On the Size of Identifying Codes in Binary Hypercubes

Svante Janson
Department of Mathematics
Uppsala University
P.O. Box 480
SE-751 06 Uppsala, Sweden
and
Tero Laihonen*
Department of Mathematics
University of Turku
FIN-20014 Turku, Finland

18 April 2008

Abstract

In this paper, we consider identifying codes in binary Hamming spaces $\mathbb{F}^n$, i.e., in binary hypercubes. The concept of identifying codes was introduced by Karpovsky, Chakrabarty and Levitin in 1998. Currently, the subject forms a topic of its own with several possible applications, for example, to sensor networks.

Let $C \subseteq \mathbb{F}^n$. For any $X \subseteq \mathbb{F}^n$, denote by $I_r(X) = I_r(C; X)$ the set of elements of C within distance r from at least one $x \in X$. Now $C \subseteq \mathbb{F}^n$ is called an $(r, \leq \ell)$-identifying code if the sets $I_r(X)$ are distinct for all $X \subseteq \mathbb{F}^n$ of size at most ℓ. Let us denote by $M_r^{(\leq \ell)}(n)$ the smallest possible cardinality of an $(r, \leq \ell)$-identifying code. In [14], it is shown for $\ell = 1$ that

$$\lim_{n\to\infty} \frac{1}{n} \log_2 M_r^{(\leq \ell)}(n) = 1 - h(\rho)$$

where $r = \lfloor \rho n \rfloor$, $\rho \in [0, 1)$ and $h(x)$ is the binary entropy function. In this paper, we prove that this result holds for any fixed $\ell \geq 1$ when $\rho \in [0, 1/2)$. We also show that $M_r^{(\leq \ell)}(n) = O(n^{3/2})$ for every fixed ℓ and r slightly less than $n/2$, and give an explicit construction of small $(r, \leq 2)$-identifying codes for $r = \lfloor n/2 \rfloor - 1$.

*Research supported by the Academy of Finland under grant 111940.

1 Introduction

Let $\mathbb{F} = \{0, 1\}$ be the binary field and denote by $\mathbb{F}^n$ the n-fold Cartesian product of it, i.e. the Hamming space. We denote by $A \triangle B$ the *symmetric difference* $(A \setminus B) \cup (B \setminus A)$ of two sets A and B. The *(Hamming) distance* $d(x, y)$ between the vectors (called words) $x, y \in \mathbb{F}^n$ is the number of coordinate places in which they differ, i.e, $x(i) \neq y(i)$ for $i = 1, 2, \dots, n$. The *support* of $x = (x(1), x(2), \dots, x(n)) \in \mathbb{F}^n$ is defined by $\mathrm{supp}(x) = \{i \mid x(i) = 1\}$. The *complement* of a word $x \in \mathbb{F}^n$, denoted by $\overline{x}$, is the word for which $\mathrm{supp}(\overline{x}) = \{1, 2, \dots, n\} \setminus \mathrm{supp}(x)$. Denote by 0 the word where all the coordinates equal zero, and by 1 the all-one word. Clearly $\overline{0} = 1$. The *(Hamming) weight* $w(x)$ of a word $x \in \mathbb{F}^n$ is defined by $w(x) = d(x, 0)$. We say that x *r-covers* y if $d(x, y) \leq r$. The *(Hamming) ball* of radius r centered at $x \in \mathbb{F}^n$ is

$$B_r(x) = \{y \in \mathbb{F}^n \mid d(x, y) \leq r\}.$$

and its cardinality is denoted by $V(n, r)$. For $X \subseteq \mathbb{F}^n$, denote

$$B_r(X) = \bigcup_{x \in X} B_r(x) = \{y \in \mathbb{F}^n \mid d(y, X) \leq r\}.$$

We also use the notation

$$S_r(x) = \{y \in \mathbb{F}^n \mid d(x, y) = r\}.$$

A nonempty subset $C \subseteq \mathbb{F}^n$ is called a *code* and its elements are *codewords*. Let C be a code and $X \subseteq \mathbb{F}^n$. We denote (the codeword r-neighbourhood of X by)

$$I_r(X) = I_r(C; X) = B_r(X) \cap C.$$

We write for short $I_r(C; \{x_1, \dots, x_k\}) = I_r(x_1, \dots, x_k)$.

Definition 1. *Let r and ℓ be non-negative integers. A code $C \subseteq \mathbb{F}^n$ is said to be* $(r, \leq \ell)$-identifying *if for all $X, Y \subseteq \mathbb{F}^n$ such that $|X| \leq \ell$, $|Y| \leq \ell$ and $X \neq Y$ we have*

$$I_r(C; X) \neq I_r(C; Y).$$

The idea of the identifying codes is that given the set $I_r(X)$ we can uniquely determine the set $X \subseteq \mathbb{F}^n$ as long as $|X| \leq \ell$.

The seminal paper [15] by Karpovsky, Chakrabarty and Levitin initiated research in identifying codes, and it is nowadays a topic of its own with different types of problems studied, see, e.g., [2],[4],[5],[6],[11],[12],[20],[22]; for an updated bibliography of identifying codes see [19]. Originally, identifying codes were designed for finding malfunctioning processors in multiprocessor systems (such as binary hypercubes, i.e., binary Hamming spaces); in this application we want to determine the set of malfunctioning processors X of size at most ℓ when the only information available is the set $I_r(C; X)$ provided by the code C. A natural goal there is to use identifying codes which is as small as possible. The theory of identification can also be applied to sensor networks, see [21]. Small identifying codes are needed for energy conservation [16]. For other applications we refer to [17].

The smallest possible cardinality of an $(r, \leq \ell)$-identifying code in $\mathbb{F}^n$ is denoted by $M_r^{(\leq \ell)}(n)$.

Let $h(x) = -x \log_2 x - (1-x) \log_2(1-x)$ be the binary entropy function and $\rho \in [0, 1)$ be a constant. Let further $r = \lfloor \rho n \rfloor$. Honkala and Lobstein showed in [14] that, when $\ell = 1$, we have

$$\lim_{n \to \infty} \frac{1}{n} \log_2 M_r^{(\leq 1)}(n) = 1 - h(\rho). \tag{1}$$

The lower bound that is part of (1) comes from the simple observation that if C is an $(r, \leq \ell)$-identifying code for any $\ell \geq 1$, then necessarily $B_r(C) = \mathbb{F}^n$ (otherwise there would be a word $x \notin B_r(C)$ and then $I_r(x) = \emptyset = I_r(\emptyset)$, so $\{x\}$ and $\emptyset$ cannot be distinguished by C) and also $|\mathbb{F}^n \setminus B_{n-r-1}(C)| \leq 1$ (otherwise there would be two words $x, y \notin B_{n-r-1}(C)$ and then $I_r(\overline{x}) = C = I_r(\overline{y})$, so $\{\overline{x}\}$ and $\{\overline{y}\}$ cannot be distinguished by C); consequently, for any $n, r, \ell \geq 1$,

$$\begin{aligned} M_r^{(\leq \ell)}(n) \geq M_r^{(\leq 1)}(n) &\geq \max\left(\frac{2^n}{|V(n,r)|}, \frac{2^n - 1}{|V(n, n-r-1)|}\right) \\ &= \max\left(\frac{2^n}{\sum_{i=0}^{r} \binom{n}{i}}, \frac{2^n - 1}{\sum_{i=0}^{n-r-1} \binom{n}{i}}\right) \end{aligned} \tag{2}$$

and the lower bound in (1) follows from Stirling's formula. Cf. [7, Chapter 12], [3], [10], [14] and [15] for this and similar arguments and related estimates.

Let us now consider any fixed $\ell > 1$. When $r = \lfloor \rho n \rfloor$, we have by (1) or (2) the same lower bound as for $\ell = 1$:

$$\liminf_{n \to \infty} \frac{1}{n} \log_2 M_r^{(\leq \ell)}(n) \geq 1 - h(\rho). \tag{3}$$

In the opposite direction, it is shown in [10] that

$$\limsup_{n \to \infty} \frac{1}{n} \log_2 M_r^{(\leq \ell)}(n) \leq 1 - (1 - 2\ell\rho) h\Big(\frac{\rho}{1 - 2\ell\rho}\Big) \tag{4}$$

where $0 \leq \rho \leq 1/(2\ell + 1)$. In this paper, we improve (4) by showing that the lower bound (3) is attained for any fixed $\ell \geq 1$ when $\rho \in [0, 1/2)$. (The proof is given in Section 2.)

Theorem 1. *Let $\ell \geq 1$ be fixed, let $\rho \in [0, 1/2)$ and assume that $r/n \to \rho$. Then*

$$\lim_{n \to \infty} \frac{1}{n} \log_2 M_r^{(\leq \ell)}(n) = 1 - h(\rho).$$

Furthermore, it is easy to see that when $\ell \geq 2$, unlike the case $\ell = 1$, no $(r, \leq \ell)$-identifying codes at all exist for $r \geq \lfloor n/2 \rfloor$. (This explains why we have to assume $\rho < 1/2$ in Theorem 1.)

Theorem 2. *If $n \geq 2$ and $r \geq \lfloor n/2 \rfloor$, then there does not exist an $(r, \leq 2)$-identifying code in $\mathbb{F}^n$.*

Proof. If $r \geq \lfloor n/2 \rfloor$, then $B_r(x) \cup B_r(\overline{x}) = \mathbb{F}^n$ and thus $I_r(x, \overline{x}) = I_r(y, \overline{y})$ for any $C \subseteq \mathbb{F}^n$ and $x, y \in \mathbb{F}^n$. $\square$

We give this theorem mainly because the proof is so simple. In fact, it is proved in [18] that any $(r, \leq \ell)$-identifying code in $\mathbb{F}^n$ has to satisfy

$$r \leq \lfloor n/2 \rfloor + 2 - \ell, \tag{5}$$

which is slightly better than Theorem 2 when $\ell > 3$.

Since $h(\rho) < 1$ unless $\rho = 1/2$, Theorem 1 implies that an $(r, \leq \ell)$-identifying code has to be exponentially large unless r is close to $n/2$. We give in Section 3 an explicit construction of a small $(r, \leq 2)$-identifying code for the largest possible r permitted by Theorem 2, viz. $r = \lfloor n/2 \rfloor - 1$.

Theorem 3. *Let $n \geq 2$. There exists an $(r, \leq 2)$-identifying code in $\mathbb{F}^n$ of size at most $n^3 - n^2$ when $r = \lfloor n/2 \rfloor - 1$.*

For comparison, it is shown in [14] that for $\ell = 1$ and $n \geq 3$,

$$M^{(\leq 1)}_{\lfloor n/2 \rfloor}(n) \leq \begin{cases} \frac{n^2-n+2}{2}, & n \text{ odd}, \\ \frac{n^2-4}{2}, & n \text{ even}. \end{cases}$$

For $\ell > 2$, we do not know any explicit constructions of small $(r, \leq \ell)$-identifying codes in $\mathbb{F}^n$, but we can show the existence of small such codes (even smaller than the one provided by Theorem 3) for every $\ell \geq 1$ when r is a little smaller than $n/2$. For $\ell = 1$, there exist by the explicit estimate in [10, Corollary 13] $(r, \leq 1)$-identifying codes in $\mathbb{F}^n$ of size $O(n^{3/2})$ for every $r < n/2$ with $r = n/2 - O(\sqrt{n})$. Our next theorem, proved in Section 2, yields a bound of the same order (although less explicit) for every fixed ℓ and certain r.

Theorem 4. *Let $\ell \geq 1$ be fixed and let $0 < a < b$. Then there exist n_0 and A such that for every $n \geq n_0$ and r with $n/2 - b\sqrt{n} \leq r \leq n/2 - a\sqrt{n}$,*

$$M^{(\leq \ell)}_r(n) \leq An^{3/2}.$$

Remark 1. This is not far from the best possible, since an $(r, \leq \ell)$-identifying code C in $\mathbb{F}^n$ trivially must satisfy $\sum_{i=0}^{\ell} \binom{2^n}{i} \leq 2^{|C|}$ and in particular $2^n < 2^{|C|}$; thus $M^{(\leq \ell)}_r(n) > n$ for r and $\ell \geq 1$. (Moreover, this argument yields $M^{(\leq \ell)}_r(n) \geq \ell n - O(1)$ for every fixed $\ell \geq 1$.)

For r closer to $n/2$, we can show a weaker result, still with a polynomial bound. (This theorem too is proved in Section 2.)

Theorem 5. *Let $\ell \geq 1$ be fixed and let L be fixed with $L \geq 2^\ell$. Then there exist n_0 and A such that for every $n \geq n_0$ and r with $r = \lfloor n/2 \rfloor - L$,*

$$M^{(\leq \ell)}_r(n) \leq An^{2^{l-1}+1}.$$

For $\ell \geq 3$, we do not know the largest possible r such that there exists an $(r, \leq \ell)$-identifying code in $\mathbb{F}^n$, but Theorem 5 leaves only a small gap to the bounds in Theorem 2 and (5).

2 Proofs of the main results

Our non-constructive upper bounds in Theorems 1, 4 and 5 are based on the following general theorem proven in [10]. Let $m_n(r,\ell)$ stand for the minimum of $|B_r(X) \triangle B_r(Y)|$ over any subsets $X, Y \subseteq \mathbb{F}^n$, $X \neq Y$ and $1 \leq |X| \leq \ell$ and $1 \leq |Y| \leq \ell$. Denote further by N_ℓ the number of (unordered) pairs $\{X, Y\}$ of subsets of $\mathbb{F}^n$ such that $X \neq Y$ and $1 \leq |X| \leq \ell$ and $1 \leq |Y| \leq \ell$.

Theorem 6 ([10])**.** *Let $r \geq 1$, $\ell \geq 1$ and $n \geq 1$. Provided that $m_n(r,\ell) > 0$, there exists an $(r, \leq \ell)$-identifying code of size K in $\mathbb{F}^n$ such that*

$$K \leq \left\lceil \frac{2^n}{m_n(r,\ell)} \ln N_\ell \right\rceil + 1.$$

Obviously,

$$N_\ell \leq \left(\sum_{i=1}^{\ell} \binom{2^n}{i} \right)^2 \leq 2^{2n\ell}$$

and thus Theorem 6 yields

$$M_r^{(\leq \ell)}(n) \leq \frac{2^{n+1}\ell n}{m_n(r,\ell)} + 2. \tag{6}$$

It remains to estimate $m_n(r,\ell)$. Using probabilistic arguments, we are able to show in Theorems 7 and 8 the following crucial result: for fixed ℓ and $\epsilon > 0$ or a, b, there exists constants n_0 and $c > 0$ such that for $n \geq n_0$ and r with $r \leq (1/2 - \epsilon)n$, or $n/2 - b\sqrt{n} \leq r \leq n/2 - a\sqrt{n}$, we have

$$m_n(r,\ell) \geq c\binom{n}{r}. \tag{7}$$

(By combing the methods of proofs below, it is possible to show that this holds in the intermediate range of r too, but we omit the details.) For ℓ fixed and $r = \lfloor n/2 \rfloor - L$, with $L \geq 2^\ell$ fixed, we prove in Theorem 9 the slightly weaker estimate

$$m_n(r,\ell) \geq cn^{-2^{\ell-1}} 2^n. \tag{8}$$

(We do not know whether (7) holds in this case too.) Combining (6), (7), (8) and standard estimates for binomial coefficients, see [7, p. 33], we obtain Theorems 1, 4 and 5.

We prove the required estimates of $m_n(r,\ell)$ in the following form. In applying the following results to obtain the bounds (7) and (8) on $m_n(r,\ell)$ just notice that we can assume that there is $x \in X \setminus Y$ and $Y \subseteq \{y_1, \ldots, y_\ell\}$.

Theorem 7. *Let $\ell \geq 1$ be fixed. For every $\varepsilon > 0$ there is a constant $c > 0$ and n_0 such that for $n \geq n_0$ and any $\ell + 1$ words x and $y_1, \ldots, y_\ell$ in $\mathbb{F}^n$, with $y_i \neq x$ for $i = 1, \ldots, \ell$, and every r with $0 \leq r \leq (1/2 - \varepsilon)n$, there exist at least $c\binom{n}{r}$ words $z \in \mathbb{F}^n$ with $d(z, x) = r$ and $d(z, y_i) > r$ for $i = 1, \ldots, \ell$.*

Theorem 8. *Let $\ell \geq 1$ be fixed. For every $a, b > 0$ there is a constant $c > 0$ and n_0 such that for $n \geq n_0$ and any $\ell + 1$ words x and $y_1, \ldots, y_\ell$ in $\mathbb{F}^n$, with $y_i \neq x$ for $i = 1, \ldots, \ell$, and every r with $n/2 - b\sqrt{n} \leq r \leq n/2 - a\sqrt{n}$, there exist at least $cn^{-1/2}2^n \geq c\binom{n}{r}$ words $z \in \mathbb{F}^n$ with $d(z, x) = r$ and $d(z, y_i) > r$ for $i = 1, \ldots, \ell$.*

Theorem 9. *Let $\ell \geq 1$ be fixed. For every $L \geq 2^\ell$ there is a constant $c > 0$ and n_0 such that for $n \geq n_0$ and any $\ell + 1$ words x and $y_1, \ldots, y_\ell$ in $\mathbb{F}^n$, with $y_i \neq x$ for $i = 1, \ldots, \ell$, and $r = \lfloor n/2 \rfloor - L$, there exist at least $cn^{-2^{\ell-1}}2^n$ words $z \in \mathbb{F}^n$ with $d(z, x) = r$ and $d(z, y_i) > r$ for $i = 1, \ldots, \ell$.*

The proofs of Theorems 7–9 are similar, although some details differ. We begin with some common considerations.

By symmetry we may assume that $x = 0$. Given $y_1, \ldots, y_\ell$, partition the index set $[n] = \{1, \ldots, n\}$ into 2^ℓ subsets A_α, indexed by $\alpha \in \mathbb{F}^\ell$, such that

$$A_\alpha = \{i \in [n] : y_j(i) = \alpha_j \text{ for } j = 1, \ldots, \ell\}.$$

Let $z \in \mathbb{F}^n$ and let further $s_\alpha = s_\alpha(z) = |\{i \in A_\alpha : z(i) = 1\}|$. Then $d(z, x) = \sum_\alpha s_\alpha$ and

$$d(z, y_j) = \sum_{\alpha:\alpha_j=0} s_\alpha + \sum_{\alpha:\alpha_j=1} (|A_\alpha| - s_\alpha) = d(z, x) + \sum_{\alpha:\alpha_j=1} (|A_\alpha| - 2s_\alpha).$$

Hence, if $d(z, x) = r$, we need also

$$\sum_{\alpha:\alpha_j=1} (|A_\alpha| - 2s_\alpha) \geq 1$$

for each $j = 1, \ldots, \ell$; then $d(z, y_j) > d(z, x) = r$.

For simplicity, we consider only z such that $s_\alpha < |A_\alpha|/2$ for every α such that $A_\alpha \neq \emptyset$; we say that such z's are *good*. Note that $\sum_{\alpha:\alpha_j=1} |A_\alpha| = d(x, y_j) \geq 1$ for each j, so $A_\alpha \neq \emptyset$ for some α with $\alpha_j = 1$, and if z is good, then $\sum_{\alpha:\alpha_j=1}(|A_\alpha| - 2s_\alpha) > 0$, and thus, as shown above, we get $d(z, y_j) > d(z, x)$ for each j. Thus, it suffices to show that the number of good words z with $d(z, x) = r$ is at least the given bounds in the theorems.

Proof of Theorem 7. It now suffices to show that there exist c and n_0 such that for any choice of $n \geq n_0$, $x = 0$, $y_1, \ldots, y_\ell$ and r with $0 \leq r \leq (1/2 - \varepsilon)n$, if z is a random word with $d(z, x) = r$, i.e., a random string of r 1's and $n - r$ 0's, then

$$\mathbb{P}(z \text{ is good}) \geq c.$$

Suppose that this is false for all c and n_0. Then there exists a sequence of such $(n, y_1, \ldots, y_\ell, r)$, say $n_\gamma, y_1^{(\gamma)}, \ldots, y_\ell^{(\gamma)} \in \mathbb{F}^{n_\gamma}$ and r_γ, $\gamma = 1, 2, \ldots$, such that $n_\gamma \to \infty$ and if $z \in \mathbb{F}^{n_\gamma}$ is a random string with r_γ 1's, then

$$\mathbb{P}(z \text{ is good}) \to 0.$$

The sets A_α depend on γ, but by selecting a subsequence, we may assume that for each $\alpha \in \mathbb{F}^\ell$, either

$$|A_\alpha| = a_\alpha \text{ for some finite } a_\alpha \tag{9}$$

or

$$|A_\alpha| \to \infty. \tag{10}$$

Let $S = \{\alpha : \alpha \text{ is of type (9)}\}$. Let z be a random word as above (length n_γ with $r_\gamma \le (1/2-\varepsilon)n_\gamma$ non-zero coordinates). Let $\mathcal{E}_1$ be the event that $s_\alpha(z) = 0$ for each α of type (9). The bits $z(i)$ for the finitely many indices $i \in A_\alpha$ for some α of type (9) are asymptotically independent and each is 0 with probability $(n_\gamma - r_\gamma)/n_\gamma > 1/2$.

Hence

$$\liminf_{\gamma\to\infty} \mathbb{P}(\mathcal{E}_1) \ge 2^{-\sum_{\alpha\in S} a_\alpha} > 0.$$

(This depends on a_α, but we have chosen them and they are now fixed). Given $\mathcal{E}_1$, for every $\alpha \notin S$ (i.e., α is of type (10)) the random variable $s_\alpha(z)$ has a hypergeometric distribution with mean

$$\frac{r_\gamma}{n_\gamma - \sum_{\alpha\in S} a_\alpha}|A_\alpha|$$

and it follows by the law of large numbers that

$$\mathbb{P}\left(\left| \frac{s_\alpha(z)}{|A_\alpha|} - \frac{r_\gamma}{n_\gamma} \right| < \varepsilon \;\Big|\; \mathcal{E}_1 \right) \to 1.$$

Since $r_\gamma/n_\gamma \le 1/2 - \varepsilon$, it follows that

$$\mathbb{P}\left(\frac{s_\alpha(z)}{|A_\alpha|} < \frac{1}{2} \;\Big|\; \mathcal{E}_1 \right) \to 1$$

for each $\alpha \notin S$. Hence, with probability $(1+o(1))\,\mathbb{P}(\mathcal{E}_1)$,

$$\begin{cases} s_\alpha(z) = 0, & \alpha \in S, \\ s_\alpha(z) < \frac{1}{2}|A_\alpha|, & \alpha \notin S, \end{cases}$$

and then z is good.

Hence

$$\liminf_{\gamma\to\infty} \mathbb{P}(z \text{ is good}) \ge \liminf_{\gamma\to\infty} \mathbb{P}(\mathcal{E}_1) > 0,$$

a contradiction. ∎

For the remaining two proofs we will use the central limit theorem in its simplest version, for symmetric binomial variables. (This was also historically the first version, proved by de Moivre in 1733 [8, 1].) We let, for $N \ge 1$, X_N denote a binomial random variable with the distribution $\mathrm{Bi}(N, 1/2)$. The central limit theorem says that $(X_N - N/2)/\sqrt{N/4}$ converges in distribution to the standard normal distribution $N(0,1)$, which means that if $Z \sim N(0,1)$, then for any interval $I \subseteq \mathbb{R}$,

$$\mathbb{P}\Big(\frac{X_N - N/2}{\sqrt{N/4}} \in I\Big) \to \mathbb{P}(Z \in I) \qquad \text{as } N \to \infty. \tag{11}$$

We will also need the more precise local central limit theorem which says that if x_N is any sequence of integers, then, as $N \to \infty$,

$$\mathbb{P}(X_N = x_N) = \binom{N}{x_N} 2^{-N} = (2/\pi N)^{1/2}\Big(e^{-2(x_N - N/2)^2/N} + o(1)\Big). \tag{12}$$

(This is a simple consequence of Stirling's formula.)

Proof of Theorem 8. Let $n_\alpha = |A_\alpha|$, and note that $\sum_{\alpha \in \mathbb{F}^\ell} n_\alpha = n$. Fix an index α_0 with $n_{\alpha_0} \geq n/2^\ell$ (for example the index maximizing n_α). Let $\mathcal{A} = \{\alpha \in \mathbb{F}^\ell : n_\alpha > 0\}$ and $\mathcal{A}' = \mathcal{A} \setminus \{\alpha_0\}$. Consider a random $z \in \mathbb{F}^n$. The numbers $s_\alpha = s_\alpha(z)$ thus are independent binomial random variables: $s_\alpha \sim \mathrm{Bi}(n_\alpha, 1/2)$. Let $\delta = 2^{-\ell-1}a$. Let $\mathcal{E}_\alpha$ be the event

$$n_\alpha/2 > s_\alpha \geq n_\alpha/2 - \lceil \delta\sqrt{n} \rceil, \tag{13}$$

for $\alpha \in \mathcal{A}'$, let $\mathcal{E}_{\alpha_0}$ be the event

$$s_{\alpha_0} = r - \sum_{\mathcal{A}'} s_\alpha, \tag{14}$$

and let $\mathcal{E} = \bigwedge_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha$. Assume in the sequel that $\sqrt{n} \geq 2^{\ell+1}/a$. If $\mathcal{E}$ holds, then

$$\begin{aligned} s_{\alpha_0} = r - \sum_{\mathcal{A}'} s_\alpha &\leq n/2 - a\sqrt{n} - \sum_{\mathcal{A}'} (n_\alpha/2 - \delta\sqrt{n} - 1) \\ &< n_{\alpha_0}/2 - a\sqrt{n} + 2^\ell \delta\sqrt{n} + 2^\ell \leq n_{\alpha_0}/2, \end{aligned} \tag{15}$$

and thus z is good; further, $d(z,x) = \sum_\alpha s_\alpha = r$. It thus suffices to prove that $\mathbb{P}(\mathcal{E}) \geq cn^{-1/2}$, since then the number of good words z with $d(z,x) = r$ is at least $\mathbb{P}(\mathcal{E})2^n \geq cn^{-1/2}2^n$, and further $\binom{n}{r} \leq n^{-1/2}2^n$ by (12) (at least for large n).

First, let

$$p_N = \mathbb{P}(N/2 > X_N \geq N/2 - \lceil \delta\sqrt{N} \rceil).$$

Note that $p_N \geq \mathbb{P}(X_N = \lfloor (N-1)/2 \rfloor) > 0$ for every $N \geq 1$, and that the central limit theorem (11) shows that as $N \to \infty$, $p_N \to \mathbb{P}(0 \geq Z \geq -2\delta) > 0$. Hence, $p_* = \inf_{N \geq 1} p_N > 0$. Consequently, for $\alpha \in \mathcal{A}'$, $\mathbb{P}(\mathcal{E}_\alpha) \geq p_{n_\alpha} \geq p_*$. Moreover, the events $\mathcal{E}_\alpha$, $\alpha \in \mathcal{A}'$, are independent, and thus

$$\mathbb{P}\Big(\bigwedge_{\alpha \in \mathcal{A}'} \mathcal{E}_\alpha\Big) = \prod_{\alpha \in \mathcal{A}'} \mathbb{P}(\mathcal{E}_\alpha) \geq p_*^{2^\ell}.$$

Secondly, if (13) holds for $\alpha \in \mathcal{A}'$, then $r - \sum_{\mathcal{A}'} s_\alpha < n_{\alpha_0}/2$ by the calculation in (15), and

$$r - \sum_{\mathcal{A}'} s_\alpha \geq n/2 - b\sqrt{n} - \sum_{\mathcal{A}'} n_\alpha/2 = n_{\alpha_0}/2 - b\sqrt{n} \geq n_{\alpha_0}/2 - b2^{\ell/2}\sqrt{n_{\alpha_0}}.$$

The random variable s_{α_0} is independent of $\{s_\alpha : \alpha \in \mathcal{A}'\}$, and $s_{\alpha_0} \sim \mathrm{Bi}(n_{\alpha_0}, 1/2)$. Thus, the local limit theorem (12) shows that for every set of numbers s_α, $\alpha \in \mathcal{A}'$, satisfying (13),

$$\begin{aligned} \mathbb{P}\big(\mathcal{E}_{\alpha_0} \mid s_\alpha,\, \alpha \in \mathcal{A}'\big) &= (2/\pi n_{\alpha_0})^{1/2}\Big(\exp\Big(-2\Big(r - \sum_{\mathcal{A}'} s_\alpha - n_{\alpha_0}/2\Big)^2/n_{\alpha_0}\Big) + o(1)\Big) \\ &\geq (2n_{\alpha_0})^{-1/2}\Big(\exp\Big(-2^{\ell+1}b^2\Big) + o(1)\Big) \geq c_1 n^{-1/2} \end{aligned}$$

for some $c_1 > 0$, provided n, and thus also $n_{\alpha_0} \geq 2^{-\ell}n$, is large enough. Consequently, for large n,

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}\Big(\bigwedge_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha\Big) = \mathbb{P}\Big(\mathcal{E}_{\alpha_0} \mid \bigwedge_{\alpha \in \mathcal{A}'} \mathcal{E}_\alpha\Big)\mathbb{P}\Big(\bigwedge_{\alpha \in \mathcal{A}'} \mathcal{E}_\alpha\Big) \geq c_1 n^{-1/2} p_*^{2^\ell} = cn^{-1/2},$$

which completes the proof. ☐

Proof of Theorem 9. Let n_α, α_0, $\mathcal{A}$ and $\mathcal{A}'$ be as in the preceding proof and consider again a random $z \in \mathbb{F}^n$. Define the numbers t_α, $\alpha \in \mathcal{A}$, by

$$t_\alpha = \begin{cases} \lfloor (n_\alpha - 1)/2 \rfloor, & \alpha \in \mathcal{A}', \\ r - \sum_{\mathcal{A}'} t_\alpha, & \alpha = \alpha_0, \end{cases}$$

and let $\mathcal{E}$ be the event

$$s_\alpha = t_\alpha, \qquad \alpha \in \mathcal{A}.$$

Note that

$$t_{\alpha_0} \leq r - \sum_{\mathcal{A}'} (n_\alpha/2 - 1) \leq n_{\alpha_0}/2 - L + |\mathcal{A}'| < n_{\alpha_0}/2,$$

and thus $\mathcal{E}$ implies that z is good and $d(z, x) = \sum_\alpha s_\alpha = r$.

Since also $t_{\alpha_0} \geq r - \sum_{\mathcal{A}'} n_\alpha/2 \geq n_{\alpha_0}/2 - L - 1$, it follows from (12) that for some constant $c_2 > 0$ (depending on L) and every $n \geq 2^\ell L$,

$$\mathbb{P}(s_\alpha = t_\alpha) \geq c_2 n_\alpha^{-1/2} \geq c_2 n^{-1/2}$$

for every $\alpha \in \mathcal{A}$, and thus

$$\mathbb{P}(\mathcal{E}) = \prod_{\alpha \in \mathcal{A}} \mathbb{P}(s_\alpha = t_\alpha) \geq c_3 n^{-2^\ell/2},$$

which completes the proof. ☐

3 Construction of small $(r, \leq 2)$-identifying codes

Proof of Theorem 3. We make an explicit construction for $r = \lfloor n/2 \rfloor - 1$. If $2 \leq n \leq 3$, then $r = 0$ and we trivially may take $C = \mathbb{F}^n$. Furthermore, the following few values are known (see [13, 9]) $M_1^{(\leq 2)}(4) = 11$, $M_1^{(\leq 2)}(5) = 16$, $M_2^{(\leq 2)}(6) \leq 22$. So, we may assume that $n \geq 7$.

Let C_0 consist of the words $c_0 = 0 \in \mathbb{F}^n$ and $c_i \in \mathbb{F}^n$ such that $\mathrm{supp}(c_i) = \{1, 2, \ldots, i\}$ for $i = 1, 2, \ldots, n$. Clearly $|C_0| = n + 1$. Let $C_u = \{a \in \mathbb{F}^n \mid a \in C_0 \text{ or } \overline{a} \in C_0\}$. Now $|C_u| = 2n$. The code which we claim to be $(r, \leq 2)$-identifying for $r = \lfloor n/2 \rfloor - 1$ is then the following

$$\begin{aligned} C &= \{c \in \mathbb{F}^n \mid 1 \leq d(c, a) \leq 2 \text{ for some } a \in C_u\} \\ &= \{c \in \mathbb{F}^n \mid d(c, a) = 2 \text{ for some } a \in C_u\}, \end{aligned} \tag{16}$$

where the equality follows since every word in C_u has two neighbours in C_u. Obviously, $|C| \leq \binom{n}{2}|C_u| = n^3 - n^2$. (We are interested in the order of growth, so this estimate is enough for our purposes. However, with some effort one can check that $|C| = n^3 - 5n^2 + 4n$ for $n \geq 7$.)

We consider separately the cases n even and n odd.

1) Let first n be odd. The code C_0 is such that from every word $x \in \mathbb{F}^n$ we have a codeword exactly at distance $(n-1)/2$. Indeed, either $d(x, 0) > (n-1)/2$ and $d(x, 1) \leq (n-1)/2$ or $d(x, 0) \leq (n-1)/2$ and $d(x, 1) > (n-1)/2$. Moving (in the first case — the second case is analogous) from the codeword $c_0 = 0$ to $c_n = 1$ visiting every codeword c_i $(i = 1, \ldots, n)$, there exists an index i such that $d(x, c_i) = (n-1)/2$, since every move between two codewords c_i and c_{i+1} changes the distance by ± 1.

Now we need to show that

$$I_r(X) \neq I_r(Y)$$

for any two distinct subsets $X \subseteq \mathbb{F}^n$ and $Y \subseteq \mathbb{F}^n$ where $|X| \leq 2$ and $|Y| \leq 2$. Assume to the contrary that $I_r(X) = I_r(Y)$ for some $X, Y \subseteq \mathbb{F}^n$ with $|X|, |Y| \leq 2$ and $X \neq Y$.

Without loss of generality, we can assume that $|X| \geq |Y|$ and that we have a word $x \in X \setminus Y$. Using the property of C_u, we know that there exists a codeword $a \in C_u$ such that $d(x, a) = (n-1)/2$ and $d(x, \overline{a}) = (n+1)/2$. We concentrate on the words in the sets $S_1(a) \cup S_2(a)$ and $S_1(\overline{a}) \cup S_2(\overline{a})$ which all belong to C. Since $I_r(x) \subseteq I_r(X)$, we know that the sets

$$I_r(X) \cap S_1(a), \quad I_r(X) \cap S_2(a) \text{ and } I_r(X) \cap S_2(\overline{a}) \tag{17}$$

are all nonempty. By the symmetry of $\mathbb{F}^n$, we can assume without loss of generality, that $a = 0$ (and so, $\overline{a} = 1$).

Since $I_r(X) \cap S_1(0)$ is nonempty, there must be $\gamma \in Y$ such that $w(\gamma) \leq (n-1)/2$.

(i) Suppose first that $w(\gamma) \leq (n-5)/2$. This implies that $S_1(0) \subseteq I_r(\gamma) \subseteq I_r(Y) = I_r(X)$. Consequently, there must exist $y \in X$ $(y \neq x)$ such that

$w(y) \leq (n-1)/2$, since x does not r-cover all of $S_1(0)$. Since $|X| \leq 2$, thus $X = \{x, y\}$.

In order to cover the (nonempty) set $I_r(X) \cap S_2(1)$, there has to be β in Y ($\beta \neq \gamma$) such that $w(\beta) \geq (n-1)/2$. Thus $Y = \{\gamma, \beta\}$. If $w(\beta) > (n-1)/2$, then $I_r(\beta)$ (and hence $I_r(Y)$) contains elements from $S_1(1)$, but the set $I_r(X) \cap S_1(1)$ is empty, immediately giving a contradiction. If $w(\beta) = (n-1)/2$, then $I_r(x)$ contains a codeword not in $I_r(Y)$. Indeed, since $x \neq \beta$ (and $w(x) = w(\beta)$), then there exists an index $j \in \operatorname{supp}(\beta)$ such that $j \notin \operatorname{supp}(x)$. This implies that the needed codeword, say c', is found in $S_2(1)$ by taking $\operatorname{supp}(\overline{c'}) = \{i, j\}$ for any $i \notin \operatorname{supp}(x)$ and $i \neq j$ — clearly, β cannot r-cover this codeword and γ cannot r-cover any word in $S_2(1)$.

(ii) Assume then that $w(\gamma) = (n-3)/2$. Now γ cannot r-cover all the words in $I_r(x) \cap S_1(0)$, so there must be $\beta \in Y$ such that $w(\beta) \leq (n-1)/2$. If $w(\beta) < (n-1)/2$, then $I_r(Y) \cap S_2(1) = \emptyset$ which contradicts $I_r(X) \cap S_2(1) \neq \emptyset$. If $w(\beta) = (n-1)/2$, we are done as in (i), using again $x \neq \beta$.

(iii) Let then $w(\gamma) = (n-1)/2$. Because there are codewords in $I_r(x) \cap S_1(0)$ which are not r-covered by γ, it follows that there exists $\beta \in Y$ with $w(\beta) \leq (n-1)/2$. By the previous cases, it suffices to consider $w(\beta) = (n-1)/2$, since otherwise we can interchange β and γ. Let $i \in \operatorname{supp}(x)$ be such that $i \notin \operatorname{supp}(\gamma)$ and $j \in \operatorname{supp}(x)$ such that $j \notin \operatorname{supp}(\beta)$. Since $x \notin Y$, such indices (it is possible that $i = j$) exist. When $i \neq j$, a codeword $c \in S_2(0)$ such that $\operatorname{supp}(c) = \{i, j\}$, gives a contradiction. If $i = j$, then we pick a codeword with $\operatorname{supp}(c) = \{i, k\}$ where $k \in \operatorname{supp}(x)$, $i \neq k$.

2) Let now n be even. Take C_u as in the odd case; it has now the analogous property that from every word $x \in \mathbb{F}^n$ there is a codeword $a \in C_u$ such that $d(x, a) = d(x, \overline{a}) = n/2$. Let C be defined also as above. We will show that it is $(r, \leq 2)$-identifying for $r = n/2 - 1$. If $I_r(X) = I_r(Y)$, we can again assume that $|X| \geq |Y|$ and choose $x \in X \setminus Y$. We know that there is $a \in C_u$ such that $d(x, a) = n/2$. The sets (17) as well as now the set $I_r(x) \cap S_1(\overline{a})$ are nonempty. Again it suffices to consider $a = 0$. Since $I_r(x) \cap S_1(0)$ is nonempty, so there must be a word $\gamma \in Y$ such that $w(\gamma) \leq n/2$.

(i) Suppose first that $w(\gamma) \leq n/2 - 2$. Then $S_1(0) \subseteq I_r(\gamma)$. Since $S_1(0) \nsubseteq I_r(x)$, this implies that there is $y \in X$, $y \neq x$, such that $w(y) \leq n/2$.

Let first $w(y) \leq n/2 - 1$. Subsequently, neither y nor γ r-covers any of the codewords of $S_1(1)$ whereas $|I_r(x) \cap S_1(1)| = n/2$. Hence there has to be $\beta \in Y$ such that $I_r(x) \cap S_1(1) = I_r(\beta) \cap S_1(1)$. However, this implies that $x = \beta$, a contradiction.

Assume next that $w(y) = n/2$. Due to the fact that γ r-covers all the codewords $S_1(0)$ we know that $y = \overline{x}$. Hence $S_1(1) \subseteq I_r(X) = I_r(Y)$. On the other hand, $S_1(1) \cap I_r(\gamma) = \emptyset$. Consequently, $S_1(1) \subseteq I_r(\beta)$ which implies $w(\beta) \geq n/2 + 2$. Thus β does not r-cover any of the words in $S_2(0)$.

- If $w(\gamma) \leq n/2 - 3$, then γ r-covers all of $S_2(0)$. However, all of $S_2(0)$ is not contained in $I_r(X)$. Indeed, take $i \in \operatorname{supp}(x)$ and $j \notin \operatorname{supp}(x)$ (notice that now $X = \{x, \overline{x}\}$). The codeword c' of $S_2(0)$ which has $\operatorname{supp}(c') = \{i, j\}$ does not belong to $I_r(X)$.

- If $w(\gamma) = n/2 - 2$, then $\mathrm{supp}(x)$ has (at least) two distinct indices, say i and j, which are not in $\mathrm{supp}(\gamma)$. Consequently, the codeword c' in $S_2(0)$ with $\mathrm{supp}(c') = \{i, j\}$ belongs to $I_r(x)$ but not to $I_r(Y)$, a contradiction.

(ii) Let now $w(\gamma) = n/2 - 1$. Now $I_r(x) \cap S_1(0)$ has one more codeword than $I_r(\gamma) \cap S_1(0)$ and, therefore, there exists $\beta \in Y$ such that $w(\beta) \leq n/2$ (or we are done). But now $I_r(x) \cap S_1(1)$ contains at least one codeword not in $I_r(\beta)$ — notice that γ does not r-cover any words in $S_1(1)$.

(iii) Assume finally that $w(\gamma) = n/2$. Since $\gamma \neq x$, there must exist an index $i \in \mathrm{supp}(x)$ such that $i \notin \mathrm{supp}(\gamma)$. Consequently, the r codewords $z_j \in S_2(0) \cap I_r(x)$ such that $\mathrm{supp}(z_j) = \{i, j\}$, where $j \in \mathrm{supp}(x)$ and $j \neq i$, do not belong to $I_r(\gamma)$ and hence must belong to $I_r(\beta)$ for some other $\beta \in Y$ or we are done. By the previous cases and symmetry with respect to a and $\overline{a}$, we can also assume that $w(\beta) = n/2$. However, this means that $\mathrm{supp}(z_j) \subseteq \mathrm{supp}(\beta)$ for all j. Subsequently, $x = \beta$ or $w(\beta) > n/2$ and we get a contradiction in both cases. $\square$

Acknowledgement. This research was done mainly during the Workshop on Codes and Discrete Probability in Grenoble, 2007.

References

[1] R. C. Archibald. A rare pamphlet of Moivre and some of his discoveries, *Isis*, 8(4):671–683, 1926.

[2] Y. Ben-Haim and S. Litsyn. Exact minimum density of codes identifying vertices in the square grid. *SIAM J. Discrete Math.*, 19(1):69–82, 2005.

[3] U. Blass, I. Honkala and S. Litsyn. On binary codes for identification. *J. Combin. Des.*, 8:151-156, 2000.

[4] I. Charon, S. Gravier, O. Hudry, A. Lobstein, M. Mollard and J. Moncel. A linear algorithm for minimum 1-identifying codes in oriented trees. *Discrete Appl. Math.*, 154(8):1246–1253, 2006.

[5] I. Charon, I. Honkala, O. Hudry and A. Lobstein. Structural properties of twin-free graphs. *Electron. J. Combin.*, 14(1):R16, 15 pp., 2007.

[6] I. Charon, O. Hudry and A. Lobstein. Extremal cardinalities for identifying and locating-dominating codes in graphs. *Discrete Math.*, 307(3-5):356–366, 2007.

[7] G. Cohen, I. Honkala, S. Litsyn and A. Lobstein. *Covering Codes*. Elsevier, Amsterdam, 1997.

[8] A. de Moivre. *Approximatio ad Summam Terminorum Binomii* $\overline{a+b}^n$ *in Seriem expansi*. 1733. English translation in A. de Moivre, *The Doctrine of Chances*, 2nd ed., London, 1738.

[9] G. Exoo, V. Junnila, T. Laihonen and S. Ranto. Improved bounds on identifying codes in binary Hamming spaces. Submitted.

[10] G. Exoo, T. Laihonen and S. Ranto. New bounds on binary identifying codes. *Discrete Appl. Math.*, to appear.

[11] A. Frieze, R. Martin, J. Moncel, M. Ruszinkó and C. Smyth. Codes identifying sets of vertices in random networks. *Discrete Math.*, 307(9-10):1094–1107, 2007.

[12] S. Gravier, J. Moncel and A. Semri. Identifying codes of cycles. *Eur. J. Comb.*, 27(5):767–776, 2006.

[13] I. Honkala, T. Laihonen and S. Ranto. On codes identifying sets of vertices in Hamming space. *Des. Codes Cryptogr.*, 24:193–204, 2001.

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

[15] 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.

[16] M. Laifenfeld and A. Trachtenberg. Disjoint identifying-codes for arbitrary graphs. In *Proceedings of International Symposium on Information Theory, 2005. ISIT 2005*, pp. 244–248, 2005.

[17] M. Laifenfeld, A. Trachtenberg and T. Y. Berger-Wolf. Identifying codes and the set cover problem. In *Proceedings of the 44th Annual Allerton Conf. on Communication, Control and Computing*, Monticello, USA, 2006.

[18] T. Laihonen and S. Ranto. Codes identifying sets of binary words with large radii. In *Proceedings of International Workshop on Coding and Cryptography WCC 2007*, pp. 215–224, 2007.

[19] A. Lobstein. Identifying and locating-dominating codes in graphs, a bibliography. Published electronically at `http://perso.enst.fr/~lobstein/debutBIBidetlocdom.pdf`.

[20] J. Moncel. Monotonicity of the minimum cardinality of an identifying code in the hypercube. *Discrete Appl. Math.*, 154(6):898–899, 2006.

[21] S. Ray, D. Starobinski, A. Trachtenberg and R. Ungrangsi. Robust location detection with sensor networks. *IEEE Journal on Selected Areas in Comminications (Special Issue on Fundamental Performance Limits of Wireless Sensor Networks)*, 22(6):1016–1025, 2004.

[22] D. L. Roberts and F. S. Roberts. Locating sensors in paths and cycles: The case of 2-identifying codes. *Eur. J. Comb.*, 29(1):72–82, 2008.