

© 2014 Daniel Cullina

CONSTANT COMPOSITION DELETION CORRECTING CODES

BY

DANIEL CULLINA

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Associate Professor Negar Kiyavash

ABSTRACT

We investigate deletion correcting codes and constant composition codes in particular. We use graph theoretic methods to characterize codes, establish bounds on code size, and describe constructions. The substring partial order has a surprising property: for any string, the number of superstrings of a particular length depends only on the length of the original string. We generalize this property to take compositions into account: for any string, the number of superstrings of a particular composition depends only on the composition of the original string. We present a bijective proof of this fact.

We apply this result to obtain a lower bound on the size of constant composition codes. We use a different technique to prove an upper bound. We construct binary constant composition single deletion correcting codes and show that they are asymptotically optimal and form an optimal coloring.

There is a natural distance on compositions that provides a lower bound on deletion distance. Unrestricted deletion correcting codes can be constructed from the union of constant composition codes as long as the set of compositions used themselves form a code. The nonbinary single deletion codes constructed by Tenengolts are a special case of this method.

We show that there is a qualitative difference between the problem of correcting a single deletion and the problem of correcting multiple deletions. In the single deletion case, the Varshamov Tenengolts codes are an optimal coloring of the confusion graph and each individual color class is asymptotically optimal. By constructing large cliques in the multiple deletion confusion graphs, we show that no construction can have both of the properties.

TABLE OF CONTENTS

CHAPTER 1	INTRODUCTION	1
1.1	Error models and the substring partial order	1
1.2	Overview of deletion correcting codes	2
1.3	Notation	4
1.4	The deletion channel and associated graphs	5
1.5	Levenshtein's lower bound	7
CHAPTER 2	SUPERSTRINGS OF FIXED COMPOSITION	10
2.1	A bijection	10
2.2	Asymptotic analysis	13
2.3	The binary case	15
2.4	Recovering Levenshtein's result	17
CHAPTER 3	APPLICATIONS TO CODES	19
3.1	Bounds on the size of constant-composition codes	19
3.2	Construction of binary constant weight single deletion correcting codes	22
3.3	An explicit coloring	23
3.4	Exact sizes of the codes in $L_{q,k,s}^R$	24
3.5	Cliques and optimal colorings	25
3.6	Fractional clique covers and optimal codes	26
3.7	Code construction by composition partitioning	28
3.8	Tenengolts q -ary single deletion correcting codes	31
3.9	Asymptotic optimal binary single deletion codes	32
CHAPTER 4	LARGE CLIQUES	34
4.1	Large cliques in $L_{q,n,s}$ and $L_{q,k,s}^R$	34
4.2	Consequences for code construction	37
REFERENCES		40

CHAPTER 1

INTRODUCTION

This thesis concerns deletion correcting codes. First we will explain the basic formulation of error correction. Then we will define the substring order relation, which is the source of the deletion error model.

1.1 Error models and the substring partial order

There are two basic approaches to modeling errors. Several concepts are common to both models. There are two spaces: an input space consisting of all of the possible values that a user can choose to store or transmit and an output space consisting of all possible error-corrupted versions of the inputs that might be read or received. A code is a subset of the input space and its elements are called codewords. A decoder for a code is a function from the output space to the code. Each codeword has a decoding region, the subset of the output space that is mapped to that codeword by the decoder.

The first type of error model is probabilistic. In such a model each point in the input space is associated with a probability distribution over the output space. When an input is selected, the output is sampled from the distribution corresponding to that input. If the output is inside the decoding region for the selected input, the output is decoded correctly. We say that a code and decoder are capable of correcting errors in this model if the probability of incorrect decoding is small for each codeword.

The second type of error model is adversarial. Each input is associated with a subset of the output space. When an input is selected, the resulting output must be from the associated subset. We say that a code and decoder are capable of correcting errors in this model if the decoding region for each codeword contains the set of possible outputs for that codeword.

In this thesis, we will only consider adversarial error models. For a partic-

ular adversarial error model, a fundamental problem is to determine the size of the largest correcting code.

Given a candidate code for a particular adversarial error model, a simple condition determines whether a decoder exists. If there is a pair of codewords that share a possible output, then no decoder can exist. If sets of possible outputs of the codewords are pairwise disjoint, then it is trivial to specify a decoder. It may still be nontrivial to find an efficient algorithm that implements the decoder, but we will not focus on this issue. Consequently, we have no need to further discuss decoders.

If a string x can be obtained from another string y by deleting some symbols, we say that x is a substring of y . In contrast to symbol erasure, symbol deletion leave no information about the location of the lost symbol. We can restate this more precisely. Let x and y be strings of length n and $n + 1$ respectively. We say that x is a substring of y if there are strings u , v , and w such that $x = u : v$, $v = u : w : v$, where $:$ indicates string concatenation. In this case, we write $x < y$. This order relation extends to strings with lengths that differ by more than one by transitivity.

1.2 Overview of deletion correcting codes

We are interested in the size of optimal deletion correcting codes. Bounds on the size of an optimal code depend on combinatorial results regarding the sets of substrings and superstrings of particular strings. Levenshtein provided the first upper and lower bounds on the size of deletion correcting codes [1]. The proofs of these results used bounds the number of substrings of a string that depend strongly on its number of runs. Recently, Liron and Langberg [2] found the best possible versions of these bounds for binary strings and identified the binary strings with the largest and smallest numbers of substrings for each length and number of runs. Cullina and Kiyavash improved upon Levenshtein's upper bound on code size [3] by obtaining analogous combinatorial results for channels that performed both deletions and insertions.

Let x be a n -symbol q -ary string. Levenshtein also showed that the number of $(n + s)$ -symbol superstrings of x depends only on n and s [4]. We extend this result to superstrings of a particular composition. The composition of a q -ary string is a vector of q non-negative integers. The entries are the number

of times that each symbol in the alphabet appears in the string. Let x be a q -ary string with composition $\mathbf{k}$. We show that the number of superstrings of x with composition $\mathbf{k} + \mathbf{s}$ depends only on q , $\mathbf{k}$, and $\mathbf{s}$ and obtain an expression for this number. We apply this to obtain upper and lower bounds on the size of constant-composition codes.

The composition is a very natural object. The set of q -ary strings under concatenation is the free monoid on q generators. The set of q -ary compositions under vector addition is the free commutative monoid on q generators. There is a natural distance on compositions that provides a lower bound for the deletion distance on strings. This lower bound can be thought of as the “trivial” component of deletion distance. By considering constant-composition spaces, we investigate the “nontrivial” component.

We can apply this idea to code construction. A deletion correcting code can be constructed from a code on compositions and a collection of constant composition codes. This construction corrects errors in two stages. First, the composition of the transmitted codeword is determined, then the codeword itself. This technique was implicitly applied by Tenengolts to construct single deletion correcting codes for non-binary alphabets [5]. We significantly generalize it.

Levenshtein investigated deletion error correction in permutations [6] and constructed single-deletion correcting codes over permutations. The permutations of an alphabet are an example of a constant composition set of strings and consequently are a special case of the problem that we consider. Gabrys et al. gave an overview of the various erasure and deletion error models for permutations [7]. If we represent permutations as square zero-one matrices, we can delete either columns or rows of the matrix. Deleting columns corresponds to deleting symbols of the word form of the permutation. Deleting rows corresponds to deleting symbols of the word form of the inverse permutation.

Multipermutations form another constant composition set. One natural model of deletion errors on multipermutations is a special case of the model we consider. Farnoud and Milenkovic as well as Sala et al. have investigated another model of deletion errors on multipermutations motivated by error [8, 9]. In this model, a multipermutation is treated as an equivalence class of permutations and errors act on these permutations rather than the multipermutations directly.

The rest of this thesis is organized as follows. The remainder of this chapter is devoted to notation, definitions, and the statement a few important existing results. We also give a short proof of Levenshtein's lower bound on the size of the optimal deletion correcting code. This proof has the same structure as one of the primary results of Chapter 3.

In Chapter 2 we determine the number of superstrings of a particular composition for any initial string. This number only depends on the composition of the initial string and the composition of the superstring. To do this, we describe a bijection between the set of superstrings and a special fixed set. We also analyze the asymptotic rate of growth of the size of these sets.

In Chapter 3, we apply the results of Chapter 2 in several ways. In Section 3.1 we obtain upper and lower bounds on the size of optimal constant-composition s -deletion-correcting codes. In Section 3.2 we construct nearly optimal constant-composition single-deletion-correcting binary codes. In Section 3.7 we construct unrestricted deletion-correcting codes from codes on compositions and constant-composition deletion-correcting codes.

In Chapter 4, we show that there is a qualitative difference between the problem of correcting a single deletion and the problem of correcting multiple deletions. In the single deletion case, the Varshamov Tenengolts codes are an optimal coloring of the confusion graph and each individual color class is asymptotically optimal. By constructing large cliques in the multiple deletion confusion graphs, we show that no construction can have both of the properties.

1.3 Notation

For a set S , let 2^S be the power set of S . Let $[n]$ be the set of non-negative integers less than n , $\{0, 1, \dots, n-1\}$. Let $[q]^n$ be the set of q -ary strings of length n . Index the symbols in a string $x \in [q]^n$ by $[n]$, so $x = (x_0, \dots, x_{n-1})$. Let $[q]^*$ be the set of q -ary strings of any length. The substring relation is a partial order on $[q]^*$, so for strings x and y , write $x \leq y$ if x is a substring of y .

Vectors in $\mathbb{N}^q$, $\mathbb{Z}^q$, and $\mathbb{Q}^q$ are denoted by bold roman letters. We will frequently use the product partial order on such vectors: $\mathbf{r} \leq \mathbf{s}$ if and only if $\mathbf{r}_i \leq \mathbf{s}_i$ for all $i \in [q]$. For a vector $\mathbf{k} \in \mathbb{Z}^q$, we let $\|\mathbf{k}\| = \sum_i |\mathbf{k}_i|$. Define

the function $\# : [q]^n \rightarrow \mathbb{N}^q$ such that $\#(x)_i$ is the number of times that i appears in x . We refer to $\#(x)$ as the composition of x . A run of the i -symbol in a string is a maximal interval in the string that contains only i -symbols. Define the function $r : [q]^n \rightarrow \mathbb{N}^q$ such that $r(x)_i$ is the number of runs of the i -symbol in x . For $q, n \in \mathbb{N}$, let $C(q, n)$ be the set of compositions with total size n :

$$C(q, n) = \{\mathbf{k} \in \mathbb{N}^q \mid \|\mathbf{k}\| = n\}.$$

It is well known that

$$|C(q, n)| = \binom{n+q-1}{n}.$$

For $\mathbf{k} = (\mathbf{k}_0, \dots, \mathbf{k}_{q-1}) \in C(q, n)$, let $M(\mathbf{k})$ be the set of strings with composition $\mathbf{k}$:

$$M(\mathbf{k}) = \{x \in [q]^n \mid \#(x) = \mathbf{k}\}.$$

It is well known that

$$|M(\mathbf{k})| = \binom{\|\mathbf{k}\|}{\mathbf{k}_0, \dots, \mathbf{k}_{q-1}} = \frac{\|\mathbf{k}\|!}{\mathbf{k}_0! \dots \mathbf{k}_{q-1}!}.$$

For a matrix A , let $A_{i,\bullet}$ be the vector corresponding to the i th row and let $A_{\bullet,j}$ be the vector corresponding to the j th column.

We need the following asymptotic notation: let $a(n) \sim b(n)$ denote that $\lim_{n \rightarrow \infty} \frac{a(n)}{b(n)} = 1$ and $a(n) \lesssim b(n)$ denote that $\lim_{n \rightarrow \infty} \frac{a(n)}{b(n)} \leq 1$. We use the following asymptotic equality frequently: for fixed c , $\binom{n}{c} \sim \frac{n^c}{c!}$.

1.4 The deletion channel and associated graphs

Let us formalize the problem of correcting deletions by defining the following sets.

Definition 1. For $x \in [q]^n$, define

$$D_s(x) = \{z \in [q]^{n-s} \mid z \leq x\},$$

the set of substrings of x that can be produced by s deletions. Similarly

$$I_s(x) = \{w \in [q]^{n+s} \mid w \geq x\},$$

the set of superstrings of x that can be produced by s insertions.

The n -symbol s -deletion channel takes a string of length n and outputs a substring of length $n - s$. Thus $D_s(x)$ is the set of possible outputs for an input x . If x is the output from the channel, $I_s(x)$ is the set of possible inputs.

To facilitate the analysis of constant-composition codes we will need the restricted-composition analogues of these sets. Throughout this thesis, our notation for sets, quantities, and graphs that somehow involve a restricted composition will include a superscript R to help distinguish these objects from their unrestricted variants.

Definition 2. For $\mathbf{k}, \mathbf{r} \in C(q, n)$, $\mathbf{k} \geq \mathbf{r}$, and $x \in M(\mathbf{k})$, let

$$D_{\mathbf{s}}^R(x) = \{z \in M(\mathbf{k} - \mathbf{s}) \mid z \leq x\}.$$

This is the set of substrings of x of composition $\mathbf{k} - \mathbf{s}$. Each of these can be produced by deleting $\mathbf{s}_i$ i -symbols from x .

$$I_{\mathbf{s}}^R(x) = \{w \in M(\mathbf{k} + \mathbf{s}) \mid w \geq x\}.$$

This is the set of superstrings of x of composition $\mathbf{k} + \mathbf{s}$. Each of these can be produced by inserting $\mathbf{s}_i$ new i -symbols into x .

When two inputs share a common output they can potentially be confused by the receiver. We are interested in codes that allow the correction of s deletions.

Definition 3. A q -ary n -symbol s -deletion correcting code is a set $C \subset [q]^n$ such that for any two distinct strings $x, y \in C$, $D_s(x) \cap D_s(y)$ is empty.

This leads to a natural distance measure on q -ary strings.

Definition 4. Let $x \in [q]^m$ and $y \in [q]^n$ and let $z \in [q]^l$ be a common substring of x and y of maximum length. Define the deletion distance between x and y to be $d_L(x, y) = m + n - 2l$.

To motivate this definition, observe that x can be transformed into z by $m - l$ deletion operations and z can be transformed into y by $n - l$ insertion

operations. It is well known that deletion distance is a metric [1]. If x and y are the same length, then the deletion distance between them is even and

$$d_L(x, y)/2 = \min\{s \in \mathbb{N} | D_s(x) \cap D_s(y) \neq \emptyset\}.$$

Thus any two codewords in an s -deletion correcting code have deletion distance strictly greater than $2s$. We will use this to give a graphical characterization of s -deletion correcting code.

In order to characterize deletion correcting codes in terms of graphs, we will need to some notation and a few definitions. All of these are sourced from West [10]. Given a graph G , let $V(G)$ denote its vertex set and let $E(G)$ denote its edge set. Given $S \subseteq V(G)$, the *subgraph induced by S* contains the vertices in S and the edges in $E(G)$ that have both endpoints in S . The *degree* of a vertex is the number of adjacent vertices. The average degree of the vertices of G is denoted by $\bar{d}(G)$. Because each edge contributes to the degree of two vertices, $\bar{d}(G) = 2|E(G)|/|V(G)|$.

An *independent set* in a graph is a set of vertices that are all nonadjacent. The size of a largest independent set in a graph G is denoted by $\alpha(G)$. A version of Turán's theorem [10, p. 122] gives the following lower bound on $\alpha(G)$:

$$\alpha(G) \geq \frac{|V(G)|}{\bar{d}(G) + 1}. \quad (1.1)$$

The following graphs give a characterization of s -deletion correcting codes and constant-composition s -deletion correcting codes.

Definition 5. For all $s, n \in \mathbb{N}$, let $L_{q,n,s}$ be a graph on the vertex set $[q]^n$. In $L_{q,n,s}$, vertices x and y are adjacent if and only if $d_L(x, y) \leq 2s$. For all $s, n \in \mathbb{N}$ and $\mathbf{k} \in C(q, n)$, let $L_{q,\mathbf{k},s}^R$ be the subgraph of $L_{q,n,s}$ induced by the vertices in $M(\mathbf{k})$.

An independent set in $L_{q,n,s}$ is an s -deletion correcting code and an independent set in $L_{q,\mathbf{k},s}^R$ is a constant-composition s -deletion correcting code.

1.5 Levenshtein's lower bound

Now that we have established some terminology and notation, we can concisely express some important existing results. Levenshtein proved the fol-

lowing asymptotic upper and lower bounds on the size of optimal s -deletion correcting codes [1]:

$$\frac{q^{n+s}}{(q-1)^{2s} \binom{n}{s}^2} \lesssim \alpha(L_{q,n,s}) \lesssim \frac{q^n}{(q-1)^s \binom{n}{s}}. \quad (1.2)$$

In this section, we present a proof of Levenshtein's lower bound on the size of the largest s -deletion correcting code. This is the lower bound of (1.2). Our proof of an analogous result for constant-composition deletion graphs follows the same basic structure, so we present this proof first to make our argument more clear.

Levenshtein showed that the number of distinct superstrings of a string produced by s insertions only depends on the length of the string [4]. For each $x \in [q]^n$, $|I_s(x)|$ is equal to the constant

$$I_{q,n,s} = \sum_{i=0}^s (q-1)^i \binom{n+s}{i}. \quad (1.3)$$

Note that for fixed s

$$I_{q,n,s} \sim (q-1)^s \binom{n}{s}. \quad (1.4)$$

We can use this to prove an upper bound on the average degree of the deletion graphs.

Lemma 1. *For all $s, n \in \mathbb{N}$ with $s \leq n$, the average degree in $L_{q,n,s}$ satisfies*

$$\bar{d}(L_{q,n,s}) \leq \frac{2}{q^s} \binom{I_{q,n-s,s}}{2} \sim \frac{(q-1)^{2s}}{q^s} \binom{n}{s}^2.$$

Proof: There are $q^{n-s} \binom{I_{q,n,s}}{2}$ triples $(z, (x, y)) \in [q]^{n-s} \times \binom{[q]^n}{2}$ such that $z \leq x$ and $z \leq y$. If $x, y \in [q]^n$ are adjacent in $L_{q,n,s}$, then they have at least one common substring of length $n-s$ and appear in at least one triple. The average degree, $\bar{d}(L_{q,n,s})$, is given by

$$\frac{2|E(L_{q,n,s})|}{|V(L_{q,n,s})|} \leq \frac{2q^{n-s}}{q^n} \binom{I_{q,n-s,s}}{2} \sim \frac{(q-1)^{2s}}{q^s} \binom{n}{s}^2,$$

where the asymptotic equality follows from (1.4). ■

Levenshtein's original lower bound follows immediately.

Theorem 1. *For all $s, n \in \mathbb{N}$, there are s -deletion correcting codes of size at*

least

$$\frac{q^{n+s}}{I_{q,n-s,s}(I_{q,n-s,s} - 1) + q^s} \sim \frac{q^{n+s}}{(q-1)^{2s} \binom{n}{s}^2},$$

where the asymptotic equality is for fixed s .

Proof: This follows immediately from Lemma 1 and Turán's theorem (1.1). ■

Levenshtein's original proof of the asymptotic version of this lower bound used a different argument based on sphere-covering [1]. He later proved the non-asymptotic version using the argument that we presented here [11].

CHAPTER 2

SUPERSTRINGS OF FIXED COMPOSITION

2.1 A bijection

Levenshtein showed that for any string, the number of distinct superstrings a particular length depends only on the length of the starting string [4]. In this section, we will show that the number of distinct superstrings of a particular composition depends only on the composition of the starting string. Furthermore, we will derive an expression for this number. This combinatorial result will be important in the analysis of the constant-composition deletion graphs $L_{q,\mathbf{k},\mathbf{s}}^R$.

More concisely, we will show that for all $\mathbf{k}, \mathbf{s} \in \mathbb{N}^q$ and all $x \in M(\mathbf{k})$, $|I_{\mathbf{s}}^R(x)|$ only depends on $\mathbf{k}$ and $\mathbf{s}$. To do this, we will define a set $J_{q,\mathbf{k},\mathbf{s}}^R$, an encoding procedure that maps $I_{\mathbf{s}}^R(x)$ to $J_{q,\mathbf{k},\mathbf{s}}^R$, and a decoding procedure that does the reverse. We show that the decoding procedure inverts encoding, so for each starting string $x \in M(\mathbf{k})$, $|I_{\mathbf{s}}^R(x)| = |J_{q,\mathbf{k},\mathbf{s}}^R|$.

The algorithms are based on the following observation. Let $x \in [q]^n$ and let $y \in I_s(x)$. Consider a subset of the indices of y , $S \in 2^{[n+s]}$. If deleting the symbols at all other indices in y produces x , call S an appearance of x in y . There may be multiple appearances of x in y , but there is a unique leftmost appearance that is found greedily. Suppose S is the leftmost appearance of x in y , $i \notin S$, $j \in S$, $i < j$, and $k \notin S$ for all $i < k < j$. If $y_i = y_j$, the greedy search would have added i to S before inspecting j . Thus $y_i \neq y_j$.

The encoding procedure works as follows. Find the leftmost appearance of x in y . The appearance partitions y into $n + 1$ (possibly empty) blocks. Add the symbol from x to end of each of the first n blocks. Group the first n blocks by the symbol of x that appears to their left. For each group, record the lengths of the blocks and concatenate them into a single string. The result is q strings produced by concatenation plus the final block of y . The

Algorithm 1 Encoding $y \in I_{\mathbf{s}}^R(x)$ as $(u, v, w) \in J_{q, \mathbf{k}, \mathbf{s}}^R$ and decoding y from (u, v, w)

ENCODE : $[q]^* \times [q]^* \rightarrow (\mathbb{N}^*)^q \times ([q]^*)^q \times [q]^*$

ENCODE(x, y)

if $x = \epsilon$ **then**

return $(\epsilon^q, \epsilon^q, y)$

else

$(l, z, y) \leftarrow \text{SPLITAT}(\text{HEAD}(x), y)$

$(u, v, w) \leftarrow \text{ENCODE}(\text{TAIL}(x), y)$

$u_i \leftarrow l : u_i.$

$v_i \leftarrow z : v_i.$

return (u, v, w)

end if

SPLITAT : $[q] \times [q]^* \rightarrow \mathbb{N} \times [q]^* \times [q]^*$

SPLITAT(i, y)

$j \leftarrow \text{HEAD}(y)$

if $i = j$ **then**

return $(0, \epsilon, \text{TAIL}(y))$

else

$(l, z, y) \leftarrow \text{SPLITAT}(i, \text{TAIL}(y))$

return $(1 + l, j : z, y)$

end if

Proof: We prove this by induction on x . In the base case, $x = \epsilon$, $\mathbf{k} = \mathbf{0}$, and $I_{\mathbf{s}}^R(x) = M(\mathbf{s})$. All of the elements of $J_{q, \mathbf{0}, \mathbf{s}}^R$ are of the form $(\epsilon, \epsilon^q, \epsilon^q, y)$ because $C(0, 0)$ contains one element while $C(0, m)$ is empty for positive m . We have $\text{ENCODE}(\epsilon, y) = (\epsilon^q, \epsilon^q, y)$ and $\text{DECODE}(\epsilon, \epsilon^q, \epsilon^q, y) = y$.

If x is nonempty, we can let $x = i : x'$. For each $y \in I_{\mathbf{s}}^R(x)$, there is a unique $z \in ([q] \setminus \{i\})^*$ such that $y = z : i : y'$. The function SPLITAT finds exactly this decomposition of y . Let $(u, v, w) \in J_{q, \mathbf{k}, \mathbf{s}}^R$. Because $\#(x)_i \geq 1$, so u_i is nonempty and we can let $u_i = l : u'_i$. The length of v_i is the sum of the entries of u_i , so we can let $v_i = z' : v'_i$ such that z' has length l . The function SPLITAFTER finds exactly this decomposition of v_i . Finally, let $u'_j = u_j$ and $v'_j = v_j$ for $j \neq i$.

If $(u', v', w') = \text{ENCODE}(x', y')$, then by the induction hypothesis, $y' = \text{DECODE}(x', u', v', w')$. If $z = z'$, then from the definition of ENCODE, $\text{ENCODE}(i : x', z : i : y') = (u, v, w')$. Similarly, $\text{DECODE}(i : x, u, v, w') = z : i : y'$. ■

Algorithm 2 Encoding $y \in I_s^R(x)$ as $(u, v, w) \in J_{q, \mathbf{k}, \mathbf{s}}^R$ and decoding y from (u, v, w)

DECODE : $[q]^* \times (\mathbb{N}^*)^q \times ([q]^*)^q \times [q]^* \rightarrow [q]^*$

DECODE(x, u, v, w)

if $x = \epsilon$ **then**

return w

else

$i \leftarrow \text{HEAD}(x)$

$(l, u_i) \leftarrow (\text{HEAD}(u_i), \text{TAIL}(u_i))$

$(z, v_i) \leftarrow \text{SPLITAFTER}(i, l, v_i)$

return $z : \text{DECODE}(\text{TAIL}(x), u, v, w)$

end if

SPLITAFTER : $[q] \times \mathbb{N} \times [q]^* \rightarrow [q]^* \times [q]^*$

SPLITAFTER(i, l, y)

if $l = 0$ **then**

return (i, y)

else

$j \leftarrow \text{HEAD}(y)$

$(z, y) \leftarrow \text{SPLITAFTER}(i, l - 1, \text{TAIL}(y))$

return $(j : z, y)$

end if

2.2 Asymptotic analysis

The following two lemmas give a better understanding of the size of $I_{q, \mathbf{k}, \mathbf{s}}^R$.

For fixed $\mathbf{s}$, we can obtain a simple asymptotic expression for the size.

Lemma 2. Fix $\mathbf{s} \in \mathbb{N}^q$ and $\mathbf{p} \in \mathbb{Q}^q$ such that $\mathbf{p} \geq \mathbf{0}$ and $\|\mathbf{p}\| = 1$. For $n \rightarrow \infty$ and $\mathbf{k} = \mathbf{p}n \in \mathbb{N}^q$,

$$I_{q, \mathbf{k}, \mathbf{s}}^R \sim \binom{n}{\|\mathbf{s}\|} |M(\mathbf{s})| \prod_{j \in [q]} (1 - \mathbf{p}_i)^{\mathbf{s}_i}.$$

Proof: From Theorem 2, $I_{q, \mathbf{k}, \mathbf{s}}^R$ is equal to

$$\sum_{A, \mathbf{b}} |M(\mathbf{b})| \prod_{j \in [q]} \binom{|M(A_{\bullet, j})| + \mathbf{k}_j - 1}{|M(A_{\bullet, j})|} |M(A_{\bullet, j})|, \quad (2.1)$$

where the sum ranges over the matrices $A \in \mathbb{N}^{q \times q}$ and the vectors $\mathbf{b} \in \mathbb{N}^q$ satisfying $A\mathbf{1} + \mathbf{b} = \mathbf{s}$ and $A_{j, j} = 0$ for all j .

To simplify this, first note that

$$\prod_{j \in [q]} \binom{\|A_{\bullet,j}\| + n\mathbf{p}_j - 1}{\|A_{\bullet,j}\|} \sim \prod_{j \in [q]} \frac{(n\mathbf{p}_j)^{\|A_{\bullet,j}\|}}{\|A_{\bullet,j}\|!}. \quad (2.2)$$

For $\mathbf{p}_i > 0$, this is the standard asymptotic equality for binomial coefficients. For $\mathbf{p}_i = 0$, this is actually an equality because $\binom{m-1}{m}$ and $\frac{0^m}{m!}$ each equal one for $m = 0$ and equal zero for all positive integers.

$$\begin{aligned} & \sum_A \prod_{j \in [q]} \frac{(n\mathbf{p}_j)^{\|A_{\bullet,j}\|} |M(A_{\bullet,j})|}{\|A_{\bullet,j}\|!} \\ \stackrel{(a)}{=} & \sum_A \prod_{i \in [q]} \prod_{j \neq i} \frac{(n\mathbf{p}_j)^{A_{i,j}}}{A_{i,j}!} \\ \stackrel{(b)}{=} & \sum_A \prod_{i \in [q]} \frac{n^{\|A_{i,\bullet}\|} |M(A_{i,\bullet})|}{\|A_{i,\bullet}\|!} \prod_{j \neq i} \mathbf{p}_j^{A_{i,j}} \\ \stackrel{(c)}{=} & \sum_A \frac{n^{\|\mathbf{s}-\mathbf{b}\|} |M(\mathbf{s}-\mathbf{b})|}{\|\mathbf{s}-\mathbf{b}\|!} \prod_{i \in [q]} |M(A_{i,\bullet})| \prod_{j \neq i} \mathbf{p}_j^{A_{i,j}} \\ \stackrel{(d)}{=} & \frac{n^{\|\mathbf{s}-\mathbf{b}\|} |M(\mathbf{s}-\mathbf{b})|}{\|\mathbf{s}-\mathbf{b}\|!} \prod_{i \in [q]} \sum_{A_{i,\bullet}} |M(A_{i,\bullet})| \prod_{j \neq i} \mathbf{p}_j^{A_{i,j}} \\ \stackrel{(e)}{=} & \frac{n^{\|\mathbf{s}-\mathbf{b}\|} |M(\mathbf{s}-\mathbf{b})|}{\|\mathbf{s}-\mathbf{b}\|!} \prod_{i \in [q]} (1 - \mathbf{p}_i)^{\mathbf{s}_i} \\ \stackrel{(f)}{\sim} & \binom{n}{\|\mathbf{s}-\mathbf{b}\|} |M(\mathbf{s}-\mathbf{b})| \prod_{i \in [q]} (1 - \mathbf{p}_i)^{\mathbf{s}_i}. \end{aligned} \quad (2.3)$$

In (a), we expand $|M(A_{\bullet,j})|$ into a product of factorials. Because $A_{i,i} = 0$ for all i , we can drop these terms from the product. In (b), we collect the $A_{i,j}!$ terms to form $|M(A_{i,\bullet})|$. In (c), we rewrite $\|A_{i,\bullet}\|$ as $\mathbf{s}_i - \mathbf{b}_i$. Then we collect factorial terms to form $|M(\mathbf{s}-\mathbf{b})|$. Each row of A has independent constraints, so in (d) we can move the sum over A inside of the product over i . In (e) we apply the multinomial theorem to each sum over $A_{i,\bullet}$. In (f) we apply the standard asymptotic equality.

From (2.1), (2.2), and (2.3), $I_{q,\mathbf{k},\mathbf{s}}^R$ is asymptotic to

$$\sum_{\mathbf{b} \in C(q,\mathbf{s})} |M(\mathbf{b})| \binom{n}{\|\mathbf{s}-\mathbf{b}\|} |M(\mathbf{s}-\mathbf{b})| \prod_{i \in [q]} (1 - \mathbf{p}_i)^{\mathbf{s}_i - \mathbf{b}_i}.$$

This is a finite sum of polynomials in n . The highest degree term corresponds to $\mathbf{c} = \mathbf{0}$, so

$$I_{q,\mathbf{k},\mathbf{s}}^R \sim \binom{n}{\|\mathbf{s}\|} |M(\mathbf{s})| \prod_{i \in [q]} (1 - \mathbf{p}_j)^{\mathbf{s}_i}.$$

■

2.3 The binary case

In the binary case, we can significantly simplify the expression for $I_{q,\mathbf{k},\mathbf{s}}^R$ without taking any asymptotics.

We will use a few well-known combinatorial results in the proof of Lemma 3. Vandermonde's identity is

$$\binom{a+b}{c} = \sum_i \binom{a}{i} \binom{b}{c-i}. \quad (2.4)$$

This bijection corresponding to this identity splits a string of length $a+b$ into a string of length a and a string of length b . The sum is over all possible distributions of c ones in the original string between the new strings.

A version of Vandermonde's identity related to compositions is

$$\binom{c+a+b-1}{c} = \sum_i \binom{i+a-1}{i} \binom{c-i+b-1}{c-i}. \quad (2.5)$$

This decomposes an element of $C(a+b, c)$ into an element of $C(a, i)$ and an element of $C(b, c-i)$. This also corresponds to breaking a string with $a+b-1$ zeros at the location of its a th zero, so $a-1$ zeros are in the first fragment and $b-1$ are in the second fragment.

Lemma 3. *For all $\mathbf{k}, \mathbf{s} \in \mathbb{N}^2$,*

$$I_{2,\mathbf{k},\mathbf{s}}^R = \sum_{i \geq 0} \binom{\mathbf{s}_0 + \mathbf{k}_1}{\mathbf{s}_0 - i} \binom{\mathbf{s}_1 + \mathbf{k}_0}{\mathbf{s}_1 - i}.$$

Proof: From Theorem 2, $I_{2,\mathbf{k},\mathbf{s}}^R$ is equal to

$$\sum_{A, \mathbf{b}} |M(\mathbf{b})| \prod_{j \in [2]} \binom{|M(A_{\bullet,j})| + \mathbf{k}_j - 1}{|M(A_{\bullet,j})|} |M(A_{\bullet,j})|. \quad (2.6)$$

Because $A_{0,0} = A_{1,1} = 0$, A is fully determined by $\mathbf{s}$ and $\mathbf{b}$: $\|A_{\bullet,1}\| = A_{0,1} = \mathbf{s}_0 - \mathbf{b}_0$ and $\|A_{\bullet,0}\| = A_{1,0} = \mathbf{s}_1 - \mathbf{b}_1$.

Thus $I_{2,\mathbf{k},\mathbf{s}}^R$ is equal to

$$\sum_{\mathbf{b}_0} \sum_{\mathbf{b}_1} \binom{\mathbf{s}_1 - \mathbf{b}_1 + \mathbf{k}_0 - 1}{\mathbf{s}_1 - \mathbf{b}_1} \binom{\mathbf{s}_0 - \mathbf{b}_0 + \mathbf{k}_1 - 1}{\mathbf{s}_0 - \mathbf{b}_0} \binom{\|\mathbf{b}\|}{\mathbf{b}_0}.$$

From Vandermonde's identity, (2.4), we have

$$\begin{aligned} & \sum_{\mathbf{b}_0} \sum_{\mathbf{b}_1} \binom{\mathbf{s}_1 - \mathbf{b}_1 + \mathbf{k}_0 - 1}{\mathbf{s}_1 - \mathbf{b}_1} \binom{\mathbf{s}_0 - \mathbf{b}_0 + \mathbf{k}_1 - 1}{\mathbf{s}_0 - \mathbf{b}_0} \times \\ & \sum_i \binom{\mathbf{b}_0}{\mathbf{b}_0 - i} \binom{\mathbf{b}_1}{i}. \end{aligned}$$

Exchanging the order of the sums yields

$$\begin{aligned} & \sum_{i \geq 0} \left(\sum_{\mathbf{b}_0} \binom{\mathbf{s}_0 - \mathbf{b}_0 + \mathbf{k}_1 - 1}{\mathbf{s}_0 - \mathbf{b}_0} \binom{\mathbf{b}_0}{\mathbf{b}_0 - i} \right. \\ & \left. \sum_{\mathbf{b}_1} \binom{\mathbf{s}_1 - \mathbf{b}_1 + \mathbf{k}_0 - 1}{\mathbf{s}_1 - \mathbf{b}_1} \binom{\mathbf{b}_1}{\mathbf{b}_1 - i} \right). \end{aligned}$$

The multiset variant of Vandermonde's identity, (2.5), eliminates the sums over $\mathbf{b}$ giving

$$\sum_{i \geq 0} \binom{\mathbf{s}_0 + \mathbf{k}_1}{\mathbf{s}_0 - i} \binom{\mathbf{s}_1 + \mathbf{k}_0}{\mathbf{s}_1 - i}.$$

Substituting $\mathbf{k} - \mathbf{s}$ for $\mathbf{k}$ yields the claimed result. ■

Finally, we will verify that the asymptotic value of this expression matches the result from Lemma 2. As there, we fix $\mathbf{p}$ and $\mathbf{r}$, let $n \rightarrow \infty$, and let $\mathbf{k} = n\mathbf{p}$. The $i = 0$ term of the sum is a degree s polynomial and all other

terms are of lower degree. Thus we have

$$\begin{aligned}
I_{2,n\mathbf{p},\mathbf{r}}^R &\sim \binom{n\mathbf{p}_1 + \mathbf{s}_0}{\mathbf{s}_0} \binom{n\mathbf{p}_0 + \mathbf{s}_1}{\mathbf{s}_1}, \\
&\sim \binom{n\mathbf{p}_1}{\mathbf{s}_0} \binom{n\mathbf{p}_0}{\mathbf{s}_1}, \\
&\sim \frac{(n\mathbf{p}_1)^{\mathbf{s}_0}}{\mathbf{s}_0!} \frac{(n\mathbf{p}_0)^{\mathbf{s}_1}}{\mathbf{s}_1!}, \\
&\sim \binom{n}{s} \binom{s}{\mathbf{s}_0} \mathbf{p}_1^{\mathbf{s}_0} \mathbf{p}_0^{\mathbf{s}_1}, \\
&= \binom{n}{s} |M(\mathbf{s})| (1 - \mathbf{p}_0)^{\mathbf{s}_0} (1 - \mathbf{p}_1)^{\mathbf{s}_1}.
\end{aligned}$$

2.4 Recovering Levenshtein's result

To finish this chapter, we will derive the value of $I_{q,n,s}$ from $I_{q,\mathbf{k},\mathbf{s}}^R$. For all $\mathbf{k} \in C(q, n)$

$$\begin{aligned}
I_{q,n,s} &= \sum_{\mathbf{s} \in C(q,s)} I_{q,\mathbf{k},\mathbf{s}}^R \\
&= \sum_{\mathbf{s} \in C(q,s)} \sum_{A, \mathbf{b}} |M(\mathbf{b})| \prod_{j \in [q]} \binom{\|A_{\bullet,j}\| + \mathbf{k}_j - 1}{\|A_{\bullet,j}\|} |M(A_{\bullet,j})|.
\end{aligned}$$

The inner sum ranges over the matrices $A \in \mathbb{N}^{q \times q}$ and the vectors $\mathbf{b} \in \mathbb{N}^q$ satisfying $A\mathbf{1} + \mathbf{b} = \mathbf{s}$ and $A_{j,j} = 0$ for all j . The outer sum ranges over $\mathbf{s} \in \mathbb{N}^q$ such that $\mathbf{1}^T \mathbf{s} = s$. If we consider both the outer and inner sums together, we can eliminate $\mathbf{s}$ and simply require that A and $\mathbf{b}$ satisfy $\mathbf{1}^T A \mathbf{1} = a$ and $\mathbf{1}^T \mathbf{b} = s - a$.

Let $C_{i,j} = A_{(i+j \bmod q),j}$, so each column of C is a rotation of the corresponding column of A . This operation preserves the column sums but potentially changes the row sums. Thus $\mathbf{1}^T C = \mathbf{1}^T A$ and $\|C_{\bullet,j}\| = \|A_{\bullet,j}\|$, but $C\mathbf{1}$ might differ from $A\mathbf{1}$. For each j , $|M(C_{\bullet,j})| = |M(A_{\bullet,j})|$ and $C_{0,j} = 0$ because $A_{j,j} = 0$.

$$\begin{aligned}
I_{q,n,s} &= \sum_a \sum_{\mathbf{b} \in C(q,s-a)} |M(\mathbf{b})| \sum_C \prod_{j \in [q]} \binom{\|C_{\bullet,j}\| + \mathbf{k}_j - 1}{\|C_{\bullet,j}\|} |M(C_{\bullet,j})| \\
&= \sum_a \sum_{\mathbf{b} \in C(q,s-a)} |M(\mathbf{b})| \sum_{C\mathbf{1} \in C(q,a)} \prod_{j \in [q]} \binom{\|C_{\bullet,j}\| + \mathbf{k}_j - 1}{\|C_{\bullet,j}\|} \sum_{C_{\bullet,j}} |M(C_{\bullet,j})| \\
&= \sum_a q^{s-a} \sum_{C\mathbf{1} \in C(q,a)} \prod_{j \in [q]} \binom{\|C_{\bullet,j}\| + \mathbf{k}_j - 1}{\|C_{\bullet,j}\|} (q-1)^{\|C_{\bullet,j}\|} \\
&= \sum_a q^{s-a} \sum_{C\mathbf{1} \in C(q,a)} (q-1)^a \prod_{j \in [q]} \binom{\|C_{\bullet,j}\| + \mathbf{k}_j - 1}{\|C_{\bullet,j}\|} \\
&\stackrel{(a)}{=} \sum_a q^{s-a} (q-1)^a \binom{a+n-1}{a} \\
&\stackrel{(b)}{=} \sum_a (q-1)^a \binom{a+n-1}{a} \sum_{i \geq 0} \binom{s-a}{s-a-i} (q-1)^{s-a-i} \\
&\stackrel{(c)}{=} \sum_{i \geq 0} \binom{s+n}{s-i} (q-1)^{s-i}
\end{aligned}$$

In (a), we use Vandermonde's identity for compositions to eliminate the sum over $C\mathbf{1}$ and the fact that $\|\mathbf{k}\| = n$. In (b), we expand q^{s-a} using the binomial theorem. In (c), we use Vandermonde's identity for compositions one final time to eliminate the sum over a . As expected, our final result matches (1.3).

CHAPTER 3

APPLICATIONS TO CODES

3.1 Bounds on the size of constant-composition codes

We used the total number of superstrings of a string to upper bound the average degree of $L_{q,n,s}$, and we will use the number of superstrings of a given composition to upper bound the average degree of $L_{q,\mathbf{k},s}^R$. Turán's theorem will translate this into a lower bound on the size of independent sets in $L_{q,\mathbf{k},s}^R$.

Lemma 4. *Fix $\mathbf{p} \in \mathbb{Q}^q$ such that $\mathbf{p} \geq \mathbf{0}$ and $\|\mathbf{p}\| = 1$ and let $\mathbf{k} = n\mathbf{p} \in \mathbb{N}^q$. Then for fixed s , the average degree of the composition- $\mathbf{k}$ deletion graph satisfies*

$$\bar{d}(L_{q,\mathbf{k},s}^R) \lesssim \binom{n}{s}^2 \sum_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|^2 \prod_{i \in [q]} (\mathbf{p}_i (1 - \mathbf{p}_i)^2)^{\mathbf{s}_i}.$$

Proof: There are $|M(\mathbf{k} - \mathbf{s})| \binom{I_{q,\mathbf{k},s}^R}{2}$ triples $(z, (x, y)) \in M(\mathbf{k} - \mathbf{s}) \times \binom{M(\mathbf{k})}{2}$ such that $z \leq x$ and $z \leq y$. If $x, y \in M(\mathbf{k})$ are adjacent in $L_{q,\mathbf{k},s}^R$, then they have at least one common substring of composition $\mathbf{k} - \mathbf{s}$ where $\mathbf{s} \in C(q, s)$. This common substring appears with x and y in at least one triple.

Thus every edge in $L_{q,\mathbf{k},s}$ is counted at least once in the sum in

$$\frac{2|E(L_{q,\mathbf{k},s})|}{|V(L_{q,\mathbf{k},s})|} \leq \frac{2}{|M(\mathbf{k})|} \sum_{\mathbf{s} \in C(q,s)} |M(\mathbf{k} - \mathbf{s})| \binom{I_{q,\mathbf{k}-\mathbf{s},s}^R}{2}. \quad (3.1)$$

The ratio of multinomial coefficients simplifies asymptotically to

$$\begin{aligned} \frac{|M(\mathbf{k} - \mathbf{s})|}{|M(\mathbf{k})|} &= \frac{(n-s)!}{n!} \prod_{i \in [q]} \frac{\mathbf{k}_i!}{(\mathbf{k}_i - \mathbf{s}_i)!} \\ &\sim \frac{1}{n^s} \prod_{i \in [q]} \mathbf{k}_i^{\mathbf{s}_i} = \prod_{i \in [q]} \mathbf{p}_i^{\mathbf{s}_i}. \end{aligned} \quad (3.2)$$

Substituting Lemma 2 and (3.2) into (3.1) gives an asymptotic upper bound on $\bar{d}(L_{q,\mathbf{k},s}^R)$ of

$$\begin{aligned} & \sum_{\mathbf{s} \in C(q,s)} \left(\prod_{i \in [q]} \mathbf{p}_i^{\mathbf{s}_i} \right) \left(\binom{n-s}{s} |M(\mathbf{s})| \prod_{i \in [q]} (1 - \mathbf{p}_i)^{\mathbf{s}_i} \right)^2 \\ & \sim \binom{n}{s}^2 \sum_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|^2 \prod_{i \in [q]} (\mathbf{p}_i (1 - \mathbf{p}_i)^2)^{\mathbf{s}_i}. \end{aligned}$$

■

We now come to our first main result of this section.

Theorem 3. Fix $\mathbf{p} \in \mathbb{Q}^q$ such that $\mathbf{p} \geq \mathbf{0}$ and $\|\mathbf{p}\| = 1$ and let $\mathbf{k} = n\mathbf{p} \in \mathbb{N}^q$. For fixed s ,

$$\alpha(L_{q,\mathbf{k},s}^R) \gtrsim \frac{|M(\mathbf{k})|}{\binom{n}{s}^2 \sum_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|^2 \prod_{i \in [q]} (\mathbf{p}_i (1 - \mathbf{p}_i)^2)^{\mathbf{s}_i}}.$$

Proof: This follows immediately from Lemma 4 and Turán's theorem (1.1). ■

Now we will work toward an upper bound on code size. In the following lemma we derive a bound on the number of substrings of a particular composition. Recall that for a string x , $r(x)_i$ is the number of runs of the i -symbol in x .

Lemma 5. For $\mathbf{k}, \mathbf{s} \in \mathbb{N}^q$ and $x \in M(\mathbf{k})$, $|D_{\mathbf{s}}^R(x)| \geq \prod_{i \in [q]} \binom{r(x)_i + \mathbf{s}_i - \|\mathbf{s}\|}{\mathbf{s}_i}$.

Proof: We delete at most one symbol from each run and ensure that no two adjacent runs both have deletions. For the purpose of counting, it is easier to replace this second condition with a more restrictive one: after the deletion of an i -symbol, for each symbol other than i , deletions are forbidden in the next run of that symbol. In other words, each time we delete a symbol from a run, we forbid deletions in up to $q - 1$ other runs, one of which is the next run. There are $r(x)_i$ places where an i -symbol could be deleted. Of these, there will be at most $\|\mathbf{s}\| - \mathbf{s}_i$ places where deletion is forbidden. Thus there are at least $\binom{r(x)_i + \mathbf{s}_i - \|\mathbf{s}\|}{\mathbf{s}_i}$ deletion patterns for i -symbols. ■

Lemma 6. There are $|M(\mathbf{k})| \binom{\|\mathbf{k}\| - \mathbf{k}_0 + 1}{\mathbf{r}_0} \binom{\mathbf{k}_0 - 1}{\mathbf{k}_0 - \mathbf{r}_0} \binom{\|\mathbf{k}\|}{\mathbf{k}_0}^{-1}$ strings in $M(\mathbf{k})$ with exactly $\mathbf{r}_0$ runs of the zero-symbol.

Proof: There are $\binom{\mathbf{k}_0-1}{\mathbf{k}_0-\mathbf{r}_0}$ ways to distribute the zero-symbols among the $\mathbf{r}_0$ runs. Other symbols can appear at the start of the string, the end, and in the $\mathbf{r}_0 - 1$ gaps between runs of zero-symbols. Thus there are $\mathbf{r}_0 + 1$ places where the other symbols can appear and $\mathbf{r}_0 - 1$ of them contain at least one symbol each, so there are $\binom{\|\mathbf{k}\|-\mathbf{k}_0-(\mathbf{r}_0-1)+\mathbf{r}_0}{\mathbf{r}_0}$ ways to distribute them. Finally, there are $|M(0, \mathbf{k}_1, \dots, \mathbf{k}_{q-1})| = |M(\mathbf{k})| \binom{\|\mathbf{k}\|}{\mathbf{k}_0}^{-1}$ possible values for the substring induced by the nonzero symbols. ■

Suppose that we select an m -element subset $S \subseteq [N]$ uniformly at random. Then $|S \cup [n]|$ is given by a hypergeometric distribution with parameters m, n , and N and

$$P[|S \cup [n]| = l] = \frac{\binom{n}{l} \binom{N-n}{m-l}}{\binom{N}{m}}.$$

Thus the number of runs of zero-symbols follows a hypergeometric distribution with parameters $\mathbf{k}_0$, $\|\mathbf{k}\| - \mathbf{k}_0 + 1$, and $\|\mathbf{k}\|$. The mean of this distribution is $\mathbf{k}_0(\|\mathbf{k}\| - \mathbf{k}_0 + 1)/\|\mathbf{k}\|$.

Theorem 4. Fix $\mathbf{p} \in \mathbb{Q}^q$ such that $\mathbf{p} \geq \mathbf{0}$ and $\|\mathbf{p}\| = 1$ and let $\mathbf{k} = n\mathbf{p} \in \mathbb{N}^q$. For fixed s ,

$$\alpha(L_{q,\mathbf{k},s}^R) \lesssim \frac{|M(\mathbf{k})|}{\binom{n}{s} \max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})| \prod_{j \in [q]} (1 - \mathbf{p}_j)^{\mathbf{s}_j}}.$$

Proof: We will use the degree threshold upper bound [12], the same technique that Levenshtein applied to $L_{q,n,s}$ [1]. This method partitions the space of channel inputs into typical and atypical inputs and bounds the number of codewords of each class separately.

From Lemma 6, the number of i -runs in a uniformly selected member of $M(\mathbf{k})$ follows a hypergeometric distribution. Call $x \in M(\mathbf{k})$ an atypical string if for some i ,

$$r(x)_i \leq \frac{\mathbf{k}_i(\|\mathbf{k}\| - \mathbf{k}_i + 1)}{\|\mathbf{k}\|} - \epsilon_i.$$

By a large-deviations bound for the hypergeometric distribution, for each i , the fraction of strings for which this inequality holds is at most $e^{-2\epsilon_i^2/\mathbf{k}_i}$ [13].

Let $\epsilon_i = \sqrt{s\mathbf{k}_i \log \|\mathbf{k}\|}$ so $e^{-2\epsilon_i^2/\mathbf{k}_i} = \|\mathbf{k}\|^{-2s} = n^{-2s}$. Then there are at most $|M(\mathbf{k})|qn^{-2s}$ atypical strings. At worst, there is some code which includes all

of these atypical strings.

Now we will bound the number of typical strings in any code. If $x, y \in M(\mathbf{k})$ appear together in a code, then $D_{\mathbf{s}}^R(x)$ and $D_{\mathbf{s}}^R(y)$, which are subsets of $M(\mathbf{k} - \mathbf{s})$, must be disjoint. If $x \in M(\mathbf{k})$ is typical then

$$r(x)_i > \frac{\mathbf{k}_i(\|\mathbf{k}\| - \mathbf{k}_i + 1)}{\|\mathbf{k}\|} - \sqrt{s\mathbf{k}_i \log \|\mathbf{k}\|} \sim n\mathbf{p}_i(1 - \mathbf{p}_i).$$

Fix some $\mathbf{s} \in C(q, s)$. From Lemma 5,

$$|D_{\mathbf{s}}^R(x)| \geq \prod_{i \in [q]} \binom{r(x)_i + \mathbf{s}_i - \|\mathbf{s}\|}{\mathbf{s}_i} \gtrsim \prod_{i \in [q]} \binom{n\mathbf{p}_i(1 - \mathbf{p}_i)}{\mathbf{s}_i}.$$

Thus the number of typical strings in any code is asymptotically at most

$$\begin{aligned} & \frac{|M(\mathbf{k} - \mathbf{s})|}{\prod_{i \in [q]} \binom{n\mathbf{p}_i(1 - \mathbf{p}_i)}{\mathbf{s}_i}} \\ & \sim \frac{|M(\mathbf{k} - \mathbf{s})|}{\prod_{i \in [q]} \frac{1}{\mathbf{s}_i!} (n\mathbf{p}_i(1 - \mathbf{p}_i))^{\mathbf{s}_i}} \\ & \sim \frac{|M(\mathbf{k} - \mathbf{s})|}{\binom{n}{s} |M(\mathbf{s})| \prod_{i \in [q]} (\mathbf{p}_i(1 - \mathbf{p}_i))^{\mathbf{s}_i}} \\ & \sim \frac{|M(\mathbf{k})|}{\binom{n}{s} |M(\mathbf{s})| \prod_{i \in [q]} (1 - \mathbf{p}_i)^{\mathbf{s}_i}}, \end{aligned}$$

where the last asymptotic equality follows from (3.2). The bound on number of typical vertices in a code is $O(|M(\mathbf{k})|n^{-s})$, so the contribution of the atypical vertices does not matter asymptotically. This argument produces a bound for each $\mathbf{s} \in C(q, s)$ and we use the best one. $\blacksquare$

3.2 Construction of binary constant weight single deletion correcting codes

In this section, we focus on the binary single-deletion case ($q = 2, s = 1$). We present an explicit construction of independent sets in the graphs $L_{2,\mathbf{k},1}^R$. This construction is closely related to the Varshamov-Tenengolts (VT) codes in $L_{2,n,1}$. Like the VT codes, our codes are asymptotically optimal.

A k -coloring of a graph assigns a color (an element of $[k]$) to each vertex. The coloring is proper if it never assigns the same color to both endpoints

of an edge. Thus a proper coloring of a graph partitions its vertices into independent sets; each independent set is assigned a single color and called a *color class*. The chromatic number of a graph G , denoted $\chi(G)$, is the smallest k for which a proper k -coloring of G exists.

In a proper coloring, at least one color class must be at least as large as the average size of the color classes. Consequently, $\alpha(G) \geq |V(G)|/\chi(G)$. However, properly coloring a graph using the minimum number of colors is not equivalent to finding the largest independent set. In general there is no guarantee that the largest color class in a particular coloring is a maximum independent set or even that any optimal coloring has a maximum independent set as a color class.

A *clique* in a graph is a set of vertices that are all adjacent. The size of a largest clique in a graph G is denoted by $\omega(G)$. In a proper coloring, each vertex in a clique must be assigned a distinct color, so for any graph G , $\chi(G) \geq \omega(G)$. A *clique cover* of G is a subset of the cliques of G such that each vertex is contained in at least one clique from the subset. In a *fractional clique cover* each clique is assigned a fractional weight. For each vertex, the total weight assigned to the cliques containing that vertex must be at least one.

3.3 An explicit coloring

The VT construction uses a weight function to partition $[2]^n$ into $n + 1$ sets. Levenshtein showed that each of these sets is a single-deletion correcting code [1], so each is an independent set in $L_{2,n,1}$. This makes the VT weight a proper coloring of $L_{2,n,1}$ using $n + 1$ colors. Because $L_{2,\mathbf{k},1}^R$ is an induced subgraph of $L_{2,n,1}$ and $\chi(L_{2,n,1}) \leq n + 1$, for any $\mathbf{k} \in C(2, n)$, $\chi(L_{2,\mathbf{k},1}^R) \leq n + 1$. Tenengolts constructed non-binary single deletion correcting codes [5]. He does not use graph theoretic terminology, but implicit in his construction is a proper coloring of $L_{q,\mathbf{k},1}^R$ using $\|\mathbf{k}\|$ colors for each q and $\mathbf{k}$.

In this section, we exhibit a proper coloring of $L_{2,\mathbf{k},1}$ using $\max(\mathbf{k}_0, \mathbf{k}_1) + 1$ colors. Both the VT coloring of $L_{2,n,1}$ and our colorings of $L_{2,\mathbf{k},1}^R$ are based on the following weight function.

Definition 6. For any $x \in [2]^n$, let $w(x) = \sum_{i=0}^{n-1} (i + 1)x_i$. Call $w(x) \bmod n + 1$ the VT weight. Let $f_{\mathbf{k}}(x) = w(x) \bmod (\max(\mathbf{k}_0, \mathbf{k}_1) + 1)$. We call $f_{\mathbf{k}}$

the modified VT weight.

Lemma 7. *The modified VT weight $f_{\mathbf{k}}$ is a proper coloring of $L_{2,\mathbf{k},1}^R$.*

Proof: Let x and y be adjacent vertices in $L_{2,n,1}$. Then x and y have a common substring z of length $n - 1$. There are $a, b \in [n]$, such that z is produced by deleting x_a from x or by deleting y_b from y . Because x and y have the same composition, $x_a = y_b$ and $a \neq b$. Without loss of generality, let $a < b$. Then for $a \leq i \leq b - 1$, $x_{i+1} = z_i = y_i$, and for $0 \leq i < a$ and $b < i \leq n - 1$, $x_i = y_i$. Then

$$\begin{aligned} w(x) - w(y) &= \sum_{i=0}^{n-1} (i+1)(x_i - y_i) \\ &= ax_a - by_b + \sum_{i=a}^{b-1} (i+2)x_{i+1} - (i+1)y_i \\ &= (a-b)y_b + \sum_{i=a}^{b-1} y_i. \end{aligned}$$

Let $l = \sum_{i=a}^{b-1} y_i$. Then l , which is the number of ones between y_a and y_{b-1} , is at most $\mathbf{k}_1$ and $b - a - l$, which the number of zeros between y_a and y_{b-1} , is at most $\mathbf{k}_0$. If $y_b = 0$, then $x_a = 1$ and $w(x) - w(y) = l > 0$. If $y_b = 1$, then $x_a = 0$ and $w(y) - w(x) = b - a - l > 0$. In both cases, $0 < |w(y) - w(x)| \leq \max(\mathbf{k}_0, \mathbf{k}_1)$, so $(w(x) - w(y)) \bmod (\max(\mathbf{k}_0, \mathbf{k}_1) + 1) \neq 0$ and $f_{\mathbf{k}}(x) \neq f_{\mathbf{k}}(y)$. ■

Corollary 1. *For all $\mathbf{k} \in \mathbb{N}^2$, $\alpha(L_{2,\mathbf{k},1}^R) \geq |M(\mathbf{k})| / \max(\mathbf{k}_0 + 1, \mathbf{k}_1 + 1)$.*

3.4 Exact sizes of the codes in $L_{q,\mathbf{k},s}^R$

In some cases, it is not too hard to determine the exact sizes of the color classes of $f_{\mathbf{k}}$.

Lemma 8. *Let $\mathbf{k} \in C(2, n)$ such that $\mathbf{k}_0 \leq \mathbf{k}_1$ and $\mathbf{k}_0$ is relatively prime to $\mathbf{k}_1 + 1$. In the coloring of $L_{q,\mathbf{k},s}^R$ by $f_{\mathbf{k}}$, each color class contains exactly $\frac{|M(\mathbf{k})|}{\mathbf{k}_1 + 1}$ vertices.*

Proof: In each $x \in M(\mathbf{k})$, the ones partition the zeros into $\mathbf{k}_1 + 1$ possibly empty blocks. This gives a bijection between $M(\mathbf{k})$ and $C(\mathbf{k}_1 + 1, \mathbf{k}_0)$.

There is a simple operation that partitions $C(\mathbf{k}_1 + 1, \mathbf{k}_0)$ into orbits of $\mathbf{k}_1 + 1$ elements: for each $y \in C(\mathbf{k}_1 + 1, \mathbf{k}_0)$, cyclically shift the first entry of y to the end. This is equivalent to deleting all of the initial zeros and the first one of the corresponding $x \in M(\mathbf{k})$, then adding a one followed by the deleted zeros to the end of x . This has a fairly simple effect on $w(x)$. Let i be the first entry of y or equivalently the number of initial zeros of x . The operation moves the leftmost one $n - 1 - 2i$ locations to the right and moves the other $\mathbf{k}_1 - 1$ ones $i + 1$ locations to the left. This increases $w(x)$ by

$$n - 1 - 2i - (\mathbf{k}_1 - 1)(i + 1) = n + 1 - (\mathbf{k}_1 + 1)(i + 1) = \mathbf{k}_0 - (\mathbf{k}_1 + 1)i.$$

Because $\mathbf{k}_0 \leq \mathbf{k}_1$, $f_{\mathbf{k}}(x) = w(x) \bmod (\mathbf{k}_1 + 1)$. Thus $f_{\mathbf{k}}(x)$ is increased by $\mathbf{k}_0$. Because $\mathbf{k}_0$ is relatively prime to $\mathbf{k}_1 + 1$, the orbits of $M(\mathbf{k})$ under this operation each contain $\mathbf{k}_1 + 1$ strings, one of each color class. ■

If n is even, then $\mathbf{k} = \frac{n}{2}\mathbf{1}$ meets the conditions of the Lemma 8. If $n + 1$ is prime, for every $\mathbf{k} \in C(2, n)$ $L_{2, \mathbf{k}, 1}^R$ is partitioned into equally sized color classes.

This argument can be generalized to calculate the size of the color classes in all case. The generalization uses Möbius inversion over the lattice of positive integers ordered by divisibility.

3.5 Cliques and optimal colorings

Both the VT coloring of $L_{2, n, 1}$ and our coloring of $L_{2, \mathbf{k}, 1}^R$ are optimal. In both cases, to show this we will exhibit cliques of matching size and use the inequality $\omega(G) \leq \chi(G)$. The following lemma constructs these cliques.

Lemma 9. *For $n \geq s$, the graph $L_{q, n, s}$ contains cliques of $I_{q, n, s}$ vertices. For each $\mathbf{s} \in C(q, s)$ such that $\mathbf{k} \geq \mathbf{s}$, $L_{q, \mathbf{k}, s}^R$ contains cliques of $I_{q, \mathbf{k}, s}^R$ vertices.*

Proof: By (1.3), each string in $[q]^{n-s}$ has $I_{q, n, s}$ superstrings in $[q]^n$. These are all adjacent in $L_{q, n, s}$, so they form a clique. By Lemma 2, each string in $M(\mathbf{k} - \mathbf{s})$ has $I_{q, \mathbf{k}, s}^R$ superstrings in $M(\mathbf{k})$. These are all adjacent in $L_{q, \mathbf{k}, s}^R$, so they form a clique. ■

The optimality of both colorings follows immediately.

Theorem 5. *For all n , the VT coloring of $L_{2,n,1}$ is optimal and*

$$\chi(L_{2,n,1}) = \omega(L_{2,n,1}) = n + 1.$$

For all n and $\mathbf{k} \in C(2, n)$ such that $\mathbf{k} \geq \mathbf{1}$, the coloring of $L_{2,\mathbf{k},1}^R$ by the modified VT weight $f_{\mathbf{k}}$ is optimal and

$$\chi(L_{2,\mathbf{k},1}^R) = \omega(L_{2,\mathbf{k},1}^R) = \max(\mathbf{k}_0, \mathbf{k}_1) + 1.$$

Proof: By Lemma 9, $L_{2,n,1}$ contains cliques of $I_{2,n-1,1} = n + 1$ vertices. The VT coloring uses $n + 1$ colors.

By Lemma 9, $L_{2,\mathbf{k},1}^R$ contains cliques of sizes $I_{2,(\mathbf{k}_0-1,\mathbf{k}_1),(1,0)} = \mathbf{k}_1 + 1$ and $I_{2,(\mathbf{k}_0,\mathbf{k}_1-1),(0,1)} = \mathbf{k}_0 + 1$. From Lemma 7 we have $\chi(L_{2,\mathbf{k},1}^R) \leq \max(\mathbf{k}_0, \mathbf{k}_1) + 1$. ■

We can show that the Tenengolts coloring of $L_{q,\mathbf{k},1}^R$ is optimal for a small set of parameters, but not in general. For $\mathbf{k} \in C(q, n)$ such that $\mathbf{k} \geq \mathbf{1}$,

$$\omega(L_{q,\mathbf{k},1}^R) \geq n + 1 - \min_{i \in [q]} \mathbf{k}_i.$$

If $\mathbf{k}_i = 1$ for some i , then the n -coloring is optimal. Otherwise there is a gap between our lower bound on the clique number and upper bound on the coloring number. The biggest gap occurs when $\mathbf{k} = \frac{n}{q}\mathbf{1}$, which makes the ratio between these bounds $\frac{qn}{(q-1)n+q} \sim \frac{q}{q-1}$.

3.6 Fractional clique covers and optimal codes

There is a gap between Levenshtein's upper and lower bounds for all numbers of deletions. However, for a single deletion, the VT construction asymptotically matches the upper bound and closes the gap. The largest VT code always contains at least $\frac{2^n}{n+1}$ codewords. This matches the asymptotic upper bound (1.2), which states that $\alpha(L_{2,n,1}) \lesssim \frac{2^n}{n}$. Kulkarni and Kiyavash [14] showed that

$$\alpha(L_{2,n,1}) \leq \frac{2^n}{n-1}. \tag{3.3}$$

Thus for length n , the ratio of the size of the largest VT code to the size of the optimal single deletion correcting code is at least $1 - \frac{2}{n+1}$.

From Theorem 4, we know that $\alpha(L_{q,\mathbf{k},1}^R) \lesssim \frac{|M(\mathbf{k})|}{\|\mathbf{k}\| - \min_i \mathbf{k}_i}$. Thus the modified VT weight produces asymptotically optimal constant composition binary codes. Now we present a non-asymptotic bound.

Lemma 10. *Let $\mathbf{k} \in \mathbb{N}^q$. If $\mathbf{k}_i = 1$, then*

$$\alpha(L_{q,\mathbf{k},1}^R) \leq \frac{|M(\mathbf{k})|}{\|\mathbf{k}\|}.$$

If $\mathbf{k}_i \geq 2$, then

$$\alpha(L_{q,\mathbf{k},1}^R) \leq \frac{|M(\mathbf{k})|\mathbf{k}_i}{(\|\mathbf{k}\| - \mathbf{k}_i + 1)(\mathbf{k}_i - 1)}.$$

Proof: Let $\mathbf{s}_i = 1$ and $s_j = 0$ for $j \neq i$. For each $x \in M(\mathbf{k} - \mathbf{s})$, the set $I_{\mathbf{s}}^R(x)$ forms a clique in $M(\mathbf{k})$. To find an upper bound on $\alpha(L_{q,\mathbf{k},1}^R)$, we will construct a fractional clique cover using this family of cliques.

If $\mathbf{k}_i = 1$, then each $y \in M(\mathbf{k})$ is a superstring of exactly one $x \in M(\mathbf{k} - \mathbf{s})$. We assign weight one to each $x \in M(\mathbf{k} - \mathbf{s})$ and get a cover with total weight $|M(\mathbf{k} - \mathbf{s})| = \frac{|M(\mathbf{k})|}{\|\mathbf{k}\|}$.

For $\mathbf{k}_i \geq 2$, we construct a slightly more complicated cover. Each $y \in M(\mathbf{k})$ has $r(y)_i$ substrings in $M(\mathbf{k} - \mathbf{s})$ where $r(y)_i$ is the number of runs of zeros in y . Thus it is contained in $r(y)_i$ cliques in our family. If each of these cliques is assigned weight at least $1/r(y)_i$, y is covered.

If y is a superstring of x , then $r(y)_i \geq r(x)_i$. We assign each $x \in M(\mathbf{k} - \mathbf{s})$ a weight of $1/r(x)_i$ so every $y \in M(\mathbf{k})$ is covered. From Lemma 6, there are $|M(\mathbf{k} - \mathbf{s})| \binom{\|\mathbf{k}\| - \mathbf{k}_i + 1}{\mathbf{r}_i} \binom{\mathbf{k}_i - 2}{\mathbf{k}_i - 1 - \mathbf{r}_i} \binom{\|\mathbf{k}\| - 1}{\mathbf{k}_i - 1}^{-1}$ strings in $M(\mathbf{k} - \mathbf{s})$ with $\mathbf{r}_i$ runs of zeros.

Because $|M(\mathbf{k} - \mathbf{s})| \binom{\|\mathbf{k}\| - 1}{\mathbf{k}_i - 1}^{-1} = |M(\mathbf{k})| \binom{\|\mathbf{k}\|}{\mathbf{k}_i}^{-1}$, the total weight of the clique cover is

$$\begin{aligned} & \sum_{\mathbf{r}_i=1}^{\mathbf{k}_i-1} \frac{1}{\mathbf{r}_i} |M(\mathbf{k})| \binom{\|\mathbf{k}\| - \mathbf{k}_i + 1}{\mathbf{r}_i} \binom{\mathbf{k}_i - 2}{\mathbf{k}_i - 1 - \mathbf{r}_i} \binom{\|\mathbf{k}\|}{\mathbf{k}_i}^{-1} \\ &= \frac{|M(\mathbf{k})|}{\mathbf{k}_i - 1} \sum_{\mathbf{r}_i=1}^{\mathbf{k}_i-1} \binom{\|\mathbf{k}\| - \mathbf{k}_i + 1}{\mathbf{r}_i} \binom{\mathbf{k}_i - 1}{\mathbf{k}_i - 1 - \mathbf{r}_i} \binom{\|\mathbf{k}\|}{\mathbf{k}_i}^{-1} \\ &\stackrel{(a)}{=} \frac{|M(\mathbf{k})|}{\mathbf{k}_i - 1} \left(\binom{\|\mathbf{k}\|}{\mathbf{k}_i - 1} - 1 \right) \binom{\|\mathbf{k}\|}{\mathbf{k}_i}^{-1} \\ &< \frac{|M(\mathbf{k})|}{\mathbf{k}_i - 1} \frac{\mathbf{k}_i}{\|\mathbf{k}\| - \mathbf{k}_i + 1}. \end{aligned}$$

In (a), we use Vandermonde's identity to eliminate the sum. The $\mathbf{r}_i = 0$

term, which would equal one, was missing from the sum. ■

Combining Corollary 1 and Lemma 10, we can see that for $\mathbf{k} \in \mathbb{N}^2$ such that $2 \leq \mathbf{k}_0 \leq \mathbf{k}_1$,

$$\frac{|M(\mathbf{k})|}{\mathbf{k}_1 + 1} \leq \alpha(L_{2,\mathbf{k},1}) \leq \frac{|M(\mathbf{k})|}{\mathbf{k}_1 + 1} \left(1 + \frac{1}{\mathbf{k}_0 - 1}\right).$$

The single-deletion correcting codes produced by the modified VT weight are nearly optimal.

3.7 Code construction by composition partitioning

We now describe a strategy for code construction for any number of deletions. This strategy is inspired by a simple bound on deletion distance.

Lemma 11. *For all $x \in M(\mathbf{a})$ and $y \in M(\mathbf{b})$, the deletion distance between them satisfies the lower bound $d_L(x, y) \geq \|\mathbf{a} - \mathbf{b}\|$.*

Proof: Let $z \in M(\mathbf{c})$ be a longest common substring of x and y . Then z has at most as many of the symbol i than either x or y , so

$$\mathbf{c}_i \leq \min(\mathbf{a}_i, \mathbf{b}_i) = (\mathbf{a}_i + \mathbf{b}_i - |\mathbf{a}_i - \mathbf{b}_i|)/2.$$

Summing over i gives

$$2\|\mathbf{c}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| - \|\mathbf{a} - \mathbf{b}\|.$$

Rearranging yields

$$\|\mathbf{a} - \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| - 2\|\mathbf{c}\| = d_L(x, y).$$

■

Definition 7. Let $C_{q,n,s}$ be a graph with vertex set $C(q, n)$. Vertices $x, y \in C(q, n)$ are adjacent if and only if $\|x - y\| \leq 2s$.

Lemma 12. For each $\mathbf{k} \in C(q, n)$, let $C_{\mathbf{k}} \subseteq M(\mathbf{k})$ be an independent set in $L_{q,\mathbf{k},s}^R$. If $A \subseteq C(q, n)$ is an independent set in $C_{q,n,s}$, then $\bigcup_{\mathbf{k} \in A} C_{\mathbf{k}}$ is an independent set in $L_{q,n,s}$.

Let $f : C(q, n) \rightarrow [m]$ be a proper m -coloring of $C_{q,n,s}$. For each $i \in [m]$, let $B_i = \{\mathbf{k} \in C(q, n) : f(\mathbf{k}) = i\}$. Then there is some $i \in [m]$ such that

$$\left| \bigcup_{\mathbf{k} \in B_i} C_{\mathbf{k}} \right| \geq \frac{1}{m} \left| \bigcup_{\mathbf{k} \in C(q, n)} C_{\mathbf{k}} \right|.$$

Proof: Let $x, y \in \bigcup_{\mathbf{k} \in B} C_{\mathbf{k}}$. If $\#(x) = \#(y)$, then $d_L(x, y) > 2s$ because $C_{\#(x)}$ is an independent set in $L_{q, \#(x), s}^R$. Otherwise, $d_L(x, y) \geq \|\#(x) - \#(y)\| > 2s$ because A is an independent set in $C_{q,n,s}$. Thus $\bigcup_{\mathbf{k} \in A} C_{\mathbf{k}}$ is an independent set in $L_{q,n,s}$.

The color classes of f are independent sets in $C_{q,n,s}$. Thus f composed with $\#$ partitions $\bigcup_{\mathbf{k} \in C(q, n)} C_{\mathbf{k}}$ into m classes, each of which is an independent set in $L_{q,n,s}$. One of these m classes must be at least as large as the average. ■

Lemma 13. *The function $f : C(q, n) \rightarrow [s+1]^q$, $f(x)_i = x_i \bmod (s+1)$, is a proper $(s+1)^{q-1}$ -coloring of $C_{q,n,s}$.*

Proof: Let $x, y \in C(q, n)$, such that $x \neq y$ and $f(x) = f(y)$. Then there are $i, j \in [q]$ such that $x_i > y_i$, $x_j < y_j$, and $i \neq j$. Thus $\|x - y\| \geq |x_i - y_i| + |x_j - y_j| \geq 2s + 2$.

If $y \in f(C(q, n))$, then $\sum_{i \in [q]} y_i = n \bmod (s+1)$. Thus $|f(C(q, n))| = (s+1)^{q-1}$. ■

Lemma 14. *For $s \leq n$, the graph $C_{q,n,s}$ has cliques of $\binom{s+q-1}{q-1}$ vertices.*

Proof: In the product partial order on $\mathbb{N}^q$, each vector in $C(q, n-s)$ is less than $|C(q, s)|$ vectors in $C(q, n)$. The maximum distance between these vectors is $2s$, so they form a clique in $C_{q,n,s}$. ■

Lemma 13 and Lemma 14 give upper and lower bounds on $\chi(C_{q,n,s})$. Note that $\binom{s+q-1}{q-1} \geq (s+1)^{q-1}/(q-1)!$. For $q = 2$, the bounds are equal and thus are tight. For larger q , both bounds can be improved. For fixed q both bounds are $O(s^{q-1})$, so these improvements only affect the multiplicative constant.

We can now use the upper bound on average degree to get a lower bound on code size.

Theorem 6. *For fixed s , the composition partitioning method produces s -deletion correcting codes that contain asymptotically at least*

$$\frac{q^{n+2s}}{(s+1)^{q-1}(q-1)^{2s} \binom{n}{s}^2 (\max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|)}$$

codewords.

Proof: From Lemma 13, there is an $(s+1)^{q-1}$ -coloring of $C_{q,n,s}$. Thus by Lemma 12, there is a code with at least

$$\alpha(L_{q,n,s}) \geq \frac{1}{(s+1)^{q-1}} \sum_{\mathbf{k} \in C(q,n)} \alpha(L_{q,\mathbf{k},s}^R)$$

codewords. From Theorem 3, $\alpha(L_{q,\mathbf{k},s}^R)$ is asymptotically at least

$$\alpha(L_{q,\mathbf{k},s}^R) \gtrsim \frac{|M(\mathbf{k})|}{\binom{n}{s}^2 \sum_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|^2 \prod_{i \in [q]} (\mathbf{p}_i(1 - \mathbf{p}_i)^2)^{\mathbf{s}_i}}.$$

The sum in the denominator can be bounded as follows:

$$\begin{aligned} & \sum_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|^2 \prod_{i \in [q]} (\mathbf{p}_i(1 - \mathbf{p}_i)^2)^{\mathbf{s}_i} \\ & \leq \left(\sum_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})| \prod_{i \in [q]} (\mathbf{p}_i(1 - \mathbf{p}_i)^2)^{\mathbf{s}_i} \right) \max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})| \\ & = \left(\sum_{i \in [q]} \mathbf{p}_i(1 - \mathbf{p}_i)^2 \right)^s \max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|. \end{aligned}$$

It is easy to verify that $\sum_{i \in [q]} \mathbf{p}_i(1 - \mathbf{p}_i)^2 \leq (q-1)^2/q^2$ and that the inequality is tight only at $\mathbf{p} = \frac{1}{q}\mathbf{1}$. Thus the composition partitioning method produces codes of size at least

$$\begin{aligned} & \sum_{\mathbf{k} \in C(q,n)} \frac{|M(\mathbf{k})|}{(s+1)^{q-1} \left(\frac{q-1}{q}\right)^{2s} \binom{n}{s}^2 (\max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|)} \\ & = \frac{q^{n+2s}}{(s+1)^{q-1}(q-1)^{2s} \binom{n}{s}^2 (\max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|)}. \end{aligned}$$

■

Corollary 2. *The ratio of the size of the codes promised by Levenshtein's lower bound to the size of the codes from Theorem 6 is at most $c_q(s+1)^{(q-1)/2}$, where c_q is some constant that depends only on q .*

Proof: Levenshtein's lower bound is

$$\alpha(L_{q,n,s}) \gtrsim \frac{q^{n+s}}{(q-1)^{2s} \binom{n}{s}^2}.$$

The ratio of Levenshtein's lower bound to the bound of Theorem 6 is

$$\frac{(s+1)^{q-1} \max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})|}{q^s}.$$

From Stirling's approximation [15],

$$\max_{\mathbf{s} \in C(q,s)} |M(\mathbf{s})| \leq \frac{e^{1/12} q^{s+q/2}}{(2\pi s)^{(q-1)/2}}.$$

Thus the ratio is at most

$$\frac{(s+1)^{q-1} e^{1/12} q^{q/2}}{(2\pi s)^{(q-1)/2}} \leq \frac{e^{1/12} q^{q/2}}{\pi^{(q-1)/2}} (s+1)^{(q-1)/2}.$$

■

3.8 Tenengolts q -ary single deletion correcting codes

The construction q -ary single deletion correcting codes by Tenengolts [5] is a special case of the composition partitioning technique. As we mentioned in Section 3.3, this construction provides a proper coloring of $L_{q,\mathbf{k},1}^R$ using $\|\mathbf{k}\|$ colors for each q and $\mathbf{k}$. This coloring implies that $\alpha(L_{q,\mathbf{k},1}^R) \geq |M(\mathbf{k})|/\|\mathbf{k}\|$.

The construction also uses the following very simple coloring of $C_{q,n,1}$.

Lemma 15. *The function $f : C(q,n) \rightarrow [q]$, $f(x) = (\sum_i i x_i) \bmod q$, is a proper q -coloring of $C_{q,n,1}$.*

Proof: Let $x, y \in C(q,n)$ be adjacent in $C_{q,n,1}$. Then $\|x - y\| = 2$ and there are $j, k \in [q]$ such that $x_j - y_j = 1$, $y_k - x_k = 1$, and $j \neq k$. Then $(f(x) - f(y)) \bmod q = \left(\sum_{i \in [q]} i(x_i - y_i) \right) \bmod q = (j - k) \bmod q \neq 0$. ■

This yields a code of size at least

$$\frac{1}{q} \sum_{\mathbf{k} \in C(q,n)} \frac{|M(\mathbf{k})|}{n} = \frac{q^n}{qn}.$$

3.9 Asymptotic optimal binary single deletion codes

We now show that taking the union of independent sets from $L_{2,\mathbf{k},1}^R$ produces an independent set in $L_{2,n,1}$ that is asymptotically of optimal size. Let $C_{\mathbf{k}}$ be a largest color class of the $f_{\mathbf{k}}$ coloring of $L_{2,\mathbf{k},1}^R$ described in Lemma 7. The graph $C_{2,n,1}$, which is just the path on $n+1$ vertices, has a proper 2-coloring. This is a special case of both the $(s+1)^{q-1}$ -coloring described in Lemma 13 and the q -coloring described in Lemma 15. The color classes of this coloring are $B_{n,i} = \{\mathbf{k} \in C(2,n) : \mathbf{k}_0 \bmod 2 = i\}$ for $i \in [2]$.

Theorem 7. *For all $n \in \mathbb{N}$ and $i \in [2]$, $\bigcup_{\mathbf{k} \in B_{n,i}} C_{\mathbf{k}}$ is an independent set in $L_{2,n,1}$ and $\left| \bigcup_{\mathbf{k} \in B_{n,i}} C_{\mathbf{k}} \right| \geq \frac{1}{n+1} (2^n - \binom{n}{k^*})$ where k^* is the integer closest to $n/2$ such that $k^* \bmod 2 \neq i$.*

Proof: From Lemma 12, $\bigcup_{\mathbf{k} \in B_{n,i}} C_{\mathbf{k}}$ is an independent set in $L_{2,n,1}$. In each graph $L_{2,\mathbf{k},1}^R$, some color class must be at least as large as the average, so from Lemma 7

$$|C_{\mathbf{k}}| \geq \frac{|V(L_{2,\mathbf{k},1}^R)|}{\chi(L_{2,\mathbf{k},1}^R)} = \binom{n}{\mathbf{k}_0} \frac{1}{\max(\mathbf{k}_0, \mathbf{k}_1) + 1}.$$

Thus

$$\begin{aligned} \sum_{\mathbf{k} \in B_{n,i}} |C_{\mathbf{k}}| &\geq \sum_{\substack{\mathbf{k} \in B_{n,i} \\ \mathbf{k}_0 < k^*}} \binom{n}{\mathbf{k}_0} \frac{1}{\mathbf{k}_1 + 1} + \sum_{\substack{\mathbf{k} \in B_{n,i} \\ \mathbf{k}_0 > k^*}} \binom{n}{\mathbf{k}_0} \frac{1}{\mathbf{k}_0 + 1} \\ &= \sum_{\substack{\mathbf{k} \in B_{n,i} \\ \mathbf{k}_0 < k^*}} \binom{n+1}{\mathbf{k}_0} \frac{1}{n+1} + \sum_{\substack{\mathbf{k} \in B_{n,i} \\ \mathbf{k}_0 > k^*}} \binom{n+1}{\mathbf{k}_0 + 1} \frac{1}{n+1} \\ &\stackrel{(a)}{=} \frac{1}{n+1} \left(\sum_{\mathbf{k}_0 < k^*} \binom{n}{\mathbf{k}_0} + \sum_{\mathbf{k}_0 > k^*} \binom{n}{\mathbf{k}_0} \right) \\ &= \frac{1}{n+1} \left(2^n - \binom{n}{k^*} \right), \end{aligned}$$

where (a) follows from $\binom{n+1}{\mathbf{k}_0} = \binom{n}{\mathbf{k}_0-1} + \binom{n}{\mathbf{k}_0}$. ■

Corollary 3. *For each n , the ratio of the size of the codes constructed in Theorem 7 to the size of optimal single deletion correcting binary codes is at least $\left(1 - \sqrt{\frac{2}{\pi n}}\right) \left(1 - \frac{2}{n+1}\right) = 1 - \Theta(n^{-1/2})$, so they are asymptotically optimal.*

Proof: By Stirling's formula [15], $\binom{n}{\lfloor n/2 \rfloor} \leq 2^n \sqrt{\frac{2}{\pi n}}$. The size of the codes from Theorem 7 is at least $\frac{2^n}{n+1} \left(1 - \sqrt{\frac{2}{\pi n}}\right)$. From (3.3) in Section 3.6, $\alpha(L_{2,n,1}) \leq \frac{2^n}{n-1} = \frac{2^n}{n+1} \left(1 - \frac{2}{n+1}\right)^{-1}$. ■

Most of the vertices of $L_{2,n,1}$ are in the subgraphs $L_{2,\mathbf{k},1}^R$ with $\mathbf{k} \approx \frac{n}{2}\mathbf{1}$. These central constant composition subgraphs are colored using just over half as many colors as in the VT coloring of $L_{2,n,1}$. This makes up for excluding half of the vertices based on composition.

In the non-binary case, we have

$$\frac{q^n}{qn} \leq \alpha(L_{q,n,1}) \leq \frac{q^n}{(q-1)(n-1)}.$$

The lower bound is from the size of the Tenengolts construction and the upper bound is due to Kulkarni and Kiyavash [14]. Thus the Tenengolts construction is within a factor of $\frac{q}{q-1}$ of asymptotic optimality. In Section 3.3, we observed that for $\mathbf{k} = \frac{n}{q}\mathbf{1}$,

$$\frac{q-1}{q}n + 1 \leq \omega(L_{q,\mathbf{k},1}^R) \leq \chi(L_{q,\mathbf{k},1}^R) \leq n.$$

Unlike the modified VT coloring, the Tenengolts coloring does not meet the limit imposed by cliques in the central constant composition subgraphs. The Tenengolts coloring cannot be improved for the subgraphs with highly unbalanced compositions. These subgraphs do not contain only a small fraction of the vertices, but they are limiting the overall performance of the code. Better codes in the central constant composition graphs could lead to asymptotically optimal q -ary single deletion correcting codes.

CHAPTER 4

LARGE CLIQUES

4.1 Large cliques in $L_{q,n,s}$ and $L_{q,\mathbf{k},s}^R$

In this section, we construct a large clique in the graph $L_{q,n,s}$. These cliques provide a lower bound on the chromatic number of graph. They present an indirect barrier to code construction.

In $L_{2,n,1}$, the VT coloring demonstrates that cliques produced by a single common substring are maximum. For $s \geq 2$, a more general construction produces larger cliques in $L_{q,n,s}$. For any string $x \in [q]^m$, consider the set of strings of length n within a deletion distance of s . By the triangle inequality, the deletion distance between any two of these strings is at most $2s$, so they form a clique in $L_{q,n,s}$. When $m = n - s$ this set is just $I_s(x)$, and when $m = n + s$ it is $D_s(x)$. Bigger cliques can be constructed by taking m to be an intermediate length.

Lemma 16. *The maximum clique size in $L_{q,n,s}$ satisfies*

$$\omega(L_{q,n,s}) \gtrsim \binom{n}{s} \max_a \binom{s}{a} (q-1)^a.$$

Proof: Let $a, k, l \in \mathbb{N}$ such that l is odd and $a \leq s \leq k$. Let $m = lk + 5(k-1)$ and $n = m + a - (s-a)$. We will construct a string $x \in [q]^m$ and a set $S \subset [q]^m$ such that $|S| = \binom{k}{s} \binom{s}{a} ((q-1)(l+1) + 1)^a l^{s-a}$ and for all $y \in S$, $d_L(x, y) \leq s$.

Let $u = 1010 \dots 1 \in [q]^l$ be the alternating string of zeros and ones that starts and ends with a one and has length l . Let $v = 01110 \in [q]^5$. Let x be the concatenation of k copies of u interspersed with $k-1$ copies of v , so $x = u : v : u : v : \dots : u$. Each member of S has the same basic structure as x , but with a copies of u replaced by a member of $I_1(u)$ and $s-a$ copies of u replaced by a member of $D_1(u)$. It follows immediately that for all $y \in S$,

$x :$	1	0	1	0	1	0	1	1	1	0	1	0	1	0	1	0	1	1	1	0	1	0	1	0	1
$y :$	1	0	1	0	0	1	1	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	0	1

Figure 4.1: Two of the strings constructed in the proof of Lemma 16: the center string x and a string $y \in S$. In y , the first copy of u has been replaced with a substring of u and the third copy has been replaced with a superstring. The parameters are $l = 5$, $k = 3$, $s = 2$, and $a = 1$. Note that the deletions and insertions have created runs of length two, but the only runs of length 3 are from the copies of $v = 01110$.

$d_L(x, y) \leq s$. By the triangle inequality, for all $y, z \in S$, $d_L(y, z) \leq 2s$. Thus the vertices in S form a clique in $L_{q,n,s}$. Figure 4.1 gives an example.

The members of $I_1(u)$ or $D_1(u)$ contain only runs of length one and two. Furthermore, it is impossible to produce a run of length three by concatenating a member of $I_1(u)$ or $D_1(u)$ with v . This property allows us to uniquely decompose any member of S into its components. Thus the construction of S does not produce any string more than once.

The string u has l substrings of length $l-1$ and $(q-1)(l+1)+1$ superstrings of length $l+1$. There are $\binom{k}{s}\binom{s}{a}$ ways to select which copies of u are replaced. Thus $|S| = \binom{k}{s}\binom{s}{a}((q-1)(l+1)+1)^a l^{s-a}$.

Taking $k \sim l \sim \sqrt{n}$ yields

$$\begin{aligned}
|S| &\sim \binom{\sqrt{n}}{s} \binom{s}{a} ((q-1)\sqrt{n})^a (\sqrt{n})^{s-a} \\
&\sim \frac{(\sqrt{n})^s}{s!} \binom{s}{a} (q-1)^a (\sqrt{n})^s \\
&\sim \binom{n}{s} \binom{s}{a} (q-1)^a.
\end{aligned}$$

To construct the largest possible clique, we optimize the choice of a . This gives the claimed bound. $\blacksquare$

If q divides s , then the size of $|S|$ is maximized when $a = \frac{q-1}{q}$. We can use Stirling's formula to estimate the size in this case.

Lemma 17. For $\alpha, \beta, n \in \mathbb{N}$,

$$\max_i \binom{(\alpha + \beta)n}{i} \left(\frac{\alpha}{\alpha + \beta} \right)^i \left(\frac{\beta}{\alpha + \beta} \right)^{(\alpha + \beta)n - i} \geq \frac{1}{3} \sqrt{\frac{\alpha + \beta}{\alpha \beta n}}.$$

Proof: One form of Stirling's approximation is [15]

$$1 \leq \frac{n!}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n} \leq e^{\frac{1}{12}}.$$

Then for $\alpha, \beta, n \in \mathbb{N}$, consider the binomial distribution produced by $(\alpha + \beta)n$ trials and success probability $\alpha/(\alpha + \beta)$. The most likely outcome is αn successes and the probability of that outcome is

$$\begin{aligned} & \max_i \binom{(\alpha + \beta)n}{i} \left(\frac{\alpha}{\alpha + \beta}\right)^i \left(\frac{\beta}{\alpha + \beta}\right)^{(\alpha + \beta)n - i} \\ &= \binom{(\alpha + \beta)n}{\alpha n} \left(\frac{\alpha}{\alpha + \beta}\right)^{\alpha n} \left(\frac{\beta}{\alpha + \beta}\right)^{\beta n} \\ &\geq \frac{\sqrt{2\pi(\alpha + \beta)n} \left(\frac{(\alpha + \beta)n}{e}\right)^{(\alpha + \beta)n}}{e^{\frac{1}{12}} \sqrt{2\pi\alpha n} \left(\frac{\alpha n}{e}\right)^{\alpha n} e^{\frac{1}{12}} \sqrt{2\pi\beta n} \left(\frac{\beta n}{e}\right)^{\beta n}} \frac{\alpha^{\alpha n} \beta^{\beta n}}{(\alpha + \beta)^{(\alpha + \beta)n}} \\ &= \frac{1}{e^{\frac{1}{6}} \sqrt{2\pi}} \sqrt{\frac{\alpha + \beta}{\alpha\beta n}} \\ &\geq \frac{1}{3} \sqrt{\frac{\alpha + \beta}{\alpha\beta n}}. \end{aligned}$$

■

Corollary 4. *If q divides s then*

$$\omega(L_{q,n,s}) \gtrsim \binom{n}{s} \frac{q^{s+1}}{3\sqrt{(q-1)s}}.$$

Lemma 18. *Let $s, n \in \mathbb{N}$ be divisible by q and let $\mathbf{k} = \frac{n}{q}\mathbf{1}$. The maximum clique size in $L_{q,\mathbf{k},s}^R$ satisfies*

$$\omega(L_{q,\mathbf{k},s}^R) \gtrsim \binom{n}{s} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\max_i \binom{s/q}{i} (q-1)^i q^{-s/q} \right)^q.$$

Proof: Let $s, a, l \in \mathbb{N}$ such that all are divisible by q . Let $k \in \mathbb{N}$ such that $k - 1$ is divisible by q . As in Lemma 16, let $m = lk + 5(k - 1)$ and $n = m + a - (s - a)$. We will construct a string $x \in M(\frac{m}{q}\mathbf{1})$ and a set $S \subseteq M(\frac{n}{q}\mathbf{1})$.

In Lemma 16, we constructed x by concatenating copies of two fixed strings. To satisfy the composition constraint, we need multiple variations of

the components, so $x = u_0 : v_0 : u_1 : v_1 : \dots : u_{k-1}$. For $i \in [k-1]$ let $b, c \in [q]$ such that $b = i \bmod q$ and $c = i + 1 \bmod q$. Let $v_i = bcccb$. For $i \in [k]$, let $u_i \in M(\frac{l}{q}\mathbf{1})$ such that all runs in u_i have length one. Select u_i such that no runs of length two are created in x .

Construct the members of S exactly as in Lemma 16, but with the restriction that exactly a/q new symbols of each type are inserted and exactly $(s-a)/q$ symbols of each type are deleted. This ensures that $S \subseteq M(\frac{n}{q}\mathbf{1})$.

For each $i \in [q]$, each string u has $\frac{l}{q}$ substrings that can be produced by deleting a single i and $l + 1 - \frac{l}{q}$ superstrings that can be produced by inserting a single i . There are $\binom{k}{s}$ ways to pick the subset of the segments that will have some error. There are $|M(\frac{s}{q}\mathbf{1})|$ ways to specify which symbol is involved in each error. Finally, there are $(\frac{s/q}{a/q})^q$ ways to specify which errors are insertions and which are deletions. Thus

$$|S| = \binom{k}{s} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\frac{s/q}{a/q} \right)^q \left(l + 1 - \frac{l}{q} \right)^a \left(\frac{l}{q} \right)^{s-a}.$$

As in Lemma 16, let $k \sim l \sim \sqrt{n}$. Then

$$\begin{aligned} |S| &\sim \frac{(\sqrt{n})^s}{s!} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\frac{s/q}{a/q} \right)^q (q-1)^a \left(\frac{\sqrt{n}}{q} \right)^s \\ &\sim \binom{n}{s} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\frac{s/q}{a/q} \right)^q \frac{(q-1)^a}{q^s}. \end{aligned}$$

■

Corollary 5. *Let $s, n \in \mathbb{N}$ be divisible by q and let $\mathbf{k} = \frac{n}{q}\mathbf{1}$. If q^2 divides s then*

$$\omega(L_{q,\mathbf{k},s}^R) \gtrsim \binom{n}{s} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\frac{1}{3} \sqrt{\frac{q^3}{(q-1)s}} \right)^q.$$

4.2 Consequences for code construction

For a graph G , there are two fundamental lower bounds on chromatic number: $\chi(G) \geq \omega(G)$ and $\chi(G) \geq \frac{|V(G)|}{\alpha(G)}$. In Section 4.1, we found lower bounds on $\omega(L_{q,n,s}^R)$ and $\omega(L_{q,\mathbf{k},s}^R)$ which immediately translate to lower bounds on chromatic number. We can compare these to lower bounds on chromatic number derived from upper bounds on independent set size.

Levenshtein's asymptotic upper bound [1] is

$$\alpha(L_{q,n,s}) \lesssim \frac{q^n}{(q-1)^s \binom{n}{s}},$$

which implies $\chi(L_{q,n,s}) \gtrsim (q-1)^s \binom{n}{s}$. From Lemma 9, we know that cliques in $L_{q,n,s}$ produced by a single common substring contain $I_{q,n,s}$ vertices. Because $I_{q,n,s} \sim (q-1)^s \binom{n}{s}$, this is the same lower bound on chromatic number again. Corollary 4 improves on these two bounds by a factor of $\left(\frac{q}{q-1}\right)^s \frac{q}{3\sqrt{(q-1)s}}$.

Let both s and n be divisible by q and let $\mathbf{k} = \frac{n}{q}\mathbf{1}$ (so $\mathbf{p} = \frac{1}{q}\mathbf{1}$). From Theorem 4,

$$\alpha(L_{q,\mathbf{k},s}^R) \lesssim \frac{|M(\mathbf{k})|}{\binom{n}{s} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\frac{q-1}{q}\right)^s}.$$

This implies

$$\chi(L_{q,\mathbf{k},s}^R) \gtrsim \binom{n}{s} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\frac{q-1}{q}\right)^s.$$

Let $\mathbf{s} = \frac{s}{q}\mathbf{1}$, Lemma 9 tells us that $L_{q,\mathbf{k},s}^R$ contains cliques of $I_{q,\mathbf{k},s}^R$ vertices. From Lemma 2, this is asymptotic to

$$\binom{n}{s} \left| M\left(\frac{s}{q}\mathbf{1}\right) \right| \left(\frac{q-1}{q}\right)^s.$$

Once again, we have derived the same bound twice. Corollary 5 improves on these two bounds by a factor of $\left(\frac{q}{q-1}\right)^s \left(\frac{1}{3}\sqrt{\frac{q^3}{(q-1)s}}\right)^q$.

The existence of these large cliques is a qualitative difference between the single deletion and multiple deletion cases. In the binary single deletion case, the VT weight function has three interesting properties. It is an asymptotically optimal coloring (in fact is it exactly optimal). All of the color classes are asymptotically equal in size. The size of the largest color class asymptotically matches the best known upper bound for independent set, demonstrating that the upper bound is tight. For $s > 1$, it is impossible for a coloring function to simultaneously have all three properties. The lower bound on clique size implies that the average size of the color classes in any proper coloring is significantly smaller than the best known upper bound of independent set.

There are a few ways around this barrier.

One approach is to abandon colorings completely and attempt to construct

independent sets directly. This sounds like an easier task: a single subset is simpler than a partition of the whole space. It is easy to algorithmically construct independent sets that are unrelated to colorings, but it is not clear how to find structured examples of independent sets that are not arising from some coloring. For example, all linear codes partition a space into cosets. A structured independent set could potentially come from an improper coloring with few defects. As long as at least one color class is to be an independent set, we do not care about the others.

Alternatively, we could attempt to construct colorings with color classes of varying sizes. In the best case, this would result in an optimal coloring in which some but not all color classes were optimal independent sets. This is a lot to ask. We are interested in coloring primarily because they can lead to more structured independent sets. Colorings with sufficient variation in the sizes of their color classes might not have this structure. A final, closely related approach is to identify the regions of the graph that contain the large clique, remove these regions, and color the remainder of the graph. This is very similar to reserving a few colors to be used only in the most difficult regions of the graph.

REFERENCES

- [1] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet Physics Doklady*, vol. 10, 1966, p. 707710.
- [2] Y. Liron and M. Langberg, “A characterization of the number of subsequences obtained via the deletion channel,” in *IEEE International Symposium on Information Theory Proceedings*, 2012, p. 503507.
- [3] D. Cullina and N. Kiyavash, “An improvement to Levenshtein’s upper bound on the cardinality of deletion correcting codes,” *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3862–3870, July 2014.
- [4] V. I. Levenshtein, “Elements of coding theory,” *Diskretnaya Matematika i Matematicheskie Voprosy Kibernetiki*, p. 207305, 1974.
- [5] G. Tenengolts, “Nonbinary codes, correcting single deletion or insertion (corresp.),” *IEEE Transactions on Information Theory*, vol. 30, no. 5, p. 766769, 1984.
- [6] V. I. Levenshtein, “On perfect codes in deletion and insertion metric,” *Discrete Mathematics and Applications*, vol. 2, no. 3, pp. 241–258, 1992.
- [7] R. Gabrys, E. Yaakobi, F. Farnoud, and J. Bruck, “Codes correcting erasures and deletions for rank modulation,” in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, p. 27592763.
- [8] F. Farnoud and O. Milenkovic, “Multipermutation codes in the Ulam metric,” in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, p. 27542758.
- [9] F. Sala, R. Gabrys, and L. Dolecek, “Deletions in multipermutations,” in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, p. 27692773.
- [10] D. B. West, *Introduction to Graph Theory*, 2nd ed. Upper Saddle River, NJ: Prentice Hall Upper Saddle River, 2001.
- [11] V. I. Levenshtein, “Bounds for deletion/insertion correcting codes,” in *IEEE International Symposium on Information Theory Proceedings*, 2002, p. 370.

- [12] D. Cullina and N. Kiyavash, “Generalized sphere-packing upper bounds on the size of codes for combinatorial channels,” in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, p. 12661270.
- [13] V. Chvátal, “The tail of the hypergeometric distribution,” *Discrete Mathematics*, vol. 25, no. 3, p. 285287, 1979.
- [14] A. A. Kulkarni and N. Kiyavash, “Non-asymptotic upper bounds for deletion correcting codes,” *IEEE Transactions on Information Theory*, 2012. [Online]. Available: <http://arxiv.org/abs/1211.3128>
- [15] W. Feller, “Stirling’s formula,” in *An Introduction to Probability Theory and Its Applications*, 3rd ed. New York: John Wiley & Sons, 1968, vol. 1, pp. 52–54.