On t-Revealing Codes in Binary Hamming Spaces

Tero Laihonen
Department of Mathematics and Statistics
University of Turku, FI-20014 Turku, Finland
terolai@utu.fi

Abstract

In this paper, we introduce t-revealing codes in the binary Hamming space $\mathbb{F}^n$. Let $C \subseteq \mathbb{F}^n$ be a code and denote by $I_t(C; \mathbf{x})$ the set of elements of C which are within (Hamming) distance t from a word $\mathbf{x} \in \mathbb{F}^n$. A code C is t-revealing if the majority voting on the coordinates of the words in $I_t(C; \mathbf{x})$ gives unambiguously $\mathbf{x}$. These codes have applications, for instance, to the list decoding problem of the Levenshtein's channel model, where the decoder provides a list based on several different outputs of the channel with the same input, and to the information retrieval problem of the Yaakobi-Bruck model of associative memories. We give t-revealing codes which improve some of the key parameters for these applications compared to earlier code constructions.

Keywords: Levenshtein's sequence reconstruction problem, list decoding, information retrieval, associative memory, majority voting on coordinates, indentifying codes

1 Introduction

Let us first define mathematically the codes we are interested in and then consider the motivations and applications of them.

Let $\mathbb{F}$ be the binary field and denote by $\mathbb{F}^n$ the Hamming space, that is, the n-fold Cartesian product $\mathbb{F}^n = \mathbb{F} \times \mathbb{F} \times \cdots \times \mathbb{F}$. As usual, the Hamming distance $d(\mathbf{x}, \mathbf{y})$ between two words $\mathbf{x} = x_1 x_2 \dots x_n$ and $\mathbf{y} = y_1 y_2 \dots y_n$ of $\mathbb{F}^n$ is the number of coordinate places in which they differ. The all-zero word is denoted by $\mathbf{0} = 00 \dots 0$ and the all-one word by $\mathbf{1} = 11 \dots 1$. The *support* of a word $\mathbf{x}$ is defined as $\mathrm{supp}(\mathbf{x}) = \{i \mid x_i \neq 0\}$. The Hamming weight $w(\mathbf{x})$ of $\mathbf{x}$ is the cardinality of the support of $\mathbf{x}$. For $\mathbf{x} \in \mathbb{F}^n$ we denote the Hamming ball of radius t and centred at $\mathbf{x}$ by

$$B_t(\mathbf{x}) = \{\mathbf{y} \in \mathbb{F}^n \mid d(\mathbf{x}, \mathbf{y}) \leq t\}.$$

The *symmetric difference* $A \triangle B$ of two sets A and B is, as usual, $(A \setminus B) \cup (B \setminus A)$. The word $\mathbf{e}_i$ is a word of weight one such that $\mathrm{supp}(\mathbf{e}_i) = \{i\}$. The *complement* of a word $\mathbf{x}$ is the word $\bar{\mathbf{x}} = \mathbf{1} + \mathbf{x}$. A *code* is a subset of $\mathbb{F}^n$ with at least two elements. Its elements are called *codewords*. The *minimum distance* of a code C is defined as

$$d_{\min}(C) = \min_{\substack{\mathbf{c}_1, \mathbf{c}_2 \in C \\ \mathbf{c}_1 \neq \mathbf{c}_2}} d(\mathbf{c}_1, \mathbf{c}_2)$$

and the *covering radius* of C as

$$R(C) = \max_{\mathbf{x} \in \mathbb{F}^n} \min_{\mathbf{c} \in C} d(\mathbf{x}, \mathbf{c}).$$

For $\mathbf{x} = x_1 x_2 \dots x_n$, let the function π_i pick the i-th coordinate, that is, $\pi_i(\mathbf{x}) = x_i$. For a subset $A \subseteq \mathbb{F}^n$, we generalize this in the following way by considering the majority voting on the i-th coordinates of the words in A. If there are more 0's (resp. 1's) among the coordinates $\pi_i(\mathbf{a})$, where $\mathbf{a} \in A$, then $\pi_i(A) = 0$ (resp. $\pi_i(A) = 1$). If there is an equal amount of 0's and 1's, the value $\pi_i(A)$ is defined to be the symbol $*$.

Let C be a code and $t \geq 1$ an integer. For any $\mathbf{x} \in \mathbb{F}^n$, we define the set of codewords within distance t from $\mathbf{x}$ as

$$I_t(\mathbf{x}) = I_t(C; \mathbf{x}) = \{\mathbf{c} \in C \mid d(\mathbf{x}, \mathbf{c}) \leq t\}.$$

We call this the *I-set* of $\mathbf{x}$.

Let $I_t(\mathbf{x})$ be non-empty for a word $\mathbf{x} = x_1 x_2 \dots x_n \in \mathbb{F}^n$. We say that the word $\mathbf{x}$ is *accessible*, if $\pi_i(I_t(\mathbf{x})) = x_i$ for all $i = 1, 2, \dots, n$. In other words, using the majority voting on the coordinates of $I_t(\mathbf{x})$ we get $\mathbf{x}$. Otherwise, we say that $\mathbf{x}$ is *non-accessible* (in particular, if I-set of $\mathbf{x}$ is empty).

Next we define a useful function $m_t(\mathbf{x})$ on an accessible word $\mathbf{x}$. Let k be the smallest integer such that if we take *any* subset $U \subseteq I_t(\mathbf{x})$ of size $|U| \geq k$, then $\pi_i(U) = x_i$ for all $i = 1, \dots n$. In other words, it is enough to take any k codewords from $I_t(\mathbf{x})$ in order to find $\mathbf{x}$ using the majority voting on the coordinates of U. The smallest such k is denoted by $m_t(\mathbf{x}) = m_t(C; \mathbf{x})$. We say that $\mathbf{x}$ is *revealed* from $I_t(\mathbf{x})$ using any $m_t(\mathbf{x})$ (or more) words of $I_t(\mathbf{x})$.

Example 1. Let the code $C = \{0000, 0100, 1100, 0110, 0111, 1011\}$. For the word $\mathbf{x} = 0100$ we have $I_1(\mathbf{x}) = \{0000, 0100, 1100, 0110\}$. Clearly, now $\pi_i(I_1(\mathbf{x})) = x_i$ for all $i = 1, 2, 3, 4$, so $\mathbf{x}$ is accessible. It is easy to check that any subset of three codewords of $I_1(\mathbf{x})$ also reveals $\mathbf{x}$ using the majority voting. Hence, $m_1(\mathbf{x}) \leq 3$. Since $U = \{0100, 1100\}$ gives $\pi_1(U) = *$, we get $m_1(\mathbf{x}) = 3$.

If $\mathbf{y} = 1111$, then $I_1(\mathbf{y}) = \{0111, 1011\}$. The word $\mathbf{y}$ is non-accessible, since $\pi_2(I_1(\mathbf{x})) = *$.

Let $\mathbb{N} = \{0, 1, \dots\}$ be the set of natural numbers. For a word $\mathbf{x} = x_1 x_2 \dots x_n$ we define a vector $\mathbf{h}_t(\mathbf{x}) = \mathbf{h}_t(C; \mathbf{x}) = (h_1, h_2, \dots, h_n) \in \mathbb{N}^n$ where h_i is the number of codewords in $I_t(\mathbf{x})$ such that their i-th coordinate differs from x_i. Hence, $\mathbf{x}$ is accessible, if

$$|I_t(\mathbf{x})| \geq 2 \max_{i=1,\dots,n} h_i + 1 \tag{1}$$

and, in that case,

$$m_t(\mathbf{x}) = 2 \max_{i=1,\dots,n} h_i + 1. \tag{2}$$

Definition 2. Let $t \geq 1$ and $n \geq 2$ be integers. A code $C \subseteq \mathbb{F}^n$ is a *coordinatewise revealing code of radius t* (a *t-revealing* code for short) if every word $\mathbf{x} \in \mathbb{F}^n$ is accessible. For such a code, denote the parameter

$$\hat{\mu}_t(C) = \max_{\mathbf{x} \in \mathbb{F}^n} m_t(C; \mathbf{x}).$$

Furthermore, let $\hat{\mu}_t(n)$ denote the minimum of $\hat{\mu}_t(C)$ over all t-revealing codes C in $\mathbb{F}^n$.

Example 3. Let $C = \mathbb{F}^3 \setminus \{000, 111\}$. For the word $\mathbf{z} = 000$, we get $\mathbf{h}_1(\mathbf{z}) = (1, 1, 1)$ and $|I_1(\mathbf{z})| = 3$. Due to (1) and (2) it follows that $m_1(\mathbf{z}) = 3$. For $\mathbf{y} = 001$, the vector $\mathbf{h}_1(\mathbf{y}) = (1, 1, 0)$ and $|I_1(\mathbf{y})| = 3$. Again $\mathbf{y}$ is accessible and $m_1(\mathbf{y}) = 3$. Similarly, one can check that $m_1(\mathbf{x}) = 3$ for all $\mathbf{x} \in \mathbb{F}^3$. Consequently, C is a 1-revealing code with $\hat{\mu}_1(C) = 3$. Later (in Theorem 7) we will see that $\hat{\mu}_1(3) = 3$.

In the sequel, we will need the following observations.

Lemma 4. *Let C be a t-revealing code, and let $\mathbf{x}$ and $\mathbf{y}$ be any distinct words in $\mathbb{F}^n$.*

(i) We have

$$|I_t(\mathbf{x}) \cap I_t(\mathbf{y})| \leq \max\{m_t(\mathbf{x}), m_t(\mathbf{y})\} - 1. \tag{3}$$

(ii) We also have

$$|I_t(\mathbf{x}) \triangle I_t(\mathbf{y})| \geq 2. \tag{4}$$

Proof. (i) Because C is a t-revealing code, the values $m_t(\mathbf{x})$ and $m_t(\mathbf{y})$ exist. Assume, without loss of generality, that $m_t(\mathbf{y}) \geq m_t(\mathbf{x})$. Suppose to the contrary that $|I_t(\mathbf{x}) \cap I_t(\mathbf{y})| \geq \max\{m_t(\mathbf{x}), m_t(\mathbf{y})\} = m_t(\mathbf{y})$. Consider the codewords in $U = I_t(\mathbf{x}) \cap I_t(\mathbf{y})$. Since C is t-revealing, we know that any subset of $m_t(\mathbf{y})$ or more codewords of $I_t(\mathbf{y})$ — in particular, the set U — reveals

$\mathbf{y}$ uniquely. Also these same codewords in U should reveal uniquely $\mathbf{x}$ because $|U| \geq m_t(\mathbf{x})$ and $U \subseteq I_t(\mathbf{x})$. However, this is a contradiction, since $\mathbf{x} \neq \mathbf{y}$.

(ii) Since $\mathbf{x} \neq \mathbf{y}$, they differ in at least one coordinate, say, $x_i \neq y_i$. By (3), we know that $I_t(\mathbf{x}) \triangle I_t(\mathbf{y})$ is non-empty, i.e., $|I_t(\mathbf{x}) \triangle I_t(\mathbf{y})| > 0$. Suppose that $|I_t(\mathbf{x}) \triangle I_t(\mathbf{y})| = 1$ and, without loss of generality, there is a codeword $\mathbf{c} \in I_t(\mathbf{y}) \setminus I_t(\mathbf{x})$. Since C is t-revealing and $I_t(\mathbf{x}) = I_t(\mathbf{x}) \cap I_t(\mathbf{y})$, we know that all the codewords in $I_t(\mathbf{x}) \cap I_t(\mathbf{y})$ reveal uniquely $\mathbf{x}$. Now these codewords together with $\mathbf{c}$ should reveal uniquely $\mathbf{y}$, because $I_t(\mathbf{y}) = \{\mathbf{c}\} \cup (I_t(\mathbf{x}) \cap I_t(\mathbf{y}))$. But this is impossible, since in majority voting for x_i to change to y_i, one needs at least two more votes — and $\mathbf{c}$ can give only one. □

It should be noted that the necessary condition of (4) is not sufficient for a code to be t-revealing. For example, the code $C = \{\mathbf{0}\} \cup (\mathbb{F}^6 \setminus (B_2(\mathbf{0}))$ is 2-revealing, and hence (4) is satisfied. However, adding one codeword, namely, $\mathbf{c} = 100000$ to C, the new code C' is not 2-revealing (since $\mathbf{0}$ is not accessible), although clearly the new code C' still satisfies (4).

Next we consider the applications and the motivations of the codes defined above. The first application is the list decoding problem of Levenshtein's channel model [13, 16], which finds its original motivation in molecular biology and chemistry, where the usual redundancy method is not feasible, and it is also relevant for recent advanced storage technologies [17]. The second application is the information retrieval in associative memories [16, 18, 9, 7, 8, 15] and the third motivation is the identification in sensor networks [11, 2, 3, 6, 5].

1) *The list decoding problem for the Levenshtein's channel model:* A codeword $\mathbf{x} \in C$ is transmitted through N channels where at most t errors can occur in each of them as illustrated in Figure 1. It is also assumed that $t > \lfloor (d_{\min}(C) - 1)/2 \rfloor$.

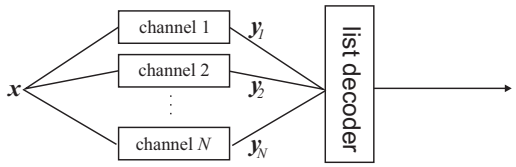

Figure 1: The channel model.

Based on the N different outputs $\mathbf{y}_1, \ldots, \mathbf{y}_N$ of the channels, the list decoder $\mathcal{D}_{\mathcal{L}}$ gives estimations $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ (where $\ell \leq \mathcal{L}$) on the transmitted word $\mathbf{x}$. In [16, 9], a successful decoder is considered (successful means that the transmitted word $\mathbf{x}$ belongs to the outputted list) and the maximal length of the list $\mathcal{L}$ is considered with respect to the number of channels N. Naturally, we would like to have as short output list as possible while keeping N small and the cardinality of the code as large as possible. In [17], it is shown that if we wish to have a unique output (that is, $\mathcal{L} = 1$), then the number of channels can be inconveniently large — see also Remark 14.

In this paper, we will focus on the case when there are only two channels, that is, $N = 2$, and we try to find large codes giving a short output list from the decoder. Suppose that C is a t-revealing code and $N = 2$. Next we see that we obtain a successful decoder with $\mathcal{L} \leq \hat{\mu}_t(C) - 1$. Two different words $\mathbf{y}_1$ and $\mathbf{y}_2$ are received from the channels and the decoder outputs all the codewords $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ of C such that $d(\mathbf{y}_j, \mathbf{x}_i) \leq t$ for all $j = 1, 2$ and $i = 1, \ldots, \ell$ (that is, all those codewords that could have been sent when the two words $\mathbf{y}_1$ and $\mathbf{y}_2$ were received). In other words, the list consists of the codewords in $I_t(\mathbf{y}_1) \cap I_t(\mathbf{y}_2)$. By (3), the length of this list is at most $\hat{\mu}_t(C) - 1$. The decoder is clearly successful, since $\mathbf{x} \in I_t(\mathbf{y}_1) \cap I_t(\mathbf{y}_2)$ due to the fact that at most t errors occurred in the channels.

2) *Information retrieval in an associative memory:* In the model of Yaakobi and Bruck [16], an associative memory is given as a (simple and undirected) graph $G = (V, E)$. A vertex in the graph corresponds to a stored information unit and if two information units are associated, then there is an edge between them. Moreover, two vertices are called *t-associated*, if the graphical distance (that is, the number of edges) between them is at most t. An unknown information unit

$x \in V$ is retrieved from the associative memory using *input clues* (provided by an information seeker) which are t-associated to x and also belong to a *reference set* $C \subseteq V$. The reference set should be such that given enough input clues, the sought information unit x can be unambiguously found. Naturally, we want the maximum number $\hat{m}$ of input clues, which are needed to retrieve any information unit from the memory, to be as small as possible.

In this paper (like in [16, 9, 7]), we concentrate on the associative memory modelled by the binary hypercube $\mathbb{F}^n$ (for other graphs see, for instance, [18, 8, 12, 10]). Here two words (i.e., information units) $\mathbf{a}$ and $\mathbf{b}$ are t-associated if and only if $d(\mathbf{a}, \mathbf{b}) \leq t$. According to the model above, we wish to find a sought information unit $\mathbf{x}$ with the aid of input clues coming from the code C (the reference set) which are t-associated to the unknown word $\mathbf{x}$. In other words, the input clues come from the set $I_t(\mathbf{x})$. If the reference set C is a t-revealing code, then we can uniquely and efficiently (due to the majority voting) find the information unit by receiving at most $\hat{\mu}_t(C)$ input clues. Therefore, the maximum number of needed input clues satisfies $\hat{m} \leq \hat{\mu}_t(C)$. Here it is natural to have as small code as possible for the reference set.

3) *Identification in sensor networks:* The motivation behind the identifying codes is locating objects in a sensor networks [11, 2]. A code C is called *t-identifying* if

$$I_t(C; \mathbf{x}) \neq I_t(C; \mathbf{y})$$

for all $\mathbf{x} \neq \mathbf{y}$. The idea is that given a set $I_t(\mathbf{x})$ we can uniquely determine $\mathbf{x}$ by comparing $I_t(\mathbf{x})$ to other sets of $I_t(\mathbf{y})$. If C is t-revealing, then it is also t-identifying due to (4). The advantage of using t-revealing codes is that we find $\mathbf{x}$ from $I_t(\mathbf{x})$ just by performing the coordinatewise majority voting and no comparison to other I-sets (or knowledge of them) is needed.

Earlier in [16, 7, 9] the length $\mathcal{L}$ of the output of the list decoder and the maximum number of input clues $\hat{m}$ in an associative memory was considered using codes $C \subseteq \mathbb{F}^n$ which are based on limiting the size of the intersections $I_t(\mathbf{x}) \cap I_t(\mathbf{y})$ while the codes have the property that $I_t(\mathbf{x}) \setminus I_t(\mathbf{y}) \neq \emptyset$ for all $\mathbf{x} \neq \mathbf{y}$ (see, for instance, Theorem 9 in [9]). In this paper, we use the idea of majority voting on coordinates in designing the codes and not the intersections. But as we saw in (3), we can still estimate the intersections (needed, for example, in the list decoding problem as explained above). We will see that the new class of t-revealing codes provides better results for the length $\mathcal{L}$ and for the number of input clues $\hat{m}$ than the earlier code constructions.

Notice that the t-revealing codes may have $I_t(\mathbf{x}) \setminus I_t(\mathbf{y}) = \emptyset$ for distinct words $\mathbf{x}$ and $\mathbf{y}$, so they are not the same codes as above in [9, 16, 7] (or in [4]) — conversely, those codes do not usually give t-revealing codes.

The structure of the paper is as follows. In Section 2 we provide optimal t-revealing codes for radii $t = 1$, $t = 2$ and $t = n - 1$. Moreover, we discuss bounds on the cardinality of the codes and a shortening method. In Section 3, we give constructions on 3-revealing codes based on error-correcting codes. In section 4, we consider constructions and lower bound for other radii.

2 Linear codes and optimal results

We can often benefit from codes being linear, so let us first recall some basic facts about them (see, for example, [14, Chapter 1] for more details).

A code C is *linear* if it is a subspace of $\mathbb{F}^n$. If C has dimension k, then C can be defined using an $(n-k) \times n$ check matrix H where

$$C = \{\mathbf{x} \in \mathbb{F}^n \mid H\mathbf{x}^T = \mathbf{0}\}$$

where $\mathbf{x}^T$ denotes the transpose of $\mathbf{x}$. Let us denote the columns of H as follows

$$H = (\mathbf{h}^{(1)} \mid \mathbf{h}^{(2)} \mid \cdots \mid \mathbf{h}^{(n)}).$$

The *syndrome* of a word $\mathbf{y} = y_1 y_2 \dots y_n$ is defined as $s(\mathbf{y}) = H\mathbf{y}^T$. Notice that the syndrome $s(\mathbf{y})$ is obtained also by summing up the columns $\mathbf{h}^{(i)}$ where $y_i = 1$. The space $\mathbb{F}^n$ can be partitioned

into cosets $\mathbf{x}+C$, $\mathbf{x}\in\mathbb{F}^n$, and two cosets $\mathbf{x}+C$ and $\mathbf{y}+C$ are equal if $s(\mathbf{x})=s(\mathbf{y})$. A word of minimum weight in a coset is chosen as the *coset leader*.

If a code C is linear, then it has the *dual code* $C^\perp=\{\mathbf{z}H \mid \mathbf{z}\in\mathbb{F}^{n-k}\}$. The *weight distribution* of a code C is the $(n+1)$-tuple $(A_0,A_1,\ldots,A_n)$ where A_i is the number of codewords of weight i in C.

For any subset $A\subseteq\mathbb{F}^n$ and a word $\mathbf{b}\in\mathbb{F}^n$ we define $d(\mathbf{b},A)=\min\{d(\mathbf{b},\mathbf{a}) \mid \mathbf{a}\in A\}$ and $\mathbf{b}+A=\{\mathbf{b}+\mathbf{a} \mid \mathbf{a}\in A\}$.

Next we will consider useful results regarding linear codes and the codes of type $\mathbf{x}+C$ (where C does not have to be linear).

Theorem 5. *(i) Let $C\subseteq\mathbb{F}^n$ be code and $\mathbf{x}\in\mathbb{F}^n$. We have $\mathbf{h}_t(\mathbf{x}+C;\mathbf{y})=\mathbf{h}_t(C;\mathbf{x}+\mathbf{y})$ and $|I_t(\mathbf{x}+C;\mathbf{y})|=|I_t(C;\mathbf{x}+\mathbf{y})|$ for all $\mathbf{y}\in\mathbb{F}^n$. If the word $\mathbf{x}+\mathbf{y}$ is accessible with respect to the code C, then $\mathbf{y}$ is accessible with respect to $\mathbf{x}+C$ and, moreover, $m_t(\mathbf{x}+C;\mathbf{y})=m_t(C;\mathbf{y}+\mathbf{x})$.*

(ii) Let C be a linear t-revealing code. Then $s(\mathbf{x})=s(\mathbf{y})$ implies that $m_t(\mathbf{x})=m_t(\mathbf{y})$. In particular, all the words in a coset have the same minimum number of revealing codewords as the coset leader.

Proof. (i) We will use the observation

$$I_t(\mathbf{x}+C;\mathbf{y})=I_t(C;\mathbf{x}+\mathbf{y})+\mathbf{x}. \tag{5}$$

Let us verify this first. The word $\mathbf{a}$ belongs to the set $I_t(\mathbf{x}+C;\mathbf{y})$ if and only if $\mathbf{a}=\mathbf{x}+\mathbf{c}$ for some $\mathbf{c}\in C$ and $d(\mathbf{x}+\mathbf{c},\mathbf{y})\le t$. This is equivalent to the fact that $\mathbf{a}=\mathbf{c}+\mathbf{x}$ for $\mathbf{c}\in C$ and $d(\mathbf{c},\mathbf{x}+\mathbf{y})\le t$. This in turn, is equivalent to the fact that $\mathbf{a}$ belongs to the set $I_t(C;\mathbf{x}+\mathbf{y})+\mathbf{x}$.

From the observation (5) it immediately follows that $|I_t(\mathbf{x}+C;\mathbf{y})|=|I_t(C;\mathbf{x}+\mathbf{y})|$. Next we show that $\mathbf{h}_t(\mathbf{x}+C;\mathbf{y})=\mathbf{h}_t(C;\mathbf{x}+\mathbf{y})$ using (5). Denote $\mathbf{h}_t(C;\mathbf{x}+\mathbf{y})=(h_1,\ldots,h_n)$ and $\mathbf{h}_t(\mathbf{x}+C;\mathbf{y})=(h'_1,\ldots,h'_n)$. We will show that these two vectors are the same. Consider any fixed coordinate i. Let $\pi_i(\mathbf{x})=x_i$ and $\pi_i(\mathbf{y})=y_i$ and thus $\pi_i(\mathbf{x}+\mathbf{y})=x_i+y_i$. Suppose first that $x_i=0$. In the set $I_t(C;\mathbf{x}+\mathbf{y})$ there are h_i codewords with the bit $1+x_i+y_i=1+y_i$ (the bit had to differ from x_i+y_i) in the i-th coordinate. Since $x_i=0$ the same is true for the set $I_t(C;\mathbf{x}+\mathbf{y})+\mathbf{x}$. In the set $I_t(\mathbf{x}+C;\mathbf{y})$ there are h'_i codewords with the bit $1+y_i$ in the corresponding coordinate. Due to (5), we obtain $h_i=h'_i$. Assume then that $x_i=1$. In the set $I_t(C;\mathbf{x}+\mathbf{y})$ there are h_i codewords with the bit $1+x_i+y_i=y_i$ in the i-th coordinate. Consequently, in the set $I_t(C;\mathbf{x}+\mathbf{y})+\mathbf{x}$ there are h_i codewords with $1+y_i$ in the i-th coordinate. As before, in $I_t(\mathbf{x}+C;\mathbf{y})$ there were h'_i codewords which have $1+y_i$ in the i-th coordinate. Again we obtain $h_i=h'_i$ by (5). This implies that $\mathbf{h}_t(C;\mathbf{x}+\mathbf{y})=\mathbf{h}_t(\mathbf{x}+C;\mathbf{y})$. Immediately it follows that if $\mathbf{x}+\mathbf{y}$ is accessible in C, then $\mathbf{y}$ is accessible in $\mathbf{x}+C$ and $m_t(\mathbf{x}+C;\mathbf{y})=m_t(C;\mathbf{x}+\mathbf{y})$.

(ii) Let now C be a t-revealing code, which is also linear. Let $\mathbf{x}$ be the leader of the coset $\mathbf{x}+C$. Since $s(\mathbf{x})=s(\mathbf{y})$, we know that $\mathbf{y}\in\mathbf{x}+C$. Therefore, $\mathbf{y}=\mathbf{x}+\mathbf{c}$ for some codeword $\mathbf{c}\in C$. Since C is linear, we know that $\mathbf{c}+C=C$. Using (i) we obtain $m_t(\mathbf{y})=m_t(C;\mathbf{x}+\mathbf{c})=m_t(\mathbf{c}+C;\mathbf{x})=m_t(C;\mathbf{x})=m_t(\mathbf{x})$ as claimed. □

Next we give some constructions to revealing codes. Before that, let us recall a couple of results concerning the $(n-k)\times n$ check matrix H of a linear code C. The covering radius $R(C)$ is the smallest integer R such that every word in $\mathbb{F}^{n-k}$ can be written as the sum of at most R columns of the matrix H (see [1, Theorem 2.1.9]). Moreover, the covering radius of C is the largest of the weights of the coset leaders (see [1, Theorem 2.1.11]). The minimum distance $d_{\min}(C)$ equals d if and only if every $d-1$ columns of H are linearly independent and there exist d columns which are linearly dependent (see [14, p. 33]).

Theorem 6. *There exist codes giving*

(i) $\hat{\mu}_1(n)\le 3$ for all $n\ge 3$,

(ii) $\hat{\mu}_2(n)\le 3$ for all $n=2^r-1-p$ where $r\ge 3$ and $0\le p\le 2^{r-1}-3$.

(iii) $\hat{\mu}_{n-1}(n) = 2^n - 1$ *for all* $n \geq 2$.

Proof. (i) Consider first the radius $t = 1$. We will show that the linear code C with $r \times n$ check matrix H such that it contains every non-zero column (of $\mathbb{F}^r$) at least 3 times and there are no zero-columns in H is 1-revealing. Since every word of $\mathbb{F}^r$ appears as a column of H, the covering radius of C equals one and, therefore, the weight of any coset leader is at most one. In addition, $d_{\min}(C) = 2$, since H contains no zero-column and there exists two identical columns. By Theorem 5(ii), it is enough to consider coset leaders when we want to calculate the values $m_t(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{F}^n$. Suppose first that the weight of the coset leader $\mathbf{x}$ equals zero, so in other words $\mathbf{x} = \mathbf{0}$. Since $d_{\min}(C) = 2$, we know that $I_1(\mathbf{0}) = \{\mathbf{0}\}$. Trivially, $\mathbf{h}_1(\mathbf{0}) = (0, 0, \ldots, 0)$, so $\max h_i = 0$ and thus, by (1) and (2), we get $m_1(\mathbf{0}) = 1$. Assume then that the coset leader $\mathbf{x}$ has weight one. Let the syndrome $s(\mathbf{x}) = \mathbf{s}$ (where $\mathbf{s} \neq \mathbf{0}$). Now the $I_1(\mathbf{x}) = \{\mathbf{x} + \mathbf{e}_i \mid i \in \mathcal{I}\}$ where $\mathcal{I}$ consists of all of those indices j for which the column $\mathbf{h}^{(j)} = \mathbf{s}$. Since H contains as a column each word of $\mathbb{F}^r$ at least three times, we get $|I_1(\mathbf{x})| \geq 3$. Now the vector $\mathbf{h}_1(\mathbf{x}) = (h_1, \ldots, h_n)$ is such that $h_i = 1$ for $i \in \mathcal{I}$ and $h_i = 0$ if $i \notin \mathcal{I}$. Therefore, by (1) and (2), we obtain $m_1(\mathbf{x}) = 3$. This yields that $\hat{\mu}_1(C) = 3$ and $\hat{\mu}_1(n) \leq 3$.

(ii) Let $t = 2$ and consider the check matrix of a Hamming code $\mathcal{H}_r$ of length $n = 2^r - 1$, that is, H contains all the non-zero columns of $\mathbb{F}^r$ exactly once. We have $d_{\min}(\mathcal{H}_r) = 3$ and $R(\mathcal{H}_r) = 1$. Suppose first that the coset leader equals $\mathbf{x} = \mathbf{0}$. Now $I_2(\mathbf{x}) = \{\mathbf{x}\}$, so $m_2(\mathbf{x}) = 1$. Let then the weight of the coset leader be one, say $\mathbf{x} = \mathbf{e}_k$, and thus, $s(\mathbf{x}) = \mathbf{h}^{(k)}$. We can partition the remaining columns $\mathbf{h}^{(j)}$ of H, $j \neq k$, using pairs $\{\mathbf{h}^{(j)}, \mathbf{h}^{(j')}\}$ where $\mathbf{h}^{(j')} = \mathbf{h}^{(j)} + \mathbf{h}^{(k)}$. Consequently, $I_2(\mathbf{x}) = \{\mathbf{x} + \mathbf{e}_k\} \cup \{\mathbf{x} + \mathbf{e}_j + \mathbf{e}_{j'} \mid \text{for all the pairs } \{j, j'\}\}$. Thus $|I_2(\mathbf{x})| = 2^{r-1}$. It also follows that $\mathbf{h}_2(\mathbf{x}) = (1, 1, 1, \ldots, 1)$. This implies that $m_2(\mathbf{x}) = 3$. Consequently, $\hat{\mu}_t(n) \leq 3$.

In order to deal with the lengths $n - p$ where $0 < p \leq 2^{r-1} - 3$ we use a shortening trick, namely, we use the code

$$C_p = \{c_1 c_2 \ldots c_{n-p} \mid \mathbf{c} = c_1 c_2 \ldots c_{n-p} 0^p \in \mathcal{H}_r\} \subseteq \mathbb{F}^{n-p}$$

where 0^p stands for p consecutive zeros. Now $I_2(C_p; \mathbf{x})$ where $\mathbf{x} = x_1 x_2 \ldots x_{n-p}$ contains exactly the codewords in $I_2(\mathcal{H}_r; \mathbf{x}0^p)$ which end in p zeros (and these zeros are removed from them). Notice that a word $\mathbf{x}0^p$ belongs to the code $\mathcal{H}_r$ if and only if $\mathbf{x}$ is in C_p. Therefore, if $\mathbf{x}$ is a codeword of C_p, then $I_2(C_p; \mathbf{x}) = \{\mathbf{x}\}$ and $m_2(\mathbf{x}) = 1$. On the other hand, if $\mathbf{x}$ is a non-codeword, then $|I_2(C_p; \mathbf{x})| \geq |I_2(\mathcal{H}_r; \mathbf{x}0^p)| - p$ because for each p last coordinates there exists one codeword in $I_2(\mathcal{H}_r; \mathbf{x}0^p)$ which has 1 in that position. Due to the fact that $p \leq 2^{r-1} - 3$, we get $|I_2(C_p; \mathbf{x})| \geq 3$ and since the maximum value of a coordinate in $\mathbf{h}_2(C_p; \mathbf{x})$ equals 1, it follows that $m_2(\mathbf{x}) = 3$.

(iii) If there is a $(n-1)$-revealing code, then it must be $C = \mathbb{F}^n$. Indeed, suppose that C is $(n-1)$-revealing and, say $\mathbf{0} \notin C$. For $\mathbf{x} \neq \mathbf{1}$, we have $B_{n-1}(\mathbf{1}) \triangle B_{n-1}(\mathbf{x}) = \{\mathbf{0}, \bar{\mathbf{x}}\}$. This implies that $|I_{n-1}(\mathbf{1}) \triangle I_{n-1}(\mathbf{x})| \leq 1$, which is a contradiction with (4). It remains to be shown that $C = \mathbb{F}^n$ is actually $(n-1)$-revealing. Since $C = \mathbb{F}^n$ is linear, we only need to consider $m_{n-1}(\mathbf{x})$ for $\mathbf{x} = \mathbf{0}$. Clearly, $|I_{n-1}(\mathbf{x})| = 2^n - 1$ and $\mathbf{h}_{n-1}(\mathbf{x}) = (2^{n-1} - 1, \ldots, 2^{n-1} - 1)$, so $m_{n-1}(\mathbf{x}) = 2^n - 1$. □

The previous constructions in (i) and (ii) are optimal according to the next result.

Theorem 7. *For* $t \geq 1$ *and* $n \geq 3$ *we have* $\hat{\mu}_t(n) \geq 3$.

Proof. Let C be a t-revealing code in $\mathbb{F}^n$, $n \geq 3$. We show that $\hat{\mu}_t(C) \geq 3$ from which the claim follows. If there exists $\mathbf{c} \in C$ such that $I_t(\mathbf{c})$ contains at least two codewords, say $\mathbf{c}$ and $\mathbf{c}'$, then they both belong to the set $I_t(\mathbf{c}) \cap I_t(\mathbf{c}')$ and hence, by (3), we know that $m_t(\mathbf{c}) \geq 3$ or $m_t(\mathbf{c}') \geq 3$. Assume therefore, that for all $\mathbf{c} \in C$ we have $I_t(\mathbf{c}) = \{\mathbf{c}\}$. Choose any $\mathbf{x} \in B_1(\mathbf{c})$ with $\mathbf{x} \neq \mathbf{c}$. The words $\mathbf{c}$ and $\mathbf{x}$ differ in exactly one coordinate, say $c_i \neq x_i$. Now $\mathbf{h}_t(\mathbf{x}) = (h_1, \ldots, h_n)$ has $h_i \geq 1$ and hence $\max_{j=1,2\ldots,n} h_j \geq 1$. By (2), we obtain $m_t(\mathbf{x}) \geq 3$. This yields the assertion $\hat{\mu}_t(C) \geq 3$. □

Let us return to the shortening trick in the proof of Theorem 6(ii) and formulate it for arbitrary $t \geq 1$ for later use. Let $C \subseteq \mathbb{F}^n$ be t-revealing. We have

$$I_t(C_p; \mathbf{x}) = \{\mathbf{c} \in \mathbb{F}^{n-p} \mid \mathbf{c}0^p \in I_t(C; \mathbf{x}0^p)\}.$$

Let $\mathbf{h}_t(C;\mathbf{x}0^p) = (h_1,\ldots,h_n)$ and $\mathbf{h}_t(C_p;\mathbf{x}) = (h'_1,\ldots,h'_{n-p})$. If there are enough codewords left in $I_t(C_p;\mathbf{x})$ after shortening, namely, if

$$|I_t(C_p;\mathbf{x})| \geq m_t(C;\mathbf{x}0^p)$$

then the requirements (1) and (2) (since $h_i \geq h'_i$ for all $i = 1,\ldots,n-p$) are satisfied for the shortened code C_p and it is t-revealing. Furthermore, we can estimate

$$|I_t(C_p;\mathbf{x})| \geq |I_t(C;\mathbf{x}0^p)| - \sum_{i=n-p+1}^{n} h_i.$$

In summary, we get the following statement.

Theorem 8. *Let $C \subseteq \mathbb{F}^n$ be a t-revealing code. Then the shortened code C_p is also t-revealing and $\hat{\mu}_t(C) \geq \hat{\mu}_t(C_p)$ provided that for all $\mathbf{x} \in \mathbb{F}^{n-p}$ we have*

$$|I_t(C;\mathbf{x}0^p)| - \sum_{i=n-p+1}^{n} h_i \geq m_t(C;\mathbf{x}0^p) \tag{6}$$

where $\mathbf{h}_t(C;\mathbf{x}0^p) = (h_1,\ldots,h_n)$.

Let us denote the cardinality of the size of the ball of radius t in $\mathbb{F}^n$ by $V(n,t)$.

Theorem 9. *Let $t \geq 1$.*

(i) If a code $C \subseteq \mathbb{F}^n$ is such that the intersection of I-sets of any distinct words $\mathbf{x}$ and $\mathbf{y}$ satisfies $|I_t(\mathbf{x}) \cap I_t(\mathbf{y})| \leq \mathcal{L}'$, then there we have the upper bound

$$|C| \leq \mathcal{L}' \frac{2^n}{V(n,t) - \binom{n-1}{t}}. \tag{7}$$

If C is a t-revealing code, then this bound holds for $\mathcal{L}' = \hat{\mu}_t(C) - 1$.

(ii) If C is t-revealing, we have a lower bound

$$|C| \geq \frac{3 \cdot 2^n}{V(n,t)+2}. \tag{8}$$

Proof. (i) For the upper bound, choose a set $S = B_t(\mathbf{0}) \cap B_t(\mathbf{e}_1)$. One obtains

$$\sum_{\mathbf{x}\in\mathbb{F}^n} |(\mathbf{x}+S) \cap C| = |S||C|.$$

Since $\mathbf{x}+S = B_t(\mathbf{x}) \cap B_t(\mathbf{x}+\mathbf{e}_1)$, and thus, $(\mathbf{x}+S)\cap C = I_t(\mathbf{x}) \cap I_t(\mathbf{x}+\mathbf{e}_1)$, we get by the assumption that $|(\mathbf{x}+S)\cap C| \leq \mathcal{L}'$. This implies that $2^n\mathcal{L}' \geq |S||C|$. For the claim (7) it suffices to notice that $|S| = V(n,t) - \binom{n-1}{t}$. By virtue of (3) we obtain the claim with $\mathcal{L}' = \hat{\mu}_t(C) - 1$ for a t-revealing code.

(ii) Next we will verify the lower bound. We examine the number M of pairs $(\mathbf{c},\mathbf{x}) \in C \times \mathbb{F}^n$ such that $d(\mathbf{x},\mathbf{c}) \leq t$. Denote $V_i = |\{\mathbf{x} \in \mathbb{F}^n \mid |I_t(\mathbf{x})| = i\}|$ for $i = 0,1,\ldots,V(n,t)$. Since C is t-revealing, $V_0 = 0$. In addition, $V_2 = 0$. Indeed, suppose that there exists $\mathbf{y}$ such that $I_t(\mathbf{y}) = \{\mathbf{c},\mathbf{c}'\}$. The codewords $\mathbf{c}$ and $\mathbf{c}'$ differ in at least one coordinate, say $c_i \neq c'_i$. However, then $\pi_i(I_t(\mathbf{y})) = *$, which is not allowed for a t-revealing code. Counting the pairs, we get

$$|C|V(n,t) = M = \sum_{i=0}^{V(n,t)} iV_i \geq 1\cdot|C| + 3\cdot(2^n - |C|),$$

since each non-codeword $\mathbf{z}$ has $m_t(\mathbf{z}) \geq 3$ giving necessarily $|I_t(\mathbf{z})| \geq 3$. □

Remark 10. Notice that the lower bound (8) can be attained (a small code is what we prefer for the information retrieval). For example, the infinite family of codes in the proof of Theorem 6(i) for the lengths $n = 3(2^r - 1)$ achieve the bound where $r \geq 1$. Indeed, each non-zero column of H appears exactly three times giving $|I_1(\mathbf{x})| = 3$ for non-codewords and $|I_1(\mathbf{x})| = 1$ for the codewords.

For $t = 2$ the above upper bound (7) gives for $\mathcal{L}' = 2$ that $|C| \leq 2^n/n$. The codes in Theorem 6(ii) give $\hat{\mu}_2(C) = 3$, so these codes satisfy $\mathcal{L}' = 2$. The ratio between the cardinality of codes $\mathcal{H}_r$ in Theorem 6(ii) and the bound (7) approaches to 1 when n tends to infinity. Large codes is what we prefer for the Levenshtein's channel problem.

Remark 11. The result $\hat{\mu}_2(n) = 3$ in Theorem 6(ii) gives the bound $\mathcal{L} = 2$ for the length of the decoder list and the bound $\hat{m} = 3$ for the maximal number of input clues in information retrieval. This improves on the known constructions [9, 7], which provide the bounds $\mathcal{L} = 4$ and $\hat{m} = 5$, respectively.

3 Optimal results for the radius $t = 3$

In this section, we consider the case of radius $t = 3$. Let $C_1 \subseteq \mathbb{F}^n$ and $C_2 \subseteq \mathbb{F}^n$ be codes (not necessarily revealing). We will utilize the following additive properties valid for all $t \geq 1$ and $\mathbf{x} \in \mathbb{F}^n$: if $C_1 \cap C_2 = \emptyset$, then

$$\mathbf{h}_t(C_1 \cup C_2; \mathbf{x}) = \mathbf{h}_t(C_1; \mathbf{x}) + \mathbf{h}_t(C_2; \mathbf{x})$$

and

$$|I_t(C_1 \cup C_2; \mathbf{x})| = |I_t(C_1; \mathbf{x})| + |I_t(C_2; \mathbf{x})|.$$

In Theorem 6(ii), we gave codes with minimum distance three and the radius was two. Recall that for the Levenshtein's channel problem, we have $t > \lfloor (d_{\min}(C) - 1)/2 \rfloor$. In the next theorem, we consider codes in the case where the minimum distance is three and the radius equals three also. These codes provide $\hat{\mu}_3(n) \leq 5$, which is shown to be optimal in Theorem 13. Moreover, the cardinality of the codes is large as pointed out in Remark 15.

Theorem 12. *We have $\hat{\mu}_3(n) \leq 5$ for $n = 2^{2r} - 1 - p$ where $r \geq 2$ and $0 \leq p \leq n/3 - 5$.*

Proof. Let the radius $t = 3$. Denote by $\mathcal{P}_r$ the punctured Preparata code [1, p. 51] of length $n = 2^{2r} - 1$ where $r \geq 2$. It is well-known that the minimum distance $d_{\min}(\mathcal{P}_r) = 5$ and the covering radius $R(\mathcal{P}_r) = 3$. The code $\mathcal{P}_r$ is non-linear. Let us first determine $m_3(\mathbf{x})$ for those words $\mathbf{x} \in \mathbb{F}^n$ that are accessible (not all are). Since the covering radius is three, we know that $d(\mathbf{x}, \mathcal{P}_r) \leq 3$.

Let first $2 \leq d(\mathbf{x}, \mathcal{P}_r) \leq 3$. Since $\mathcal{P}_r$ is a nearly perfect code [1, p. 313], we have $|I_3(\mathcal{P}_r; \mathbf{x})| = n/3$. Let us consider $\mathbf{h}_3(\mathbf{x}) = (h_1, \ldots, h_n)$. We will see that $h_i \leq 1$ for all $i = 1, \ldots, n$. Indeed, suppose to the contrary that $h_i \geq 2$ for some i. Consequently, there are (at least) two codewords $\mathbf{c}$ and $\mathbf{c}'$ in $I_3(\mathcal{P}_r; \mathbf{x})$ such that they differ from $\mathbf{x}$ in the coordinate i. But now $d(\mathbf{c}, \mathbf{c}') \leq 4$ and this is a contradiction with $d_{\min}(\mathcal{P}_r) = 5$. Moreover, since $|I_3(\mathcal{P}_r); \mathbf{x})| = n/3$, in the vector $\mathbf{h}_3(\mathbf{x})$ all entries h_i are equal to 1 or exactly one is 0 and the others are 1. Therefore, by (1) and (2), we get $m_3(\mathbf{x}) = 3$.

Let then $0 \leq d(\mathbf{x}, \mathcal{P}_r) \leq 1$. If $\mathbf{x} \in \mathcal{P}_r$ we obtain $\mathbf{h}_3(\mathbf{x}) = (0, \ldots, 0)$ and $|I_3(\mathbf{x})| = 1$ due to the fact that the minimum distance is five. Thus, $m_3(\mathbf{x}) = 1$. If $d(\mathbf{x}, \mathcal{P}_r) = 1$, then $\mathbf{x}$ is *not* accessible (and $m_3(\mathbf{x})$ does not exist), since $I_3(\mathcal{P}_r; \mathbf{x}) = \{\mathbf{c}\}$ where $\mathbf{x} \neq \mathbf{c}$ and $\mathbf{h}_3(\mathbf{x})$ contains zeros except 1 in the position where $\mathbf{x}$ and $\mathbf{c}$ differ. Hence the code $\mathcal{P}_r$ is not 3-revealing.

As we saw, there are three types of words in $\mathbb{F}^n$ with respect to the code $\mathcal{P}_r$. Those words which have $m_3(\mathbf{x}) = 3$ and $|I_3(\mathbf{x})| = n/3$ we call *type 3* words. The (code)words with $m_t(\mathbf{x}) = 1$ and $|I_3(\mathbf{x})| = 1$ are called *type 1* words. The rest of the words (the non-accessible ones) are of *type 0*.

In order to find a 3-revealing code we take advantage of the additive properties mentioned above and consider the code $C = \mathcal{P}_r \cup (\mathbf{g} + \mathcal{P}_r)$ where $\mathbf{g}$ is a word of weight three such that $d(\mathbf{g}, \mathcal{P}_r) = 3$

(for such words, see [14, p. 475]). Due to the fact that $d_{\min}(\mathcal{P}_r) = 5$ we have $\mathcal{P}_r \cap (\mathbf{g} + \mathcal{P}_r) = \emptyset$, so we can use the additive properties. By [14, p. 475], we know that $d_{\min}(C) = 3$.

Next we estimate $m_3(C; \mathbf{y})$ for $\mathbf{y} \in \mathbb{F}^n$ (as we will see, all words will be accessible with respect to C) by considering the different types of the words. We will make use of Theorem 5(i) — namely, if $\mathbf{g}+\mathbf{y}$ is accessible in $\mathcal{P}_r$, then $\mathbf{y}$ is accessible in $\mathbf{g}+\mathcal{P}_r$ and, moreover, $m_3(\mathbf{g}+\mathcal{P}_r; \mathbf{y}) = m_3(\mathcal{P}_r; \mathbf{g}+\mathbf{y})$. In addition, $|I_3(\mathbf{g} + \mathcal{P}_r; \mathbf{y})| = |I_3(\mathcal{P}_r; \mathbf{g} + \mathbf{y})|$. If $\mathbf{g} + \mathbf{y}$ is non-accessible in $\mathcal{P}_r$, then we have $|I_3(\mathbf{g} + \mathcal{P}_r; \mathbf{y})| = |I_3(\mathcal{P}_r; \mathbf{g} + \mathbf{y})|$ and $\mathbf{h}_t(\mathbf{g} + \mathcal{P}_r; \mathbf{y}) = \mathbf{h}_t(\mathcal{P}_r; \mathbf{g} + \mathbf{y})$. Thus, the words in $\mathbb{F}^n$ have the same three types with respect the code $\mathbf{g} + \mathcal{P}_r$ as they had in the code $\mathcal{P}_r$.

If a word $\mathbf{y}$ is of type 3 in $\mathcal{P}_r$ and also of type 3 in $\mathbf{g} + \mathcal{P}_r$, then by the additive properties we get $|I_3(C; \mathbf{y})| = |I_3(\mathcal{P}_r; \mathbf{y})| + |I_3(\mathbf{g} + \mathcal{P}_r; \mathbf{y})| = 2n/3$ and $\mathbf{h}_3(C; \mathbf{y}) = \mathbf{h}_3(\mathcal{P}_r; \mathbf{y}) + \mathbf{h}_3(\mathbf{g} + \mathcal{P}_r; \mathbf{y}) = (h_1, \ldots, h_n)$, where the maximal h_i is equal to 2. Consequently, $m_3(C; \mathbf{y}) = 5$.

Suppose next that $\mathbf{y}$ is of type 3 in $\mathcal{P}_r$ and it is of type 0 or 1 in $\mathbf{g} + \mathcal{P}_r$. In this case, we have $|I_3(C; \mathbf{y})| = n/3 + 1$ and $\mathbf{h}_3(C; \mathbf{y}) = (h_1, \ldots, h_n)$ where $h_i \leq 2$ for $i = 1, \ldots, n$. Thus, we get $m_3(C; \mathbf{x}) \leq 5$. The same is true if $\mathbf{y}$ is of type 3 in $\mathbf{g} + \mathcal{P}_r$ it is of type 0 or 1 in $\mathcal{P}_r$.

Now the only possibility left to be studied is when $\mathbf{y}$ is of type 0 or 1 in both of the subcodes of C. This means, by the definition of types 0 and 1, that there would be codewords $\mathbf{c} \in \mathcal{P}_r$ and $\mathbf{g} + \mathbf{c}' \in \mathbf{g} + \mathcal{P}_r$ such that $d(\mathbf{y}, \mathbf{c}) \leq 1$ and $d(\mathbf{y}, \mathbf{g} + \mathbf{c}') \leq 1$. But then, by the triangle inequality, we get $d(\mathbf{c}, \mathbf{g} + \mathbf{c}') \leq 2$, which contradicts the fact that $d_{\min}(C) = 3$. Therefore, there does not exist such a possibility for the word $\mathbf{y}$. Consequently, all the words are accessible and C is 3-revealing with the parameter $\hat{\mu}_3(C) \leq 5$. Hence $\hat{\mu}_3(n) \leq 5$ for $n = 2^{2r} - 1$, $r \geq 2$.

In order to get the result for the lengths $n - p$, where $0 < p \leq n/3 - 5$, we use the shortening of Theorem 8. Notice that there are two classes of word in $\mathbb{F}^n$ with respect to C. Those with $|I_3(C; \mathbf{y})| = 2n/3$ and those with $|I_3(C; \mathbf{y})| = n/3 + 1$. In both cases, $m_3(C; \mathbf{y}) \leq 5$.

We need to show that shortening the code C will still leave enough (that is, at least five) codewords to $I_3(C_p; \mathbf{x})$ for any $\mathbf{x} \in \mathbb{F}^{n-p}$. If $\mathbf{x}$ is such that the word $\mathbf{x}0^p$ has $|I_3(C; \mathbf{x}0^p)| = 2n/3$, then in the vector $\mathbf{h}_3(\mathbf{x}0^p)$ we have $h_i \leq 2$ for all $i = n - p + 1, \ldots n$. On the other hand, if $\mathbf{x}0^p$ is such $|I_3(C; \mathbf{c})| = n/3 + 1$, then all h_i's, where $i = n - p + 1, \ldots, n$, are equal to 1 except maybe one which is at most two. Since $p \leq n/3 - 5$, the condition (6) is satisfied and C_p is 3-revealing with $\hat{\mu}_3(C_p) \leq 5$. □

The result $\hat{\mu}_3(n) \leq 5$ found in the previous theorem is actually optimal for $t = 3$ as will be seen next.

Theorem 13. *For $t \geq 3$ and $n \geq 5$ we have $\hat{\mu}_t(n) \geq 5$.*

Proof. Let C be a t-revealing code in $\mathbb{F}^n$ with $t \geq 3$ and $n \geq 5$. Take a word $\mathbf{y} \in \mathbb{F}^n$ such that $d(\mathbf{y}, \mathbf{c}) = 1$ for some $\mathbf{c} \in C$. Clearly, $m_t(\mathbf{y}) \geq 3$, and if $m_t(\mathbf{y}) > 3$ we are done, because $m_t(\mathbf{y})$ is always odd (see (2)). Suppose then that $m_t(\mathbf{y}) = 3$. Clearly, $|I_t(\mathbf{y})| \geq 3$. Let us consider three codewords $\mathbf{c}, \mathbf{c}_1, \mathbf{c}_2 \in I_t(\mathbf{y})$. If $\mathrm{supp}(\mathbf{c}_1) \cap \mathrm{supp}(\mathbf{c}_2) = \emptyset$, then a word $\mathbf{z} = \mathbf{y} + \mathbf{e}_i + \mathbf{e}_j$ such that $i \in \mathrm{supp}(\mathbf{c}_1)$ and $j \in \mathrm{supp}(\mathbf{c}_2)$ has distance at most t to all the three codewords. Consequently, $\mathbf{c}, \mathbf{c}_1, \mathbf{c}_2 \in I_t(\mathbf{z})$ and, by (3), $m_t(\mathbf{z}) \geq 4$ and we are done because $m_t(\mathbf{z})$ is odd. Suppose then that $\mathrm{supp}(\mathbf{c}_1) \cap \mathrm{supp}(\mathbf{c}_2)$ is non-empty and contains, say, the index i. Then $\mathbf{z} = \mathbf{y} + \mathbf{e}_i$ gives $\mathbf{c}, \mathbf{c}_1, \mathbf{c}_2 \in I_t(\mathbf{z})$ and again $m_t(\mathbf{z}) \geq 5$. In summary, $\hat{\mu}_t(C) \geq 5$, which yields the assertion. □

Remark 14. For the radius $t = 3$, the construction in Theorem 12 gives $\mathcal{L} = 4$ for the length of the list decoder and $\hat{m} = 5$ for the information retrieval. In earlier constructions, the best results [9] for $t = 3$ are $\mathcal{L} = 6$ and $\hat{m} = 7$. Recall that these results on the list decoding are for the case when we use only two channels, $N = 2$. If we would like to find the transmitted word uniquely [13] (that is, $\mathcal{L} = 1$) we would need as many as $N = 6n - 9$ channels to do that when the minimum distance is three as for the codes in Theorem 12.

Remark 15. The upper bound of Theorem 9 for the maximal size of intersection $\mathcal{L}' = 4$ and for lengths $n = 2^{2r} - 1$ equals

$$|C| \leq \frac{2^{4^r+1}}{16^r - 3 \cdot 4^r + 4}.$$

The codes of length $n = 2^{2r} - 1$ in Theorem 12 give $\mathcal{L}' = \hat{\mu}_t(C) - 1 = 4$ with the cardinality $2^{4^r - 4r + 1}$ (twice the cardinality of the punctured Preparata code 2^{n-4r+1}, see [1, p. 313]). The ratio between the cardinality of these codes and the upper bound approaches to 1 as r tends to infinity. Therefore, these codes are good also in this respect for the Levenshtein's list decoding problem.

4 Results for other radii

In this section, we show how to get from the results of the previous sections bounds on large values of the radius t. We also consider 4-revealing codes.

4.1 Large radii

In this section, we will study how to get $(n-t-1)$-revealing codes from t-revealing ones.

Theorem 16. *Let $C \subseteq \mathbb{F}^n$ be such a t-revealing code that each coordinate has 0 in exactly half of the codewords in any given coordinate. Then C is also $(n-t-1)$-revealing with*

$$m_{n-t-1}(\mathbf{x}) = |C| - 2|I_t(\bar{\mathbf{x}})| + m_t(\bar{\mathbf{x}})$$

for all $\mathbf{x} \in \mathbb{F}^n$.

Proof. The words in $\mathbb{F}^n$ can be partitioned into two balls, namely, $B_{n-t-1}(\mathbf{x})$ and $B_t(\bar{\mathbf{x}})$ for any $\mathbf{x} \in \mathbb{F}^n$. Consequently, $I_{n-t-1}(\mathbf{x}) = C \setminus I_t(\bar{\mathbf{x}})$ and hence $|I_{n-t-1}(\mathbf{x})| = |C| - |I_t(\bar{\mathbf{x}})|$. Let $\mathbf{h}_{n-t-1}(\mathbf{x}) = (h_1, \ldots, h_n)$ and $\mathbf{h}_t(\bar{\mathbf{x}}) = (h'_1, \ldots, h'_n)$. There is the relation

$$h_i = \frac{|C|}{2} - |I_t(\bar{\mathbf{x}})| + h'_i \qquad \forall i = 1, \ldots, n$$

due to the fact that there are all in all $|C|/2$ codewords with coordinates $c_i \neq x_i$ in C and $|I_t(\bar{\mathbf{x}})| - h'_i$ of them are in $I_t(\bar{\mathbf{x}})$. The condition (1) for the radius $n-t-1$ is now satisfied. Indeed, by combining

$$2 \max_{j=1,\ldots,n} h_j + 1 = |C| - 2|I_t(\bar{\mathbf{x}})| + 2 \max_{j=1,\ldots,n} h'_j + 1$$

with the inequality $|I_t(\bar{\mathbf{x}})| \geq 2 \max_{j=1,\ldots,n} h'_j + 1$ (which is true because C is t-revealing), we get $|I_{n-t-1}(\mathbf{x})| \geq 2 \max_{j=1,\ldots,n} h_j + 1$. Consequently, $m_{n-t-1}(\mathbf{x}) = |C| - 2|I_t(\bar{\mathbf{x}})| + m_t(\bar{\mathbf{x}})$. □

Remark 17. If we do not have the property that there are equal number of 0's and 1's in each of the coordinates of C, then a t-revealing code may not be $(n-t-1)$-revealing. Indeed, the code $C = \{\mathbf{0}\} \cup (\mathbb{F}^6 \setminus (B_2(\mathbf{0}))$ of length 6 is 2-revealing, but it is not 3-revealing. The cardinality of C equals 43, so there cannot be equal amount of 0's and 1's in any coordinate.

The *distance distribution* of a code C is the $(n+1)$-tuple $(B_0, B_1, \ldots, B_n)$ where

$$B_i = \frac{1}{|C|} |\{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x}, \mathbf{y} \in C, d(\mathbf{x}, \mathbf{y}) = i\}|.$$

The *MacWilliams transform* of the distance distribution is the $(n+1)$-tuple $(B_0^\perp, B_1^\perp, \ldots, B_n^\perp)$ with

$$B_s^\perp = \frac{1}{|C|} \sum_{i=0}^{n} B_i K_s(i), \qquad s = 0, 1, \ldots, n, \tag{9}$$

where the Krawtchouk polynomial of degree s

$$K_s(x) = \sum_{j=0}^{s} (-1)^j \binom{x}{j} \binom{n-x}{s-j}.$$

The *dual distance* $d^\perp$ of a code is defined as the non-zero index i such that $B_i^\perp \neq 0$ and $B_j^\perp = 0$ for indices $0 < j < i$. If the code C is linear, then the MacWilliams transform of the distance distribution of C gives the distance distribution of the dual code $C^\perp$ (see [1, p. 25] and thus $d^\perp = d_{\min}(C^\perp)$.

Let us recall a fact related to the dual distance [14, p. 139]: Let $d^\perp$ be the dual distance of a code $C \subseteq \mathbb{F}^n$ (not necessarily linear). Then any set of $j \leq d^\perp - 1$ coordinates in C contains each j-tuple exactly $|C|/2^j$ times.

Using the codes from the previous sections, we get the following corollary.

Corollary 18. *We have*

(i) $\hat{\mu}_{n-2}(n) \leq 2^{n-r} - 1$ *for all* $n = 3(2^r - 1) + s$ *where* $r \geq 1$ *and* $0 \leq s \leq 3 \cdot 2^r - 1$

(ii) $\hat{\mu}_{n-3}(n) \leq 2^{n-r} - 1$ *for all* $n = 2^r - 1 - p$ *where* $r \geq 3$ *and* $0 \leq p \leq 2^{r-1} - 3$.

(iii) $\hat{\mu}_{n-4}(n) \leq 2^{n-4r+1} - 2(n/3 + 1) + 5$ *for all* $n = 2^{2r} - 1$, $r \geq 2$.

Proof. (i) Linear codes have the property that either all codewords or exactly half of the codewords have 0 in any given coordinate. Clearly, the latter is true for the codes in the proof of Theorem 6(i) of radius one. Therefore, we can apply Theorem 16. For these codes we have for codewords $|I_1(\mathbf{x})| = 1$ and $m_1(\mathbf{x}) = 1$ and for non-codewords $|I_1(\mathbf{x})| \geq 3$ and $m_1(\mathbf{x}) = 3$. Now, using Theorem 16, we get $\hat{\mu}_{n-2}(C) = \max_{\mathbf{x} \in \mathbb{F}^n} m_{n-t-1}(\mathbf{x}) = |C| - 1 = 2^{n-r} - 1$.

(ii) Similarly, as above, we get for $t = n - 3$ the result when we use the codes in the proof of Theorem 6(ii) of radius two together with Theorem 16.

(iii) For $t = n - 4$ we use the codes from Theorem 12. Since the dual distance of the punctured Preparata code $\mathcal{P}_r$ is greater than two [14, p. 472], there are, by the fact mentioned above, in the codewords equal amount of 0's and 1's in any of the coordinates. The same is true for the union $\mathcal{P}_r \cup (\mathbf{x} + \mathcal{P}_r)$. Therefore, we can again utilize Theorem 16. □

Notice that although results (i) and (ii) seem similar they are not the same. For example, when $n = 10$, the first bound is $\hat{\mu}_8(10) \leq 255$ and the second one $\hat{\mu}_7(10) \leq 63$.

Next we will provide a lower bound on $\hat{\mu}_t(n)$ which is useful when the radius t is large compared to n.

Theorem 19. *We have*

$$\hat{\mu}_t(n) \geq \max\left\{\frac{3}{V(n,t)+2}, \frac{1}{V(n,n-t-1)} - 1\right\}\left(V(n,t) - \binom{n-1}{t}\right) + 1.$$

Proof. Let C be a t-revealing code. Besides the lower bound (8) in Theorem 9, we can give another lower bound on $|C|$. Due to (4) we know that a t-revealing code is also an t-identifying code. For the smallest possible size of an t-identifying codes in $\mathbb{F}^n$ (denoted by $M_t(n)$) we get by [5]

$$M_t(n) \geq \frac{2^n}{V(n,n-t-1)} - 1.$$

Combining this and (8) with the upper bound (7), we obtain the claim. □

4.2 Radius $t = 4$

For the radius $t = 4$ and codes in $\mathbb{F}^n$, the best known upper bound on the parameters $\mathcal{L}$ and $\hat{m}$ is of order n^3 (see [7]). In this section, we show that there are codes giving an upper bound of (linear) order n.

Let X be a set with v elements. A *t-design* is a collection of distinct subsets of k elements (called *blocks*) of X with the property that any t-subset of X is contained in exactly λ blocks. Denote the number of blocks by b. Each element of X occurs in r blocks [14, p. 60] and, for a 2-design, there are the relations

$$bk = vr \tag{10}$$

and

$$\lambda(v-1) = r(k-1). \tag{11}$$

For a linear code, the distance distribution and the weight distribution coincide [1, p. 25]. The weight distribution $(A_0, A_1, \ldots, A_n)$ of a code C is thus linked to the weight distribution of the dual code $(A_0^\perp, A_1^\perp, \ldots, A_n^\perp)$ via (9). The number of subscripts $i > 0$ such that $A_i^\perp \neq 0$ is called the *external distance* s' of a code. Let us recall a fact about designs obtained from codewords [14, p. 175]: Let C be a code with $d_{\min}(C) = d$ and the external distance s'. Then the codewords of any fixed weight in a code C form a $(d - s')$-design provided that $d - s' \leq s' < d$.

In the next proof, we will use the following observation. If $\mathbf{h}_t(\mathbf{x}) = (h_1, \ldots, h_n)$ for $\mathbf{x} \in \mathbb{F}^n$ and a code $C \subseteq \mathbb{F}^n$, then (by the definition of h_i)

$$h_i \leq |I_{t-1}(\mathbf{x} + \mathbf{e}_i)|, \qquad i = 1, \ldots, n. \tag{12}$$

Theorem 20. *We have $\hat{\mu}_4(n) \leq (n-1)/3 + 1$ for $n = 2^{2r+1} - 1$ when $r > 2$.*

Proof. Let us consider the radius $t = 4$. Denote by $\mathcal{C}_r$ a double-error-correcting binary, narrow-sense and primitive BCH-code [14, p. 202] of length $n = 2^{2r+1} - 1$, $r > 2$. The code $\mathcal{C}_r$ is linear, and moreover, $R(\mathcal{C}_r) = 3$ and $d_{\min}(\mathcal{C}_r) = 5$. The code is also strongly uniformly packed [1, p. 313]— a word at distance two or three from the code contains exactly $(n-1)/6$ codewords within distance three. Clearly, the words at distance less than two from the code have exactly 1 codeword within distance three. We shall also need the weight distribution of the dual code of $\mathcal{C}_r$ (see [14, p. 451]): $A_0^\perp = 1$, $A_{2^{2r}-2^r}^\perp = (2^{2r+1}-1)(2^{2r-1}+2^{r-1})$, $A_{2^{2r}}^\perp = (2^{2r+1}-1)(2^{2r}+1)$, $A_{2^{2r}+2^r}^\perp = (2^{2r+1}-1)(2^{2r-1}-2^{r-1})$ and other $A_i^\perp$'s are zeros.

Since the code is linear, it suffices to determine $m_4(C; \mathbf{x})$ for coset leaders $\mathbf{x}$. Due to the covering radius, the weight of a coset leader is at most three. Let the coset leader be $\mathbf{0}$. We have $I_4(\mathbf{0}) = \{\mathbf{0}\}$ because the minimum distance is five. Consequently, $m_4(\mathbf{0}) = 1$. Suppose then that the coset leader $\mathbf{x}$ has $1 \leq w(\mathbf{x}) \leq 3$ and $\mathbf{h}_4(\mathbf{x}) = (h_1, \ldots, h_n)$. Utilizing the observation (12) we can estimate $h_i \leq |I_3(\mathbf{x} + \mathbf{e}_i)|$ and, as discussed above, we obtain

$$h_i \leq (n-1)/6.$$

In order to prove that for a coset leader $\mathbf{x}$, $\mathbf{x} \neq \mathbf{0}$,

$$m_4(\mathbf{x}) \leq \frac{n-1}{3} + 1$$

it now suffices to show, due to (1) and (2), that

$$|I_4(\mathbf{x})| \geq (n-1)/3 + 1. \tag{13}$$

We divide the investigation according to the weight of the coset leader.

Assume first that $w(\mathbf{x}) = 1$, say $\mathbf{x} = \mathbf{e}_i$. We will benefit from the fact that all the words of $\mathcal{C}_r$ form a 2-design (since $d_{\min}(\mathcal{C}_r) = 5$ and $s' = 3$). The set $I_4(\mathbf{x})$ for $\mathbf{x} = \mathbf{e}_i$ consists of the such codewords of weight five that have 1 in the position i. Consequently, using the relation (10), we get $|I_4(\mathbf{x})| = 5 \cdot A_5/n$ where, due to the weight distribution above of the dual code and (9),

$$A_5 = \frac{1}{120}\left(8^{2r+1} - 11 \cdot 4^{2r+1} + 13 \cdot 2^{2r+2} - 16\right).$$

It is straightforward to check that (13) is satisfied.

Assume then that $w(\mathbf{x}) = 2$ and let $\mathrm{supp}(\mathbf{x}) = \{i, j\}$. The set $I_4(\mathbf{x})$ contains such words of weight six which have i and j in their support (there are also codewords of weight five in the set, but we do not need to consider them). The words of weight six form also a 2-design. Therefore, using (10) and (11), we get $|I_4(\mathbf{x})| \geq 6 \cdot 5 \cdot A_6/(n(n-1))$ where, due to (9),

$$A_6 = \frac{1}{720}\left(16^{2r+1} - 17 \cdot 8^{2r+1} + 23 \cdot 4^{2r+2} - 43 \cdot 2^{2r+3} + 96\right).$$

Thus, it is easy to verify that (13) holds.

Finally, we need to consider the case where the coset leader $\mathbf{x}$ has weight three. Let $\mathrm{supp}(\mathbf{x}) = \{i, j, k\}$. Denote by $\mathbf{y}_1$ (resp. $\mathbf{y}_2$ and $\mathbf{y}_3$) the word with $\mathrm{supp}(\mathbf{y}_1) = \{i, j\}$ (resp. $\mathrm{supp}(\mathbf{y}_2) = \{i, k\}$ and $\mathrm{supp}(\mathbf{y}_3) = \{j, k\}$). Clearly, each $I_3(\mathbf{y}_s) \subseteq I_4(\mathbf{x})$ for $s = 1, 2, 3$. We show that we can find for (13) enough distinct words from the sets $I_3(\mathbf{y}_s)$, $s = 1, 2, 3$. Since $w(\mathbf{y}_s) = 2$ for $s = 1, 2, 3$, we know that $|I_3(\mathbf{y}_s)| = (n-1)/6$. Moreover, since $d_{\min}(\mathcal{C}_r) = 5$, for any two codewords c_1 and c_2 in $I_3(\mathbf{y}_s)$ have $\mathrm{supp}(c_1) \cap \mathrm{supp}(c_2) = \emptyset$. Clearly, $\mathbf{0} \in I_3(\mathbf{y}_s)$, $s = 1, 2, 3$. We consider those codewords in $I_3(\mathbf{y}_1)$ whose k-th coordinate equals zero. Besides $\mathbf{0}$, there are no such codewords in $I_3(\mathbf{y}_2)$ or in $I_3(\mathbf{y}_3)$. Consequently, the set $I_3(\mathbf{y}_1)$ give at least $(n-1)/6 - 2$ codewords to $I_4(\mathbf{x})$ which are not in $I_3(\mathbf{y}_s)$, $s = 2, 3$, and the k-th coordinate equals zero. Analogously, we can reason for the sets $I_3(\mathbf{y}_s)$, $s = 2, 3$. Therefore, we have (including now $\mathbf{0}$) that

$$|I_4(\mathbf{x})| \geq 3 \cdot ((n-1)/6 - 2) + 1.$$

The proof of the assertion is now completed by noticing that again (1) is satisfied. □

Acknowledgement: The author would like to thank the referee for useful comments.

References

[1] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein. *Covering codes*, volume 54 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, 1997.

[2] N. Fazlollahi, D. Starobinski, and A. Trachtenberg. Connected identifying codes. *IEEE Trans. Inform. Theory*, 58(7):4814–4824, 2012.

[3] S. Gravier, A. Parreau, S. Rottey, L. Storme, and É. Vandomme. Identifying codes in vertex-transitive graphs and strongly regular graphs. *Electron. J. Combin.*, 22(4):Paper 4.6, 26, 2015.

[4] I. Honkala and T. Laihonen. On a new class of identifying codes in graphs. *Inform. Process. Lett.*, 102(2-3):92–98, 2007.

[5] I. Honkala and A. Lobstein. On identifying codes in binary Hamming spaces. *J. Combin. Theory Ser. A*, 99(2):232–243, 2002.

[6] O. Hudry and A. Lobstein. More results on the complexity of identifying problems in graphs. *Theoret. Comput. Sci.*, 626:1–12, 2016.

[7] V. Junnila and T. Laihonen. Codes for information retrieval with small uncertainty. *IEEE Trans. Inform. Theory*, 60(2):976–985, 2014.

[8] V. Junnila and T. Laihonen. Information retrieval with unambiguous output. *Inform. and Comput.*, 242:354–368, 2015.

[9] V. Junnila and T. Laihonen. Information retrieval with varying number of input clues. *IEEE Trans. Inform. Theory*, 62(2):625–638, 2016.

[10] V. Junnila and T. Laihonen. Minimum number of input clues in robust information retrieval. *Fund. Inform.*, 145(3):243–256, 2016.

[11] M. G. Karpovsky, K. Chakrabarty, and L. B. Levitin. On a new class of codes for identifying vertices in graphs. *IEEE Trans. Inform. Theory*, 44(2):599–611, 1998.

[12] T. Laihonen. Information retrieval and the average number of input clues. *Adv. Math. Commun.*, 11(1):203–223, 2017.

[13] V. I. Levenshtein. Efficient reconstruction of sequences. *IEEE Trans. Inform. Theory*, 47(1):2–22, 2001.

[14] F. J. MacWilliams and N. J. A. Sloane. *The theory of error-correcting codes*, volume 16 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, 1977.

[15] K. Thulasiraman, M. Su, Y. Xiao, and X. Hu. Vertex identifying codes for fault isolation in communication networks. In *Proceedings of the International Conference on Discrete Mathematics and Applications (ICDM 2006)*, Bangalore, 2006.

[16] E. Yaakobi and J. Bruck. On the uncertainty of information retrieval in associative memories. In *Proceedings of 2012 IEEE International Symposium on Information Theory*, pages 106–110, 2012.

[17] E. Yaakobi, J. Bruck, and P. Siegel. Constructions and decoding of cyclic codes over b-symbol read channels. *IEEE Trans. Inform. Theory*, 62(4):1541–1551, 2016.

[18] E. Yaakobi, M. Schwartz, M. Langberg, and J. Bruck. Sequence reconstruction for grassmann graphs and permutations. In *Proceedings of 2013 IEEE International Symposium on Information Theory*, pages 874–878, 2013.