletion and insertion, respectively. The following concepts are central to this paper.

Definition 1.1 *A single-deletion (respectively, -insertion) correcting code for string length n is a set $C \subseteq \mathbb{F}_2^n$ such that $D(x) \cap D(y) = \emptyset$ (respectively, $I(x) \cap I(y) = \emptyset$) for all $x, y \in C$. An optimal single-deletion (respectively, -insertion) correcting code is a single-deletion correcting code of largest cardinality.*

Definition 1.2 *A deletion (respectively, insertion) cover for string length n is a set of strings $C \subseteq \mathbb{F}_2^n$ of such that their deletion (respectively, insertion) sets cover $\mathbb{F}_2^{n-1}$ (respectively, $\mathbb{F}_2^{n+1}$), i.e., if $\cup_{x \in C} D(x) = \mathbb{F}_2^{n-1}$ (respectively, $\cup_{x \in C} I(x) = \mathbb{F}_2^{n+1}$). An optimal deletion (respectively, insertion) cover is a deletion (respectively, insertion) cover of smallest cardinality. A perfect deletion (respectively, insertion) matching is a set of binary strings such that their deletion (respectively, insertion) sets cover $\mathbb{F}_2^{n-1}$ (respectively, $\mathbb{F}_2^{n+1}$) and are pairwise disjoint.*

We consider the hypergraphs

$$\mathcal{H}_n^{\mathsf{D}} = (\mathbb{F}_2^{n-1}, \{D(x) | x \in \mathbb{F}_2^n\}) \qquad \text{and} \qquad \mathcal{H}_n^{\mathsf{I}} = (\mathbb{F}_2^{n+1}, \{I(x) | x \in \mathbb{F}_2^n\}). \tag{2}$$

We call $\mathcal{H}_n^{\mathsf{D}}$ the *deletion* hypergraph and $\mathcal{H}_n^{\mathsf{I}}$ the *insertion* hypergraph. And we consider the relation of the Varshamov-Tenengolts construction to the following problems on $\mathcal{H}_n^{\mathsf{D}}$.

D1. Matching: Maximum number of strings in $\mathbb{F}_2^n$ with disjoint deletion sets.

D2. Covering: Minimum number of strings in $\mathbb{F}_2^n$ such that their deletion sets cover $\mathbb{F}_2^{n-1}$.

D3. Perfect Matching: Does there exist a perfect deletion matching?

D4. Strong Vertex Coloring: Minimum number of colors required to color strings in $\mathbb{F}_2^{n-1}$ such that no deletion set of a string in $\mathbb{F}_2^n$ contains strings of the same color.

D5. Edge Coloring: Minimum number of colors required to color strings in $\mathbb{F}_2^n$ such that no two strings with intersecting deletion sets have the same color.

The aim of this note is to show that the Varshamov-Tenengolts (VT) construction solves *all* of these problems, at least to a good approximation.

It was first noticed by Levenshtein [10] that for any a, the set $\mathrm{VT}_a(n)$ forms a single-deletion correcting code. Furthermore, more recently Levenshtein [11] showed that for every a, the set $\mathrm{VT}_a(n)$ is a cover of $\mathbb{F}_2^{n-1}$, making $\mathrm{VT}_a(n)$ a perfect deletion matching for any a. Thus this construction solves **D3**. We show that the sets $\mathrm{VT}_0(n), \ldots, \mathrm{VT}_n(n)$ collectively are an optimal edge coloring of $\mathcal{H}_n^{\mathsf{D}}$ (problem **D5**) and the sets $\mathrm{VT}_0(n-1), \ldots, \mathrm{VT}_{n-1}(n-1)$ collectively are an optimal strong vertex coloring of $\mathcal{H}_n^{\mathsf{D}}$ (problem **D4**). Although it is known that the set $\mathrm{VT}_a(n)$ is a deletion cover for any a, the role of the Varshamov-Tenengolts construction in relation to the *minimum covering*, i.e., problem **D2** is unknown. We show that $\mathrm{VT}_1(n)$ (the smallest of these sets) is within a factor of $\left(\frac{n^2-n}{n^2-3n+4-\frac{4}{2^n}}\right)$ of the solution of **D2**. Our numerical results confirm that it is indeed the smallest deletion cover for $n \leq 8$. Finally, we show $\mathrm{VT}_0(n)$ (the largest of these sets) is within a factor of $\left(\frac{(n+1)(1-\frac{2}{2^n})}{n-1}\right)$ of the largest single-deletion correcting code (problem **D1**).

In summary, these results show that the Varshamov-Tenengolts (VT) construction is an edge coloring of $\mathcal{H}_n^{\mathsf{D}}$ with the remarkable property that each of the color classes is a perfect matching, the largest color class is (nearly) the largest matching, the smallest color class is (nearly) the smallest cover and collectively, the color classes are an optimal coloring. Indeed, when $n+1$ is a power of 2, $|\mathrm{VT}_1(n)| = |\mathrm{VT}_0(n)|$, so in this case there are $n+1$ sets of hyperedges each of which appears to be both the largest matching and the smallest cover of $\mathcal{H}_n^{\mathsf{D}}$. Our numerical results confirm this for $n = 1, 3, 7$.

Some results of this flavor, where VT construction is related to other problems, have been obtained by Sloane [17]. Sloane shows that the size of $\mathrm{VT}_0(n)$ is the same as the number of certain shift register sequences and the size of $\mathrm{VT}_1(n)$ is the same as the number of certain necklaces, but in both cases, no bijections are known. The relation of the VT construction to the problems we study above adds to the list of combinatorial properties that these codes have. These facts put together point to the existence of a deeper property that this construction enjoys which may not be obvious if these problems are seen in isolation. Identifying this property remains open.

Let **I1**, . . . , **I5** be analogous problems where insertion sets are considered instead of deletion sets. For example,

I1. Maximum number of strings in $\mathbb{F}_2^n$ with disjoint *insertion* sets.

It was shown in [10] a set of strings is a single-deletion correcting code if and only if it is a single-insertion correcting code. Consequently, problems **D1** and **I1** are equivalent and our bound for **D1** applies to **I1** too. Furthermore, we show that the hypergraphs $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_{n-1}^{\mathsf{I}}$ are duals of each other. This implies that solutions of problems **I4** and **I5** (strong vertex coloring and edge coloring of $\mathcal{H}_n^{\mathsf{I}}$) are also provided by the Varshamov-Tenengolts construction. Although the structure of an insertion cover is not known, we show that the size of the smallest such set is at most $(1+\log(n+1))\frac{2^n-2}{n-1}$. This provides a bound on **I2**. It turns out that $\mathcal{H}_{n-1}^{\mathsf{I}}$ is a $(n+1)$-uniform hypergraph. From this it is easy to show that there do not, in general, exist perfect insertion covers (i.e. a set of strings in $\mathbb{F}_2^n$ with disjoint insertion sets that cover $\mathbb{F}_2^{n+1}$), which answers **I3** in the negative.

For a long time no non-asymptotic bounds or approximation factors for these problems were known. Levensthein's original paper [10] showed that $\mathrm{VT}_0(n)$ is asymptotically optimal, i.e., if $\mathcal{C}_n^*$ is the largest code, Levensthein showed that $\frac{|\mathrm{VT}_0(n)|}{|\mathcal{C}_n^*|} \to 1$ as $n \to \infty$. But since $|\mathcal{C}_n^*|$ is exponential in n, this result says little about the quality of $\mathrm{VT}_0(n)$ for any particular n. Levenshtein later found the following non-asymptotic bound [13] by the same argument from [10], but replacing certain asymptotic formulae with exact bounds.

$$|\mathcal{C}_n^*| \leq \frac{2^{n-1}}{r+1} + 2\sum_{i=0}^{r-1}\binom{n-1}{i}, \tag{3}$$

where r is any integer satisfying $1 \leq r+1 \leq n$. This bound weaker than our bound besides being more complicated. Finding upper bounds for deletion-correcting codes is notoriously hard, as emphasized by Sloane [17]: *"It is more difficult to obtain upper bounds for deletion-correcting codes than for conventional error-correcting codes, since the disjoint balls $D_e(u)$ (deletion sets) associated with the codewords ... do not all have the same size. Furthermore ... there is no obvious linear programming bound."* One would presume that the same difficulty also makes finding lower bounds on the minimum covering hard.

Our approximation factors are obtained by exploiting the linear programming relaxations of these problems. Since deletion correction and insertion correction are equivalent, both insertion and deletion hypergraphs, are relevant to this problem. On each of hypergraphs, the solution of the matching problem provides a single-deletion correcting code. The fractional matching on the insertion hypergraph provides a weak bound for the matching problem. But thanks to the duality between the insertion and deletion hypergraphs, this provides a good lower bound for the size of the smallest cover. This leads to the approximation factor for $\mathrm{VT}_1(n)$ for problem **D2**. The fractional matching on the deletion hypergraph provides the approximation factor for $\mathrm{VT}_0(n)$ to solve the matching problem **D1**. Classical asymptotic results of Levenshtein [10, 11] on the largest code and smallest deletion cover follow as corollaries of these results.

The more conventional approach [17, 4] to these problems appears to be through the use of the graph L_n, and as such the hypergraph approach seems new. We devote a portion of this paper to prove supplementary properties of the hypergraphs involved. In Section 2 we introduce some hypergraph concepts and the hypergraph formulation of the problem. Our bounds and approximation factors are derived in Section 3. We conclude in Section 4.

2 Preliminaries and background

We begin with a brief survey of the literature of the field.

2.1 A brief survey

The VT construction was originally proposed for correcting asymmetric errors [21]. The earliest work on the binary single-deletion channel appears to be that of Levenshtein [10], who observed that each set in the VT construction serves as a code for correcting a single deletion and established its asymptotic optimality of the largest set in this construction. Indeed, Levenshtein also considered a more general version of this construction to address substitution errors together with deletion and insertion. This construction was conceptually generalized for larger number of asymmetric errors later by Varshamov [20], and a variant of it for correcting a single deletion for arbitrary alphabet size was proposed by Tenengolts [19]. The perfectness property of the VT construction was noticed and studied by Levenshtein [11]. Ginsburg [7] obtained a closed form expression for the sizes of the sets $\mathrm{VT}_a(n)$ and showed that for any n, the set of smallest size was $\mathrm{VT}_1(n)$ and the one of the largest size was $\mathrm{VT}_0(n)$ (see also the survey by Sloane [17]). The binary single-deletion channel has many profound combinatorial questions associated with it, most of which remain unanswered. The survey by Sloane [17] discusses these problems in depth and draws connections of the VT construction with other combinatorial problems. Sloane also maintains a website [18] for archiving results and numerical computations pertaining to this problem.

For a complete survey of multiple deletion problems, we refer the reader to Mercier et al. [14]. We note that for multiple deletions a generalization of the VT construction was made by Helberg and Ferreira [8, 2] but the resulting code has smaller size than the known lower bound. An asymptotically optimal code construction that also corrects transpositions and where the number of errors grows linearly with the length of the string was presented by Schulman and Zuckerman [16].

2.2 Hypergraphs

Below, we briefly recall some hypergraph concepts, sourced mainly from Berge [3].

Definition 2.1 *A hypergraph $\mathcal{H}$ is a tuple $(X, \mathcal{E})$, where X is a finite set and $\mathcal{E}$ is a collection of subsets of X. X is called the vertex set, its elements are called vertices and the elements of $\mathcal{E}$ are called hyperedges. When a vertex belongs to a hyperedge, we say it is covered by the hyperedge.*

We consider the following extremal concepts on hypergraph $\mathcal{H} = (X, \mathcal{E})$. A *covering* is a collection of hyperedges that cover the vertex set. The covering number $\kappa(\mathcal{H})$ is the smallest size of a covering of $\mathcal{H}$. A *packing* is a collection of vertices such that no two vertices are covered by the same hyperedge. The size of the largest packing is the packing number $p(\mathcal{H})$. A *matching* is a collection of hyperedges no two of which intersect. The matching number $\nu(\mathcal{H})$ is the largest size of a matching of $\mathcal{H}$. A matching is said to be *perfect* if the edges in the matching cover the vertex set. A *transversal* is a set of vertices that intersects every hyperedge. The transversal number $\tau(\mathcal{H})$ is the smallest size of a transversal. A *strong vertex coloring* of $\mathcal{H}$ is a partition of its vertex set into k classes $X_1, X_2, \ldots, X_k$ such that no color appears twice in the same hyperedge. i.e., $|E \cap X_i| \leq 1$ for $i = 1, \ldots, k$ for any hyperedge $E \in \mathcal{E}$. The *strong chromatic number* $\gamma(\mathcal{H})$ is the smallest integer k for which a strong coloring exists. An *edge coloring* of $\mathcal{H}$ is a coloring of the hyperedges of $\mathcal{H}$ such that any two intersecting hyperedges are colored differently. The *chromatic index* $q(H)$ is the least number of colors required for an edge coloring.

We now define hypergraph duality to relate these concepts.

Definition 2.2 *Let $\mathcal{H} = (X, \mathcal{E})$ be a hypergraph where $X = \{x_1, \ldots, x_{|X|}\}$ and $\mathcal{E} = \{E_1, \ldots, E_{|\mathcal{E}|}\}$. The dual of $\mathcal{H}$, denoted $\mathcal{H}^*$, is a hypergraph whose vertices $e_1, \ldots, e_{|\mathcal{E}|}$, $e_i = E_i$, are the hyperedges of $\mathcal{H}$ and whose hyperedges are the sets $Y_i = \{e_j | x_i \in E_j \text{ in } \mathcal{H}\}, i = 1, \ldots, |X|$.*

Between hypergraph $\mathcal{H}$ and its dual $\mathcal{H}^*$, the following relationships hold:

$$\nu(\mathcal{H}^*) = p(\mathcal{H}), \qquad \tau(\mathcal{H}^*) = \kappa(\mathcal{H}), \qquad \text{and} \qquad q(\mathcal{H}^*) = \gamma(\mathcal{H}). \tag{4}$$

Suppose $X = \{x_1, \ldots, x_{|X|}\}$ and $\mathcal{E} = \{E_1, \ldots, E_{|\mathcal{E}|}\}$. Let the incidence matrix of $\mathcal{H}$ be a matrix $A \in \{0,1\}^{|X|\times|\mathcal{E}|}$, where the element in the k^{th} row and ℓ^{th} column, $A[k,\ell]$, is defined as

$$A[k,\ell] = \begin{cases} 1 & \text{if } x_k \in E_\ell, \\ 0 & \text{otherwise.} \end{cases}$$

The numbers $\kappa(\mathcal{H})$, $p(\mathcal{H})$, $\nu(\mathcal{H})$ and $\tau(\mathcal{H})$ are solutions of integer linear programs. By mathematical programming duality, we have $p(\mathcal{H}) \leq \kappa(\mathcal{H})$ and $\nu(\mathcal{H}) \leq \tau(\mathcal{H})$. These inequalities are, in general, not tight and sandwiched between them are the linear programming relaxations of the problem. By $\kappa^*(\mathcal{H}), p^*(\mathcal{H}), \nu^*(\mathcal{H}), \tau^*(\mathcal{H})$ denote the values of the these relaxations. We thus have the following proposition.

Proposition 2.1 *For any hypergraph $\mathcal{H}$, we have*

$$p(\mathcal{H}) \leq p^*(\mathcal{H}) = \kappa^*(\mathcal{H}) \leq \kappa(\mathcal{H}), \tag{5}$$

$$\nu(\mathcal{H}) \leq \nu^*(\mathcal{H}) = \tau^*(\mathcal{H}) \leq \tau(\mathcal{H}). \tag{6}$$

The equalities in (5), (6) are due to strong duality in linear programming.

Following are two concepts of derived graphs from the hypergraph $\mathcal{H}$ that we will employ. The *line graph* of $\mathcal{H}$, $L(\mathcal{H})$ is a graph whose vertices are the hyperedges of $\mathcal{H}$. Two vertices in $L(\mathcal{H})$ are adjacent if they intersect as hyperedges in $\mathcal{H}$. The *generated graph* of $\mathcal{H}$, $\Gamma(\mathcal{H})$ is a graph whose vertices are the vertices of $\mathcal{H}$. Two vertices in $\Gamma(\mathcal{H})$ are adjacent if they are covered by a hyperedge in $\mathcal{H}$. An edge coloring of $\mathcal{H}$ is equivalent to a vertex coloring of $L(\mathcal{H})$. A strong vertex coloring of $\mathcal{H}$ is equivalent to a vertex coloring of $\Gamma(\mathcal{H})$.

2.3 Hypergraph formulations for codes, covers and colorings

Let $\mathbb{F}_2^* = \bigcup_{n=0}^{\infty} \mathbb{F}_2^n$ be the set of all binary strings including the empty string. For any $x, y \in \mathbb{F}_2^*$ define $d(x,y)$ to be minimum number of insertions or deletions required to obtain x from y. d is known as the Levenshtein distance or edit distance. For any $x, y \in \mathbb{F}_2^*$, let $l(x), l(y)$ be the lengths of x and y, let $\bar{l}(x,y)$ be the minimum length of a string z such that both x and y can be obtained from z by the deletion of bits, and let $\underline{l}(x,y)$ be the maximum length of a string z such that z can be obtained from both x, y by the deletion of bits. For any two strings x, y, the empty string can be obtained from both x, y by the deletion of bits, and from the concatenation of x and y both x, y can be obtained by deletion of bits, whereby the functions $\bar{l}(\cdot,\cdot)$ and $\underline{l}(\cdot,\cdot)$ are well-defined and finite on $\mathbb{F}_2^* \times \mathbb{F}_2^*$. The Levenshtein distance d between x and y has the following characterization [11]:

$$d(x,y) = l(x) + l(y) - 2\underline{l}(x,y) = 2\bar{l}(x,y) - l(x) - l(y) = \bar{l}(x,y) - \underline{l}(x,y). \tag{7}$$

This leads us to a fundamental equivalence relationship between deletion-correction and insertion-correction.

Lemma 2.1 *Let $n \in \mathbb{N}$. For any $x, y \in \mathbb{F}_2^n$, the following three statements are equivalent.*

1. *$d(x,y) \leq 2$,*
2. *There exists $z \in \mathbb{F}_2^{n-1}$ such that $z \in D(x) \cap D(y)$,*
3. *There exists $z \in \mathbb{F}_2^{n+1}$ such that $z \in I(x) \cap I(y)$.*

Proof : By definition, $\underline{l}(x,y) \geq n-1$ if and only if there exists a string z of length $n-1$, such that $z \in D(x) \cap D(y)$. Similarly, $\bar{l}(x,y) \leq n+1$ if and only if there exists a $z \in \mathbb{F}_2^{n+1}$ such that $z \in I(x) \cap I(y)$. Now using $l(x) = l(y) = n$ in (7), the result follows. ■

Note that this equivalence is valid only if strings x, y have the same length. Based on the Levenshtein distance, one can define the following graph. This graph is also employed in [5, 6].

Definition 2.3 *Define the indistinguishability graph L_n as the graph with vertices $\mathbb{F}_2^n$ where two vertices $x, y \in \mathbb{F}_2^n$ are adjacent if and only if $d(x,y) \leq 2$.*

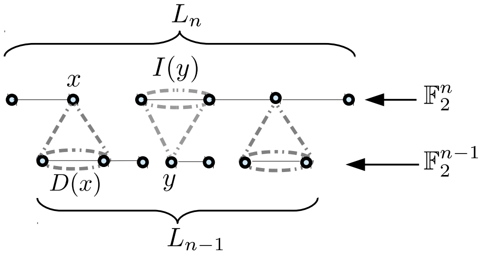

Figure 1: The figure pictorially depicts the insertion hypergraph $\mathcal{H}^{\mathsf{I}}_{n-1}$ and deletion hypergraph $\mathcal{H}^{\mathsf{D}}_n$. Bold-faced circles are strings – the top layer depicts $\mathbb{F}_2^n$ and the bottom layer depicts $\mathbb{F}_2^{n-1}$. A string $x \in \mathbb{F}_2^n$ is shown and the cone drawn below it with dotted lines is its deletion set $D(x) \subseteq \mathbb{F}_2^{n-1}$. This set is a hyperedge in $\mathcal{H}^{\mathsf{D}}_n$. Also shown is a string $y \in \mathbb{F}_2^{n-1}$ and the inverted cone drawn with dotted lines is its insertion set $I(y) \subseteq \mathbb{F}_2^n$. This is a hyperedge in $\mathcal{H}^{\mathsf{I}}_{n-1}$. The derived graphs L_n and L_{n-1} have their vertices as the sets $\mathbb{F}_2^n$ and $\mathbb{F}_2^{n-1}$, respectively. Also shown are some edges in L_n and L_{n-1}, joining strings in $\mathbb{F}_2^n$ and $\mathbb{F}_2^{n-1}$, respectively.

Recall the hypergraphs $\mathcal{H}^{\mathsf{D}}_n, \mathcal{H}^{\mathsf{I}}_n$ defined in (2). Fig 1 pictorially depicts these hypergraphs. By using Lemma 2.1, and by the definitions of the line graph and the generated graph of a hypergraph, it follows that

$$\Gamma(\mathcal{H}^{\mathsf{I}}_{n-1}) = L(\mathcal{H}^{\mathsf{D}}_n) = L(\mathcal{H}^{\mathsf{I}}_n) = \Gamma(\mathcal{H}^{\mathsf{D}}_{n+1}) = L_n. \tag{8}$$

The derived graphs of hypergraphs $\mathcal{H}^{\mathsf{D}}_n$ and $\mathcal{H}^{\mathsf{I}}_{n-1}$ are depicted in Fig 1; notice there that L_n is both the line graph of $\mathcal{H}^{\mathsf{D}}_n$ the generated graph of $\mathcal{H}^{\mathsf{I}}_{n-1}$. For a clearer appreciation of (8), note that $\mathcal{H}^{\mathsf{D}}_n$ is the dual of $\mathcal{H}^{\mathsf{I}}_{n-1}$; we show this in the next section in Lemma 3.2. This explains the first and third equality, whereas the second and fourth equality follows from Lemma 2.1.

It follows from definition that for string length n, a deletion (insertion) cover is equivalent to a covering of $\mathcal{H}^{\mathsf{D}}_n$ (of $\mathcal{H}^{\mathsf{I}}_n$) and a deletion- (insertion-) correcting code is equivalent to a matching of $\mathcal{H}^{\mathsf{D}}_n$ (of $\mathcal{H}^{\mathsf{I}}_n$). However, thanks to Lemma 2.1, we have a stronger statement, namely that these matchings are equivalent.

Lemma 2.2 *Let $n \in \mathbb{N}$ and let $C \subseteq \mathbb{F}_2^n$. Then C is a single-deletion correcting code if and only if it is a single-insertion correcting code. Consequently,*

$$\nu(\mathcal{H}^{\mathsf{D}}_n) = \nu(\mathcal{H}^{\mathsf{I}}_n) = \alpha(L_n), \tag{9}$$

where $\alpha(L_n)$ is the size of the maximum independent set in L_n.

Proof : By Definition 1.1 and Lemma 2.1, we get that a set $C \subseteq \mathbb{F}_2^n$ is a single-deletion correcting code if and only if it is a single-insertion correcting code. Furthermore, since L_n is the line graph of $\mathcal{H}^{\mathsf{D}}_n$, an independent set of L_n is equivalent to a matching of $\mathcal{H}^{\mathsf{D}}_n$. The equalities in (9) follow. ■

Likewise, by Lemma 2.1, we have that for any n, an edge coloring of $\mathcal{H}^{\mathsf{D}}_n$ is equivalent to an edge coloring of $\mathcal{H}^{\mathsf{I}}_n$. Furthermore, by (8), these are both equivalent to a strong vertex coloring of the hypergraphs $\mathcal{H}^{\mathsf{D}}_{n+1}$, $\mathcal{H}^{\mathsf{I}}_{n-1}$, and a vertex coloring of the graph L_n. Consequently, we have the relations

$$\gamma(\mathcal{H}^{\mathsf{D}}_n) = \gamma(\mathcal{H}^{\mathsf{I}}_n) = q(\mathcal{H}^{\mathsf{D}}_{n+1}) = q(\mathcal{H}^{\mathsf{I}}_{n-1}) = \chi(L_n). \tag{10}$$

This relation can better appreciated by taking note of the duality $\mathcal{H}^{\mathsf{I}}_{n-1} = (\mathcal{H}^{\mathsf{D}}_n)^*$ which we show later in Lemma 3.2. With this we conclude our preliminaries. In the following section we present our main results.

3 Main new results

3.1 Structure of insertion and deletion hypergraphs

We begin by making some observations about the structure of the insertion and deletion hypergraphs. We first note the sizes of insertion and deletion sets. The size of $I(x)$ is a constant independent of x, but is a function only of the length of x [15]. Specifically,

$$|I(x)| = n + 1 := I_n, \qquad \forall\ x \in \mathbb{F}_2^{n-1}. \tag{11}$$

A *run* of a string is a maximal contiguous subsequence of identical symbols. For example, $1, 0, 1, 000, 1$ are each a run of the string 1010001. The size of the deletion set of x, $D(x)$, is equal to the number of runs of x, denoted by $r(x)$ [11]:

$$|D(x)| = r(x), \qquad \forall\ x \in \mathbb{F}_2^*. \tag{12}$$

Next, we show that hypergraphs $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_{n-1}^{\mathsf{I}}$ are duals of each other. To show this, we need the following lemma.

Lemma 3.1 *Distinct strings of length $n \geq 2$ have distinct deletion sets and distinct insertion sets.*

Proof : Let $n \in \mathbb{N}, n \geq 2$. Theorem 2 and Eq 38, and Theorem 3 and Eq 63 in [12] give bounds on the maximum number of common subsequences and supersequences for any pair of strings as follows

$$\max_{x,y\in\mathbb{F}_2^n, x\neq y} |D(x) \cap D(y)| = 2, \qquad \max_{x,y\in\mathbb{F}_2^n, x\neq y} |I(x) \cap I(y)| = 2. \tag{13}$$

Since $|I(x)| = n+2\ \forall\ x \in \mathbb{F}_2^n$, (cf. (11)) it follows that there cannot exist distinct $x, y \in \mathbb{F}_2^n$, such that $I(x) = I(y)$.

Suppose there exist distinct $x, y \in \mathbb{F}_2^n$ such that $D(x) = D(y)$. Then it follows from (13) that $|D(x)| = |D(y)| \leq 2$. From (12), it follows that x and y must have the same number of runs and at most 2 runs each. This means there are two possibilities: a) x, y belong to the set $\{10\ldots0, 0\ldots01, 1\ldots10, 01\ldots1\}$ (of strings with two runs) or b) x, y belong to the set $\{0\ldots0, 1\ldots1\}$ (of strings with one run). It can be easily checked that no two strings in either set have identical deletion sets. ■

The (set-valued) maps $D(\cdot)$ and $I(\cdot)$ are inverses of each other, i.e., $x \in D(y) \iff y \in I(x)$ for all $x, y \in \mathbb{F}_2^*$. This observation and Lemma 3.1 allows for a bijection between strings and their deletion sets and strings and their insertion sets (cf. Fig 1). Consequently, without loss of generality, we identify hyperedges in $\mathcal{H}_n^{\mathsf{D}}$ (respectively, $\mathcal{H}_n^{\mathsf{I}}$) with the strings in $\mathbb{F}_2^n$ of which they are the deletion (respectively, insertion) sets. In the rest of the paper the equality between strings and their deletion (or insertion) sets will be understood as equality upon appropriate application of this bijection.

Lemma 3.2 *Let $n \in \mathbb{N}, n \geq 2$ and consider the hypergraphs $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_n^{\mathsf{I}}$ defined in* (2)*. We have $\mathcal{H}_{n-1}^{\mathsf{I}} = (\mathcal{H}_n^{\mathsf{D}})^*$, i.e., the hypergraph $\mathcal{H}_{n-1}^{\mathsf{I}}$ is isomorphic to the dual of $\mathcal{H}_n^{\mathsf{D}}$.*

Proof : The vertices of $(\mathcal{H}_n^{\mathsf{D}})^*$ are the sets $D(y)$, $y \in \mathbb{F}_2^n$. By Lemma 3.1, there is a bijection ϕ between the vertices of $(\mathcal{H}_n^{\mathsf{D}})^*$ and $\mathbb{F}_2^n$, the vertices of $\mathcal{H}_{n-1}^{\mathsf{I}}$. The hyperedges of $(\mathcal{H}_n^{\mathsf{D}})^*$ are the sets $\{D(y)|x \in D(y)\}, x \in \mathbb{F}_2^{n-1}$. Since $D(\cdot)$ is the inverse of $I(\cdot)$ we get that for any $x \in \mathbb{F}_2^{n-1}$, the image under ϕ of a hyperedge of $(\mathcal{H}_n^{\mathsf{D}})^*$ is $\{\phi(D(y))|x \in D(y)\} = \{\phi(D(y))|y \in I(x)\} = \{y|y \in I(x)\} = I(x)$, a hyperedge of $\mathcal{H}_{n-1}^{\mathsf{I}}$. Conversely, for any $x \in \mathbb{F}_2^{n-1}$ the image under ϕ^{-1} of a hyperedge of $\mathcal{H}_{n-1}^{\mathsf{I}}$ is $\{\phi^{-1}(y)|y \in I(x)\} = \{D(y)|x \in D(y)\}$, which is a hyperedge of $(\mathcal{H}_n^{\mathsf{D}})^*$. ■

The insertion and deletion hypergraphs we have defined fall into two well-known, but broad categories of hypergraphs.

Definition 3.1 *A hypergraph $\mathcal{H} = (X, \mathcal{E})$ is said to be k-uniform if each of its hyperedges $E_i \in \mathcal{E}$ is of size k. $\mathcal{H}$ is said to be k-partite if its vertex set X is the disjoint union of k sets $X_1, \ldots, X_k$ and its hyperedges E satisfy $|E \cap X_i| = 1$ for $i = 1, \ldots, k$. A hypergraph $\mathcal{H}$ is said to be k-regular if its dual $\mathcal{H}^*$ is k-uniform.*

Since the size of the insertion set (cf. (11)) is the same for all strings of the same length, the hypergraph $\mathcal{H}_{n-1}^{\mathsf{I}}$ is $(n+1)$-uniform and its dual $\mathcal{H}_n^{\mathsf{D}}$ is $(n+1)$-regular. Later in section 3.2.3, we show that $\mathcal{H}_{n-1}^{\mathsf{I}}$ is also $(n+1)$-partite.

One of the key challenges encountered in dealing with the deletion channel and its related problems on strings is the structure of these hypergraphs. Apart from the statement that $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_{n-1}^{\mathsf{I}}$ are respectively, uniform and regular, very little can be discerned about their structure. In the remainder of this section we elaborate on the structure of these hypergraphs. Fortunately, due to the natural recursive nature of deletion and insertion, these hypergraphs admit some recursive structure. However this also means these hypergraphs are, in general, not balanced (recall from Berge [3] that a hypergraph is said to be balanced if its incidence matrix is balanced).

Let A_n denote the incidence matrix of $\mathcal{H}^{\mathsf{I}}_{n-1}$. As a consequence of Lemma 3.2, the incidence matrix of $\mathcal{H}^{\mathsf{D}}_n$ is $A_n^\top$. For example, for $n = 4$, arranging the vertices $\mathbb{F}_2^4$ and hyperedges $\mathbb{F}_2^3$ in lexicographic order, we get

$$A_4 = \begin{array}{c|cccc:cccc|}
 & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\
0000 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0001 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0010 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0011 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0100 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0101 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\
0110 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
0111 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ \hdashline
1000 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1001 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
1010 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\
1011 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
1100 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
1101 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1110 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1111 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array} \;. \tag{14}$$

Notice that the upper half of A_4 is lower triangular and the lower half is upper triangular. Also, the top left and bottom right submatrices are identical. The following lemma shows that this structure holds for arbitrary n, and it is recursive.

Lemma 3.3 *Let $n \in \mathbb{N}$. Let the columns and the rows of A_n be arranged in lexicographic order (increasing decimal value) (i.e., the first row is $0 \dots 0 \in \mathbb{F}_2^n$, last row is $1 \dots 1 \in \mathbb{F}_2^n$ and the first column is $0 \dots 0 \in \mathbb{F}_2^{n-1}$, last column is $1 \dots 1 \in \mathbb{F}_2^{n-1}$). Then we have the following recursion:*

$$A_n = \left[\begin{array}{c|c} A_{n-1} & \begin{matrix}\mathbf{0}\\ \mathbf{I}\end{matrix} \\ \hline \begin{matrix}\mathbf{I}\\ \mathbf{0}\end{matrix} & A_{n-1} \end{array}\right], \tag{15}$$

where $\mathbf{0}$ is a matrix of all zeros of appropriate size, $\mathbf{I}$ is the identity matrix of appropriate size. Furthermore,

$$A_n = \left[\begin{array}{c} A_n^u \\ \hline A_n^\ell \end{array}\right],$$

where $A_n^u, A_n^\ell \in \{0,1\}^{2^{n-1} \times 2^{n-1}}$, and A_n^u is lower triangular with each diagonal entry 1 and A_n^ℓ is upper triangular with each diagonal entry 1.

Proof : Let $y \in \mathbb{F}_2^{n-1}, x \in \mathbb{F}_2^{n-2}$. Consider the following cases:

1. $0y \in I(0x)$: This holds if and only if at least one of the following is true: a) $y = 0x$ or b) $y \in I(x)$. Indeed, the latter case subsumes the former, whereby $0y \in I(0x) \iff y \in I(x)$. This shows that the upper left submatrix of A_n is A_{n-1}.

2. $0y \in I(1x)$: This holds if and only if $y = 1x$. It follows that the upper right submatrix is $\begin{pmatrix}\mathbf{0}\\ I\end{pmatrix}$.

Cases of $1y \in I(0x)$ and $1y \in I(1x)$ follow similarly.

We argue the second claim by induction. Clearly, the claim is true for $n = 1$. Assume it holds for $n - 1$. Suppose $A_{n-1} = \left[\begin{array}{c} A_{n-1}^u \\ \hline A_{n-1}^\ell \end{array}\right]$ where A_{n-1}^u is lower triangular and A_{n-1}^ℓ is upper triangular with diagonal entries 1. Now by the recursion showed above, $A_n^u = \left[\begin{array}{cc} A_{n-1}^u & \mathbf{0} \\ \hline A_{n-1}^\ell & \mathbf{I} \end{array}\right]$ and $A_n^\ell = \left[\begin{array}{cc} \mathbf{I} & A_{n-1}^u \\ \hline \mathbf{0} & A_{n-1}^\ell \end{array}\right]$. It follows that A_n^u and A_n^ℓ must be upper and lower triangular respectively and have diagonal entries 1. ■

As a consequence of this recursion, certain properties of A_3 are inherited by A_n for all $n \geq 3$. As a specific consequence, this shows that A_n is not a balanced matrix.

Proposition 3.1 *For any $n \geq 3$, the matrix A_n is not balanced and thereby the hypergraphs $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_n^{\mathsf{I}}$ are not balanced.*

Proof : By (15), it suffices to show that A_3 has a submatrix corresponding to an odd cycle [3]. A_3 is the submatrix of A_4 in the top left half. Consider the submatrix generated by columns (in A_4) corresponding to $001, 010$ and 011, and rows (in A_4) corresponding to $0011, 0010$ and 0110. This submatrix is a submatrix of A_3 and is an odd cycle of size 3, whereby A_3 is not balanced. ■

3.2 Hypergraph problems

We now return to problems **D1,...,D5** stated in the introduction. Recall that our aim is to establish that in each of these problems, some aspect of the Varshamov-Tenengolts construction serves to solve it at least to a good approximation. We then address problems **I1,..., I5**.

Levenshtein [11] showed that the VT codes are 'perfect from below'. This means that, for any $a = 0, \ldots, n$ the union $\bigcup_{x \in \mathrm{VT}_a(n)} D(x) = \mathbb{F}_2^{n-1}$, in addition to the fact that the set $\mathrm{VT}_a(n)$ is a deletion-correcting code for each a. As a consequence, each of these sets is a perfect matching for the hypergraph $\mathcal{H}_n^{\mathsf{D}}$. Notice that this answers **D3** in the affirmative. Furthermore, we have the following proposition.

Proposition 3.2 *For any $n \geq 2$, each set $\mathrm{VT}_a(n), a = 0, \ldots, n$ is a transversal of $\mathcal{H}_{n-1}^{\mathsf{I}}$ and a matching of $\mathcal{H}_n^{\mathsf{D}}$. Furthermore,*

$$\kappa(\mathcal{H}_n^{\mathsf{D}}) = \tau(\mathcal{H}_{n-1}^{\mathsf{I}}) \leq |\mathrm{VT}_1(n)| = \min_{a=0,\ldots,n} |\mathrm{VT}_a(n)| \leq \frac{2^n}{n+1},$$
$$\nu(\mathcal{H}_n^{\mathsf{D}}) = p(\mathcal{H}_{n-1}^{\mathsf{I}}) \geq |\mathrm{VT}_0(n)| = \max_{a=0,\ldots,n} |\mathrm{VT}_a(n)| \geq \frac{2^n}{n+1}.$$

Proof : In both relations, the leftmost equality follows from the duality between $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_{n-1}^{\mathsf{I}}$ proved in Lemma 3.2. As recalled above, we have from [11] that each set $\mathrm{VT}_a(n), a = 0, \ldots, n$ is 'perfect from below'. Consequently, each set $\mathrm{VT}_a(n), a = 0, \ldots, n$ is a transversal of $\mathcal{H}_{n-1}^{\mathsf{I}}$ and a matching of $\mathcal{H}_n^{\mathsf{D}}$. This proves the first inequalities in the above relations. The rightmost equality follows from Ginsburg [7], where it is shown that $\mathrm{VT}_1(n)$ and $\mathrm{VT}_0(n)$ are respectively the smallest and the largest of the sets $\mathrm{VT}_a(n), a = 0, \ldots, n$. And the next inequalities follow from the fact that the smallest (largest) of these sets is at most (at least) the size of the average size of these sets. ■

Due to Lemma 3.2 and Proposition 3.2 there is no loss of generality in considering only the matching and transversal problems (and ignoring the covering and packing problems) on hypergraphs $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_n^{\mathsf{I}}$. In the rest of this paper we will adopt this convention.

We first derive approximation factors for $\mathrm{VT}_0(n)$ to solve **D1** and for $\mathrm{VT}_1(n)$ to solve **D2**. Our bounds are obtained from the following linear programming relaxations pertaining to problems **D1** and **D2**:

$$\begin{aligned} \nu^*(\mathcal{H}_{n-1}^{\mathsf{I}}) &= \max\{\mathbf{1}^\top z | A_n z \leq \mathbf{1}, z \geq 0\}, & \nu^*(\mathcal{H}_n^{\mathsf{D}}) &= \max\{\mathbf{1}^\top z | A_n^\top z \leq \mathbf{1}, z \geq 0\}, \\ \tau^*(\mathcal{H}_{n-1}^{\mathsf{I}}) &= \min\{\mathbf{1}^\top w | A_n^\top w \geq \mathbf{1}, w \geq 0\}, & \tau^*(\mathcal{H}_n^{\mathsf{D}}) &= \min\{\mathbf{1}^\top w | A_n w \geq \mathbf{1}, w \geq 0\}. \end{aligned} \tag{16}$$

Balanced hypergraphs are one of the most general classes for which fractional problems have been studied [3]. Since our hypergraphs are not balanced, there are very few standard results that can be leveraged to obtain bounds. In particular, a direct argument using Proposition 3.2 that exploits the fact that $\mathrm{VT}_1(n)$ is a transversal of $\mathcal{H}_{n-1}^{\mathsf{I}}$, leads to

$$\frac{2^{n-1}}{n} \leq |\mathrm{VT}_0(n-1)| \leq \nu(\mathcal{H}_{n-1}^{\mathsf{I}}) \leq \nu^*(\mathcal{H}_{n-1}^{\mathsf{I}}) = \tau^*(\mathcal{H}_{n-1}^{\mathsf{I}}) \leq \tau(\mathcal{H}_{n-1}^{\mathsf{I}}) \leq |\mathrm{VT}_1(n)| \leq \frac{2^n}{n+1}. \tag{17}$$

The leftmost and rightmost bounds are clearly not tight. A similar situation results for $\mathcal{H}_n^{\mathsf{D}}$; indeed for $\nu^*(\mathcal{H}_n^{\mathsf{D}})$ there is no simple upper bound.

In order to derive our bounds we need a simple observation about the number of runs of a string.

Lemma 3.4 *For any $x \in \mathbb{F}_2^*$ and any $y \in I(x)$, $r(x) \leq r(y) \leq r(x) + 2$.*

Proof : Let $r(x) = r$ and write x as a concatenation of its runs, $x = X_1 X_2 \ldots X_r$, where each X_i denotes a run of x. Let $\bar{x}$ be the inserted bit, let $y = x_1 \ldots x_i \bar{x} x_{i+1} \ldots x_n$ the obtained supersequence and consider the following cases.

1. $\bar{x}$ is inserted at the end of x (i.e., $i = n$ or $i = 0$). Therefore $y = \bar{x} x_1 \ldots x_n$ or $y = x_1 \ldots x_n \bar{x}$. In the former case, it is easy to see that
$$r(y) = \begin{cases} r(x) & \text{if } \bar{x} = x_1, \\ r(x) + 1 & \text{otherwise.} \end{cases}$$
The latter case follows similarly.

2. $\bar{x}$ is inserted between two runs, whereby $y = X_1 \ldots X_k \bar{x} X_{k+1} \ldots X_r$, where $1 < k < r$, and x_i is the last bit of run X_k. Notice that $\bar{x}$ is either equal to x_i or it is equal to x_{i+1} (the first bit of run X_{k+1}). Consequently, this insertion leads only to an elongation of one of the runs of x and no change in the number of runs. Thus $r(y) = r(x)$.

3. $\bar{x}$ is inserted inside a run, say run $X_k = x_j \ldots x_\ell$. In this case $y = x_1 \ldots x_i \bar{x} x_{i+1} \ldots x_n$ and the adjacent bits x_i and x_{i+1} are equal. Thus y can be written as $y = X_1 \ldots X_{k-1} x_j \ldots x_i \bar{x} x_{i+1} \ldots x_\ell X_{k+1} \ldots X_r$, where bits $x_j = x_{j+1} \ldots = x_i = x_{i+1} \ldots = x_\ell$, since they are from the run X_k of x. It is easy to see that in this case,
$$r(y) = \begin{cases} r(x) & \text{if } \bar{x} = x_i, \\ r(x) + 2 & \text{if otherwise.} \end{cases}$$

Summarizing the above cases, we see that $r(y)$ is no less than $r(x)$ and no greater than $r(x) + 2$. ■

The left inequality in this lemma ($r(x) \leq r(y)$) is also noted in [9].

3.2.1 The covering problem

We are now ready to prove the first of our main results.

Lemma 3.5 *For $n \in \mathbb{N}$, $n \geq 2$, the transversal number of $\mathcal{H}_{n-1}^{\mathsf{I}}$ admits the following bound:*

$$\tau(\mathcal{H}_{n-1}^{\mathsf{I}}) \geq \tau^*(\mathcal{H}_{n-1}^{\mathsf{I}}) \geq \frac{2^n}{n+1} - \frac{2^n(2n-4)+4}{n^3-n}. \tag{18}$$

Proof : By Proposition 2.1, $\tau(\mathcal{H}_{n-1}^{\mathsf{I}}) \geq \tau^*(\mathcal{H}_{n-1}^{\mathsf{I}}) = \nu^*(\mathcal{H}_{n-1}^{\mathsf{I}}) \geq \mathbf{1}^\top z$ for any z such that $z \geq 0$ and $A_n z \leq \mathbf{1}$. To prove the bound, we construct a suitable fractional matching. For any $x \in \mathbb{F}_2^{n-1}$, let $z(x) = \frac{1}{r(x)+2}$. It follows that for any $y \in \mathbb{F}_2^n$,
$$\sum_{x \in D(y)} z(x) = \sum_{x \in D(y)} \frac{1}{r(x)+2} \leq \sum_{x \in D(y)} \frac{1}{r(y)} = 1,$$
where the inequality follows from Lemma 3.4 and the equality is due to (12). Consequently, z is a fractional matching of $\mathcal{H}_{n-1}^{\mathsf{I}}$. The number of strings in $\mathbb{F}_2^{n-1}$ with r runs is $2 \times \binom{n-2}{r-1}$, whereby the weight of the fractional

matching z is $2\sum_{r=1}^{n-1}\binom{n-2}{r-1}\frac{1}{r+2}$. To obtain the above expression, simplify as follows

$$\begin{aligned} 2\sum_{r=1}^{n-1}\binom{n-2}{r-1}\frac{1}{r+2} &= 2\sum_{r=1}^{n-1}\frac{(n-2)!}{(n-r-1)!(r-1)!}\left[\frac{1}{r}+\frac{1}{r+2}-\frac{1}{r}\right] \\ &= 2\sum_{r=1}^{n-1}\frac{(n-2)!}{(n-r-1)!r!} - 4\sum_{r=1}^{n-1}\frac{(n-2)!}{(n-r-1)!r!}\left[\frac{1}{r+1}+\frac{1}{r+2}-\frac{1}{r+1}\right] \\ &= 2\sum_{r=1}^{n-1}\frac{(n-2)!}{(n-r-1)!r!} - 4\sum_{r=1}^{n-1}\frac{(n-2)!}{(n-r-1)!(r+1)!} + 4\sum_{r=1}^{n-1}\frac{(n-2)!}{(n-r-1)!(r+2)!}. \end{aligned}$$

We now obtain these sums in closed form, by writing the ratios of factorials in terms of binomial coefficients. Specifically, we get that the weight of the fractional matching is

$$\begin{aligned} &= \frac{2}{n-1}\sum_{r=1}^{n-1}\binom{n-1}{r} - \frac{4}{n(n-1)}\sum_{r=1}^{n-1}\binom{n}{r+1} + \frac{4}{(n+1)(n)(n-1)}\sum_{r=1}^{n-1}\binom{n+1}{r+2} \\ &= \frac{2^n-2}{n-1} - 4\frac{2^n-1-n}{n(n-1)} + 4\frac{2^{n+1}-1-(n+1)-n(n+1)/2}{(n+1)(n)(n-1)} \end{aligned}$$

Adding and subtracting $\frac{2^n}{n+1}$ and simplifying the above expression we get

$$\sum_{x\in\mathbb{F}_2^{n-1}} z(x) = \frac{2^n}{n+1} - \frac{2^n(2n-4)+4}{n^3-n}.$$

The result follows. ■

As a consequence, we have the following approximation factor for $\mathrm{VT}_1(n)$ to be a transversal of $\mathcal{H}^{\mathsf{I}}_{n-1}$.

Theorem 3.3 *Let $n\in\mathbb{N}, n\geq 2$. The set $\mathrm{VT}_1(n)$ is nearly an optimal transversal for $\mathcal{H}^{\mathsf{I}}_{n-1}$. Specifically,*

$$1 \leq \frac{|\mathrm{VT}_1(n)|}{\tau(\mathcal{H}^{\mathsf{I}}_{n-1})} \leq \frac{n^2-n}{n^2-3n+4-\frac{4}{2^n}}.$$

In particular, $\frac{|\mathrm{VT}_1(n)|}{\tau(\mathcal{H}^{\mathsf{I}}_{n-1})} \overset{n}{\to} 1$.

Proof : By Proposition 3.2, $\mathrm{VT}_1(n)$ is a transversal of $\mathcal{H}^{\mathsf{I}}_{n-1}$. To show the right inequality, notice that the set $\mathrm{VT}_1(n)$ being the smallest amongst the sets $\mathrm{VT}_0(n),\ldots,\mathrm{VT}_n(n)$ has size at most $\frac{2^n}{n+1}$, their average. Indeed this size is achieved for that n for which $n+1$ is a power of 2 (see Equations 7,8 in Sloane's [17]). Now using (18), $\frac{|\mathrm{VT}_1(n)|}{\tau(\mathcal{H}^{\mathsf{I}}_{n-1})}$ is upper bounded by $\frac{\frac{2^n}{n+1}}{\frac{2^n}{n+1}-\frac{2^n(2n-4)+4}{n^3-n}}$. Dividing by $\frac{2^n}{n+1}$ and simplifying gives the result. ■

3.2.2 The matching problem

Our next bound is on the matching problem on $\mathcal{H}^{\mathsf{D}}_n$. Note that this bound follows as a special case of our results in [9] where upper bounds are obtained for arbitrary number of deletions and arbitrary alphabet.

Lemma 3.6 *For $n\geq 2$, the matching number of $\mathcal{H}^{\mathsf{D}}_n$ admits the following upper bound,*

$$\nu(\mathcal{H}^{\mathsf{D}}_n) \leq \frac{2^n-2}{n-1}. \tag{19}$$

Proof : By Proposition 2.1, it suffices to show that there exists a fractional transversal of $\mathcal{H}_n^{\mathsf{D}}$ with this value. Consider the function w, where for any $x \in \mathbb{F}_2^{n-1}, w(x) = \frac{1}{r(x)}$. By Lemma 3.4, for any $y \in \mathbb{F}_2^n$,

$$\sum_{x \in D(y)} w(x) = \sum_{x \in D(y)} \frac{1}{r(x)} \geq \sum_{x \in D(y)} \frac{1}{r(y)} = 1,$$

where the equality is by (12). Consequently, w is a transversal of $\mathcal{H}_n^{\mathsf{D}}$. Now, as in Theorem 3.5, we calculate the value of this transversal as

$$2\sum_{r=1}^{n-1} \binom{n-2}{r-1} \frac{1}{r} = \frac{2}{n-1} \sum_{r=1}^{n-1} \frac{(n-1)!}{(n-r-1)!r!} = \frac{2^n - 2}{n-1},$$

as required. ■

The resulting approximation factor is as follows.

Theorem 3.4 *For $n \in \mathbb{N}, n \geq 2$, the set* $\mathrm{VT}_0(n)$ *is nearly an optimal matching for* $\mathcal{H}_n^{\mathsf{D}}$*. Specifically,*

$$1 \leq \frac{\nu(\mathcal{H}_n^{\mathsf{D}})}{|\mathrm{VT}_0(n)|} \leq \frac{(n+1)(1-\frac{2}{2^n})}{n-1}.$$

In particular $\frac{\nu(\mathcal{H}_n^{\mathsf{D}})}{|\mathrm{VT}_0(n)|} \xrightarrow{n} 1$.

Proof : By Proposition 3.2, $\mathrm{VT}_0(n)$ is a matching of $\mathcal{H}_n^{\mathsf{D}}$, which gives the left inequality. By Proposition 2.1, the upper bound on $\nu^*(\mathcal{H}_n^{\mathsf{D}})$ is an upper bound on $\nu(\mathcal{H}_n^{\mathsf{D}})$. Furthermore, the size of $\mathrm{VT}_0(n)$ is at least $\frac{2^n}{n+1}$ (Equation 8, Sloane [17]). Now using (19) and simplifying, we get the desired result. ■

3.2.3 Edge and vertex coloring

We now consider edge coloring and strong vertex coloring of $\mathcal{H}_n^{\mathsf{D}}$. For this, we recall a result of Cullina et al. from [5].

Lemma 3.7 (Section II.D [5]) *For any $n \in \mathbb{N}$, the sets* $\mathrm{VT}_0(n), \dots, \mathrm{VT}_n(n)$ *are an optimal coloring of* L_n*. Consequently,* $\chi(L_n) = n+1$.

As consequence, we have that the Varshamov-Tenengolts construction solves the edge coloring and strong vertex coloring on $\mathcal{H}_n^{\mathsf{D}}$.

Theorem 3.5 *For any $n \in \mathbb{N}, n \geq 2$ the sets* $\mathrm{VT}_0(n), \dots, \mathrm{VT}_n(n)$ *are an optimal edge coloring of* $\mathcal{H}_n^{\mathsf{D}}$ *and the sets* $\mathrm{VT}_0(n-1), \dots, \mathrm{VT}_{n-1}(n-1)$ *are an optimal strong vertex coloring of* $\mathcal{H}_n^{\mathsf{D}}$*. Consequently,* $q(\mathcal{H}_n^{\mathsf{D}}) = n+1$ *and* $\gamma(\mathcal{H}_n^{\mathsf{D}}) = n$.

Proof : An edge coloring of $\mathcal{H}_n^{\mathsf{D}}$ is equivalent to a vertex coloring of its line graph $L(\mathcal{H}_n^{\mathsf{D}})$, which by (8) is L_n. Thus, from Lemma 3.7, $\mathrm{VT}_0(n), \dots, \mathrm{VT}_n(n)$ is an optimal edge coloring of $\mathcal{H}_n^{\mathsf{D}}$. It follows that $q(\mathcal{H}_n^{\mathsf{D}}) = \chi(L_n) = n+1$. Since by (8), L_{n-1} is the line graph of $\mathcal{H}_{n-1}^{\mathsf{I}}$, $\mathrm{VT}_0(n-1), \dots, \mathrm{VT}_{n-1}(n-1)$ is an optimal edge coloring of $\mathcal{H}_{n-1}^{\mathsf{I}}$. But $\mathcal{H}_{n-1}^{\mathsf{I}} = (\mathcal{H}_n^{\mathsf{D}})^*$, whereby the edge coloring $\mathrm{VT}_0(n-1), \dots, \mathrm{VT}_{n-1}(n-1)$ of $\mathcal{H}_{n-1}^{\mathsf{I}}$ is also a strong vertex coloring of $\mathcal{H}_n^{\mathsf{D}}$. Therefore, $\gamma(\mathcal{H}_n^{\mathsf{D}}) = q(\mathcal{H}_{n-1}^{\mathsf{I}}) = \chi(L_{n-1}) = n$. ■

As a corollary, we get the following result about $\mathcal{H}_{n-1}^{\mathsf{I}}$.

Corollary 3.6 *For $n \in \mathbb{N}, n \geq 2$, the hypergraph* $\mathcal{H}_{n-1}^{\mathsf{I}}$ *is* $(n+1)$*-partite.*

Proof : The vertex set of $\mathcal{H}_{n-1}^{\mathsf{I}}$ is $\mathbb{F}_2^n$. Partition the vertex set into the $n+1$ sets $\mathrm{VT}_0(n), \dots, \mathrm{VT}_n(n)$. We need to show that each hyperedge of $\mathcal{H}_{n-1}^{\mathsf{I}}$ contains exactly one string from each of these sets. Suppose this is not true, i.e., suppose there a hyperedge of $\mathcal{H}_{n-1}^{\mathsf{I}}$ corresponding to a string $x \in \mathbb{F}_2^{n-1}$ such that (i) either there exists an a such that $I(x)$ contains no string from $\mathrm{VT}_a(n)$ or (ii) there exists an a such that $I(x)$ contains more than one string from $\mathrm{VT}_a(n)$. Out of these cases, (i) means that the string x is not covered by the deletion set of any string in $\mathrm{VT}_a(n)$. This is not possible since each set $\mathrm{VT}_k(n), k = 0, \ldots, n$ is a deletion cover of $\mathcal{H}_n^{\mathsf{D}}$. (ii) means that x is covered by the deletion set of two strings in $\mathrm{VT}_a(n)$. This violates the fact that each set $\mathrm{VT}_k(n), k = 0, \ldots, n$ is an edge color-class of $\mathcal{H}_n^{\mathsf{D}}$. Consequently, $\mathcal{H}_{n-1}^{\mathsf{I}}$ is $(n+1)$-partite. ■

3.2.4 Problems on the insertion hypergraph

We now consider the problems **I1**, . . . **I5**. Let us first consider problem **I3**: does there exist a perfect insertion matching? It is quite easy to argue that, in general, the answer is no. The size of insertion sets is the same for all string of the same length (cf. (11)). If there were a perfect insertion matching of $\mathcal{H}_{n-1}^{\mathsf{I}}$, then its size would be $\frac{2^n}{n+1}$. Clearly, this is not possible if $n+1$ is not a power of 2. However the absence of a perfect insertion matching is true for any n large enough; see, e.g., Levenshtein [11]. Thus, the answer to **I3** is negative.

Since by (9), $\nu(\mathcal{H}_n^{\mathsf{D}}) = \nu(\mathcal{H}_n^{\mathsf{I}})$, Theorem 3.6 provides an approximation factor for $\mathrm{VT}_0(n)$ to solve **I1** (the matching problem on $\mathcal{H}_n^{\mathsf{I}}$) as well. Now consider problems **I4, I5** on the edge and strong vertex coloring of $\mathcal{H}_n^{\mathsf{I}}$. By Theorem 3.5 and (10), the set $\mathrm{VT}_0(n), \ldots, \mathrm{VT}_n(n)$ is an optimal edge coloring of $\mathcal{H}_n^{\mathsf{I}}$ and $\mathrm{VT}_0(n+1), \ldots, \mathrm{VT}_{n+1}(n+1)$ is an optimal strong vertex coloring of $\mathcal{H}_n^{\mathsf{I}}$. Specifically, $q(\mathcal{H}_n^{\mathsf{I}}) = n+1$ and $\gamma(\mathcal{H}_n^{\mathsf{I}}) = n+2$.

There remains the problem **I2**, namely the covering problem of $\mathcal{H}_{n-1}^{\mathsf{I}}$. The object of interest here is the covering number of $\mathcal{H}_{n-1}^{\mathsf{I}}$, i.e., the smallest set of strings in $\mathbb{F}_2^{n-1}$ such that their insertion sets cover $\mathbb{F}_2^n$, and denoted $\kappa(\mathcal{H}_{n-1}^{\mathsf{I}})$. This is the same as the smallest set of strings in $\mathbb{F}_2^{n-1}$ that meet every deletion set from strings in $\mathbb{F}_2^n$, and is hence the transversal number of $\mathcal{H}_n^{\mathsf{D}}$, $\tau(\mathcal{H}_n^{\mathsf{D}})$. In addition to the size, the structure of the optimal cover is also of interest. Unfortunately, there is little we can say about the structure. Since each of the sets $\mathrm{VT}_a(n-1)$, for $a = 0, \ldots, n-1$ is a matching of $\mathcal{H}_{n-1}^{\mathsf{I}}$, and no perfect matching exists, it follows that these sets are not insertion covers.

We derive an upper bound on $\tau(\mathcal{H}_n^{\mathsf{D}})$ by invoking the following result from Berge [3, Theorem 12, p. 100].

Lemma 3.8 *For a hypergraph $\mathcal{H}$ with maximum degree Δ,*

$$\tau(\mathcal{H}) \leq (1 + \log \Delta)\tau^*(\mathcal{H}).$$

Using the upper bound on $\tau^*(\mathcal{H}_n^{\mathsf{D}})$ we have previously derived, we get the following result.

Theorem 3.7 *Let $n \in \mathbb{N}, n \geq 2$. The transversal number of $\mathcal{H}_n^{\mathsf{D}}$ satisfies*

$$\frac{2^n}{n+1} \leq \tau(\mathcal{H}_n^{\mathsf{D}}) \leq (1 + \log(n+1))\frac{2^n - 2}{n-1}.$$

Proof : The leftmost inequality is follows from observing that $\tau(\mathcal{H}_n^{\mathsf{D}}) \geq \nu(\mathcal{H}_n^{\mathsf{D}}) \geq |\mathrm{VT}_0(n)| \geq \frac{2^n}{n+1}$. Recall that $\mathcal{H}_n^{\mathsf{D}}$ is a regular hypergraph with degree $n+1$, therefore its maximum degree $\Delta(\mathcal{H}_n^{\mathsf{D}}) = n+1$. Now using Lemma 3.8 and Theorem 3.6, the result follows. ■

To the best of our knowledge this is a new bound. Clearly, the upper and lower bounds are not asymptotically equal and there is scope for obtaining tighter bounds.

3.3 Comparison with numerical results

Table 1 shows the values obtained when the linear programs corresponding to the optimal fractional matching and transversal of $\mathcal{H}_n^{\mathsf{D}}$ and $\mathcal{H}_{n-1}^{\mathsf{I}}$ were solved numerically. Also indicated are values of the analytical lower bound on $\tau(\mathcal{H}_{n-1}^{\mathsf{I}})$ (denoted `lb`) from (18) and the analytical upper bound on $\nu(\mathcal{H}_n^{\mathsf{D}})$ (denoted `ub`) from (19), and the sizes of sets $\mathrm{VT}_1(n)$ and $\mathrm{VT}_0(n)$. The results in the right half of the table have also been reported in [9]. The exact values of $\tau^*(\mathcal{H}_{n-1}^{\mathsf{I}})$ and $\nu^*(\mathcal{H}_n^{\mathsf{D}})$ indicate that $\mathrm{VT}_1(n)$ and $\mathrm{VT}_0(n)$ are very close to being optimal for transversal problem on $\mathcal{H}_{n-1}^{\mathsf{I}}$ and the matching problem on $\mathcal{H}_n^{\mathsf{D}}$ respectively, at least for $n \leq 14$. This suggests that quite likely, $\mathrm{VT}_1(n)$ is the smallest deletion cover (**D2**) and $\mathrm{VT}_0(n)$ is the largest deletion matching (**D1**). Indeed since for $n \leq 8$, $|\mathrm{VT}_1(n)| = \tau^*(\mathcal{H}_n^{\mathsf{D}})$, $\mathrm{VT}_1(n)$ does solve **D2** for $n \leq 8$. $\mathrm{VT}_0(n)$ has been confirmed to solve **D1** for $n \leq 10$ [18].

n	`lb`	$\lceil\tau^*(\mathcal{H}^{\mathsf{I}}_{n-1})\rceil$	$\lvert\mathrm{VT}_1(n)\rvert$	`ub`	$\lfloor\nu^*(\mathcal{H}^{\mathsf{D}}_n)\rfloor$	$\lvert\mathrm{VT}_0(n)\rvert$
1	–	1	1	–	1	1
2	1	1	1	2	2	2
3	2	2	2	3	2	2
4	3	3	3	4	4	4
5	4	5	5	7	6	6
6	7	9	9	12	10	10
7	13	16	16	21	17	16
8	23	28	28	36	30	30
9	42	50	51	63	53	52
10	77	92	93	113	96	94
11	143	169	170	204	175	172
12	268	312	315	372	321	316
13	503	580	585	682	593	586
14	949	1085	1091	1260	1104	1096

Table 1: Sizes of $\mathrm{VT}_1(n), \mathrm{VT}_0(n)$ and rounded values of linear programs obtained from exact the solution on Matlab. Also indicated are the lower bound from (18) in "`lb`" and the upper bound from (19) in "`ub`".

Our results augment the results reported for "challenge problems" on Sloane's website [18]. Intriguingly, however, neither the sequences obtained from our bounds nor the sequences obtained from the numerical solution of the linear programs show a match with the Online Encyclopedia of Integer Sequences [1].

4 Conclusions

This paper has modeled the problem of single-deletion/insertion correction of binary strings on hypergraphs. It was observed that the Varshamov-Tenengolts construction is an optimal edge-coloring where each color-class is a perfect matching of the deletion hypergraph, the largest color class is the maximum matching and the smallest color class is the minimum covering, to a good approximation. In addition, thanks to the duality between insertion and deletion hypergraphs, the VT construction also provided an optimal strong vertex coloring for these hypergraphs. These results indicate that perhaps there is a *meta-problem*, as yet undiscovered, that the VT construction solves, from which the solution of all of these hypergraph problems would follow. The deletion and insertion hypergraphs do not fall in any known category, other than the fact that they are regular and uniform, respectively. Yet they seem have the interesting property that certain edge-color classes of the deletion hypergraph also solve the matching and covering problem. This suggests the existence of a fascinating new class of hypergraphs with this property.

References

[1] *The On-Line Encyclopedia of Integer Sequences*, Dec 2012. URL: `http://oeis.org/`.

[2] K. Abdel-Ghaffar, F. Palunčić, H. Ferreira, and W. Clarke, *On Helberg's generalization of the levenshtein code for multiple Deletion/Insertion error correction*, IEEE Transactions on Information Theory, 58 (2012), pp. 1804 –1808.

[3] C. Berge, *Hypergraphs, Volume 45: Combinatorics of Finite Sets*, North Holland, 1 ed., Aug. 1989.

[4] S. Butenko, P. Pardalos, I. Sergienko, V. Shylo, and P. Stetsyuk, *Finding maximum independent sets in graphs arising from coding theory*, in Proceedings of the 2002 ACM symposium on Applied computing, SAC '02, New York, NY, USA, 2002, ACM, p. 542546.

[5] D. Cullina, A. A. Kulkarni, and N. Kiyavash, *A coloring approach to constructing deletion correcting codes from constant weight subgraphs*, in Proceedings of the ISIT, Cambridge, USA, 2012.

[6] ——, *Two approaches to the construction of deletion correcting codes: Weight partitioning and optimal colorings*, CoRR, abs/1211.4056 (2012). URL: `http://arxiv.org/abs/1211.4056`.

[7] B. D. Ginsburg, *On one number theory function applicable in the coding theory*, Problemy Kibernetiki, (1967), pp. 249–252.

[8] A. Helberg and H. Ferreira, *On multiple insertion/deletion correcting codes*, IEEE Transactions on Information Theory, 48 (2002), pp. 305 –308.

[9] A. A. Kulkarni and N. Kiyavash, *Non-asymptotic upper bounds on deletion correcting codes*, IEEE Transactions on Information Theory, 59 (2013), pp. 5115–5130.

[10] V. I. Levenshtein, *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet Physics Doklady, 10 (1966), pp. 707–710.

[11] ——, *On perfect codes in deletion and insertion metric*, Discrete Mathematics and Applications, 2 (1992), pp. 241–258.

[12] ——, *Efficient reconstruction of sequences from their subsequences or supersequences*, J. Comb. Theory, 93 (2001), pp. 310–332.

[13] ——, *Bounds for deletion/insertion correcting codes*, in 2002 IEEE International Symposium on Information Theory, 2002. Proceedings, Lausanne, Switzerland, 2002, p. 370.

[14] H. Mercier, V. Bhargava, and V. Tarokh, *A survey of error-correcting codes for channels with symbol synchronization errors*, IEEE Communications Surveys Tutorials, 12 (2010), pp. 87 –96.

[15] D. Sankoff and J. B. Kruskal, eds., *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Addison-Wesley Pub. Co., Advanced Book Program, 1983.

[16] L. Schulman and D. Zuckerman, *Asymptotically good codes correcting insertions, deletions, and transpositions*, IEEE Transactions on Information Theory, 45 (1999), pp. 2552 –2557. doi:10.1109/18.796406.

[17] N. J. A. Sloane, *On single-deletion-correcting codes*, in Codes and Designs: Proceedings of a Conference Honoring Professor Dijen K. Ray-Chaudhuri on the Occasion of His 65th Birthday, The Ohio State University, May 18-21, 2000, Walter de Gruyter, 2002.

[18] ——, *Challenge problems: Independent sets in graphs*, July 2011. URL: `http://neilsloane.com/doc/graphs.html`.

[19] G. M. Tenengolts, *Nonbinary codes, correcting single deletion or insertion*, Information Theory, IEEE Transactions on, 30 (1984), pp. 766 – 769. doi:10.1109/TIT.1984.1056962.

[20] R. R. Varshamov, *A class of codes for asymmetric channels and a problem from the additive theory of numbers*, IEEE Transactions on Information Theory, 19 (1973), pp. 92 – 95.

[21] R. R. Varshamov and G. M. Tenengolts, *Codes which correct single asymmetric errors (in Russian)*, Avtomatika i Telemekhanika, 6 (1965).