A Combinatorial Problem Related to Sparse Systems of Equations

Peter Horak [1*] Igor Semaev [2†] Zsolt Tuza [3,4‡]

[1] School of Interdisciplinary Arts & Sciences, University of Washington, Tacoma, USA

[2] Department of Informatics, University of Bergen, Norway

[3] Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary

[4] University of Pannonia, Veszprém, Hungary

Abstract

Nowadays sparse systems of equations occur frequently in science and engineering. In this contribution we deal with sparse systems common in cryptanalysis. Given a cipher system, one converts it into a system of sparse equations, and then the system is solved to retrieve either a key or a plaintext. Raddum and Semaev proposed new methods for solving such sparse systems. It turns out that a combinatorial MaxMinMax problem provides bounds on the average computational complexity of sparse systems. In this paper we initiate a study of a linear algebra variation of this MaxMinMax problem.

MSC: 05C65, 68Q25, 94A60

Keywods: sparse systems of equations; MaxMinMax problem; gluing algorithm.

1 Introduction

Sparse objects such as sparse matrices and sparse systems of (non-)linear equations occur frequently in science and engineering. Nowadays sparse systems are

*Research of P. Horak was supported in part by a grant from SIAS, University of Washington, Tacoma.

†Research of P. Horak and I. Semaev was supported in part by a grant SPIRE program in 2013-2015 from University of Bergen. Research of I. Semaev was also partly supported by the EEA Grant SK06-IV-01-001, and the state budget of the Slovak Republic from the EEA Scholarship Programme Slovakia.

‡Research of Zs. Tuza was supported in part by the grant TÁMOP-4.2.2.B-15/1/KONV-2015-0004.

often studied in algebraic cryptanalysis as well. First, given a cipher system, one converts it into a system of equations. Second, the system of equations is solved to retrieve either a key or a plaintext. As pointed out in [1], this system of equations will be sparse, since efficient implementations of real-word systems require a low gate count.

There are plenty of papers on methods for solving a sparse system of equations

$$f_i(X_i) = 0 (1 \le i \le m) \tag{1}$$

over $GF(q)$. The worst case complexity bounds on (1) are attained in the case of sparse systems describing the SAT problem. These bounds are exponential with respect to the number of unknowns in (1). For example, in the case of the binary field, the bounds are 2^{cn}, where the constant c is close to 1, and depends on the size of $|X_i|'$s, see [3]. In [7] the so-called Gluing Algorithm was designed to solve such systems over any finite field $GF(q)$. If the set S_k of solutions of the first k equations together with the next equation $f_{k+1} = 0$ is given then the algorithm constructs the set S_{k+1}. It is shown there that the average complexity of finding all solutions to the original system is $O(mq^{\max|\cup_1^k X_j|-k})$, where m is the total number of equations, and $\cup_1^k X_j$ is the set of all unknowns actively occurring in the first k equations. Clearly, the complexity of finding all solutions to the system by the Gluing Algorithm depends on the order of equations. Hence one is interested to find a permutation π that minimizes the average complexity, and also to describe the worst-case scenario, i.e., the system of equations for which the average complexity of the method is maximum. Therefore, Semaev [8] suggested to study the following combinatorial MaxMinMax problem.

Let $\mathcal{S}_{n,m,t}$ be the family of all collections of sets $\mathcal{X} = \{X_1, ..., X_m\}$, where the X_i are subsets of an underlying n-set X, and $|X_i| \le t$ holds for all $i \in [m]$; we allow that some set may occur in $\mathcal{X}$ more than once. Then we define

$$f_t(n,m) := \max_{\mathcal{X}} \min_{\pi} \max_{1 \le k \le m} \left(\left| \bigcup_{i=1}^{k} X_{\pi(i)} \right| - k \right) \tag{2}$$

where the minimum runs over all permutations π on $[m]$, and the maximum is taken over all families $\mathcal{X}$ in $\mathcal{S}_{n,m,t}$.

In [2] the authors confined themselves to the case $|X_i| \le 3$ for all $i \in [m]$. It was shown there that, for $n \ge 2$ and all $m \le n-1$, $f_2(n,m)$ equals the maximum number of non-trivial components in a simple forest on n vertices with m edges; otherwise $f_2(n,m) = 1$. The main result of that paper claims that $f_3(n,n)$ grows linearly. More precisely, the following estimates are valid.

Theorem 1 *For all n sufficiently large, $f_3(n,n) \ge \frac{n}{12.2137}$ holds, while $f_3(n,n) \le \left\lceil \frac{n}{4} \right\rceil + 2$ for all $n \ge 3$.*

Later, an asymptotically better upper bound was proved in [8]; we note that the proof of the bound is algorithmic, and the needed permutation π is constructed in polynomial time.

Theorem 2 *For all n, $f_3(n,n) \leq \frac{n}{5.7883} + 1 + 2\log_2 n$.*

As a corollary we get: Let $\mathcal{X}$ be fixed. If $|X_i| \leq 3$, $m = n$, then the average complexity of finding all solutions in $GF(q)$ to polynomial equation system (1) is at most $q^{\frac{n}{5.7883}+O(\log n)}$ for arbitrary $\mathcal{X}$ and q.

In [6] a new method for representing and solving systems of algebraic equations common in cryptanalysis has been proposed. This method differs from the others in that the equations are not represented as multivariate polynomials, but as a system of Multiple Right-Hand Sides (MRHS) linear equations. The results overcome significantly what was previously achieved with Gröbner Basis related algorithms. We point out that equations describing the Data Encryption Standard (DES) [5] and the Advanced Encryption Standard (AES) can be expressed in MRHS form as well.

AES is likely the most commonly used symmetric-key cipher; AES became effective as a US federal government standard on May 26, 2002 after approval by the Secretary of Commerce. It is the first publicly accessible and open cipher approved by the National Security Agency (NSA) for top secret information when used in an NSA-approved cryptographic module. An application of MRHS equations enables one to improve on the linear cryptanalysis of DES [9].

Let X be a column n-vector of unknowns over $GF(q)$. Then an MRHS system of equations is a system of inclusions

$$A_i X \in \{b_{i_1}, \ldots, b_{i_{s_i}}\}, \quad i = 1, \ldots, m, \tag{3}$$

where the A_i are matrices over $GF(q)$ of size $t_i \times n$ and of rank t_i, and the b_{ij} are column vectors of length t_i. An $X = X_0$ is a solution to (3) if it satisfies all inclusions in (3). Methods to solve an arbitrary MRHS system of equations were introduced in [6] as well.

One of the main goals of our paper is to get asymptotic bounds on the average complexity of solving (3). As noted by Semaev, such bounds can be obtained by studying a generalization of the combinatorial problem described in (2). In particular, when matrices A_i are fixed and the right hand side columns in (3) are generated according to a probabilistic model presented in the next section, we will prove

Theorem 3 *Let $t_i \leq t$ for some fixed t and n tends to infinity. Then the average complexity of solving* (3) *is*

$$O(m\, q^{n-\lceil n/t \rceil}).$$

Proof. The proof of the statement follows directly from Theorem 7 and Lemma 8 that will be stated and proved in Section 2 and Section 3, respectively. ■

2 Probabilistic Model

We will use a generalization of the probabilistic model used in [8]. For $i = 1, ..., m$ we choose in random, independent and uniform way polynomials $g_i(y_1, \ldots, y_{t_i})$

over $GF(q)$. The degree of the polynomial in each of its variables is at most $q-1$. The right-hand sides $\{b_{i_1}, \ldots, b_{i_{s_i}}\}$ in (3) are zeroes of polynomials $g_i(y_1, \ldots, y_{t_i})$ over $GF(q)$. So that (3) is equivalent to a system of polynomial equations

$$g_i(A_i X) = 0, i = 1, ..., m.$$

Lemma 4 *The probability of a fixed vector $b \in GF(q)^t$ to be a zero of a random polynomial g over $GF(q)$ equals $\frac{1}{q}$.*

Proof. The set of the polynomials in t variables over $GF(q)$ and of degree at most $q-1$ in each of its variables is in one-to-one natural correspondence with the set of all mappings from $GF(q)^t$ to $GF(q)$. The sought probability equals $\frac{1}{q}$ as the overall number of such polynomials (mappings) is q^{q^t}, and the number of the polynomials (mappings) g, where $g(b) = 0$, is q^{q^t-1}. ■

A random selection of elements from a set of cardinality n with probability p, yields a set whose expected cardinality is equal to pn. In particular, for our problem under consideration we immediately obtain:

Lemma 5 *Let $AX \in S$ be a multiple right-hand side equation, where A is of size $t \times n$ and of rank t. Suppose the right-hand sides S are taken randomly such that $\mathbf{Pr}(b \in S) = p$. Then the average size of S is $p\, q^t$.*

For two equations $A_1 X \in S_1$ and $A_2 X \in S_2$, one can construct an equation $AX \in S$ such that its solutions are precisely the common solutions to the original two equations. In [6] the operation is called gluing, this is a linear algebra generalization of gluing earlier introduced in [7]. As above let A_i be of size $t_i \times n$ and of rank t_i. The matrix A is constructed by writing A_2 under A_1 and eliminating dependent rows. Therefore, $t = \mathbf{rank}(A) = \mathbf{rank}(A_1, A_2)$ is also the number of rows in A. Assume that S_1, S_2 are randomly generated.

Lemma 6 *Let $p_i = \mathbf{Pr}(b_i \in S_i)$ for any column vectors b_i of size t_i. If the sets S_1 and S_2 are generated independently, then for any column vector b of size t we have $\mathbf{Pr}(b \in S) = p_1 p_2$.*

Proof. Consider the system of linear equations $A_1 X = b_1, A_2 X = b_2$. We can express them as

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} X = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

If the system is consistent, then it is equivalent to $AX = b$ for some b. Due to the equivalence, $b \in S$ is constructed from only one pair $b_1 \in S_1$ and $b_2 \in S_2$. This proves the statement. ■

The lemma implies that the average number of the right-hand sides (the size of S) in $AX \in S$ is

$$p_1 p_2\, q^{\mathbf{rank}(A_1, A_2)}.$$

We say the system (3) is solved if it is represented by only one equation $AX \in S$. In turn, this is equivalent to solving $|S|$ systems of ordinary linear equations over $GF(q)$; its complexity is neglected here.

Let r_t denote the rank of all row vectors in $A_1, A_2, \ldots, A_t$.

Theorem 7 *Assume that the right-hand sides of the equations in* (3) *are generated in a random, independent, and a uniform way. Then the average complexity of* (3) *is*

$$O(m \max_t q^{r_t - t})$$

Proof. The system is solved in aggregate by at most $m-1$ applications of gluing operation. The complexity of one gluing is proportional to the number of right-hand sides in the equations to glue and the number of the resulting right-hand sides, see [6]. By Lemma 4 we know that a column vector appears on the right-hand side in (3) with probability $1/q$. Further, by Lemmas 5 and 6, after the t-th gluing the average number of the right-hand sides is $\frac{1}{q^k} q^{\mathbf{rank}(A_1,\ldots,A_k)}$. The average complexity of the system is then the sum of the average number of the right-hand sides after each application:

$$\sum_{k=1}^{m} \frac{1}{q^t} q^{\mathbf{rank}(A_1,\ldots,A_t)} \le m \max_t q^{r_t - t}.$$

That proves the statement. ■

3 Corresponding Combinatorial Problem

In this section we formulate a combinatorial problem related to an MRHS system of equations. It turns out that the complexity of the problem can be described by a generalization of the function $f_t(n,m)$ defined in (2), namely the size of the union of the first k sets is replaced by the rank of vectors belonging to the first k matrices. Formally, let $\mathcal{S}_{n,m,t,V}$ be the family of all collections of sets of **vectors** $\mathcal{X} = \{X_1, ..., X_m\}$ in an n-dimensional vector space V, over any finite on infinite field, under the restriction $|X_i| \le t$ for all $i \in [m]$. We set

$$F_t(n,m) := \max_{\mathcal{X}} \min_{\pi} \max_{1 \le k \le m} (\mathbf{rank}\left(\bigcup_{i=1}^{k} X_{\pi(i)}\right) - k), \tag{4}$$

where the minimum runs over all permutations π on $[m]$, and the maximum is taken over all families $\mathcal{X}$ in $\mathcal{S}_{n,m,t,V}$. We note that the definition of the function $F_t(n,m)$ reflects the fact that the order of matrices A_i's is important in the Gluing Algorithm.

Although functions $f_t(n,m)$ and $F_t(n,m)$ are defined in a similar way, their behavior is dramatically different.

We start with a rather simple upper bound, used in the proof of Theorem 3.

Lemma 8 *Let n, t be natural numbers. Then*

$$F_t(n,n) \leq n - \left\lceil \frac{n}{t} \right\rceil \tag{5}$$

Proof. Let $n = st + k$, where $0 \leq k < t$. We have

$$\mathbf{rank}(X_1, \ldots, X_i) - i \leq i(t-1) \leq s(t-1) = n - k - s \leq n - \lceil n/t \rceil$$

for $i \leq s$. Moreover,

$$\mathbf{rank}(X_1, \ldots, X_i) - i \leq n - \lceil n/t \rceil$$

for $i \geq s+1$, as $i \geq \lceil n/t \rceil$ in this case, and the upper bound (5) follows. ■

It turns out that bounds on $F_t(m,n)$ constitute a challenge for most vector spaces V even for $t = 2$, although in the case of the function $f_t(n,m)$ we know its exact value in the case of $t = 2$. Surprisingly, the upper bound (5) can be attained in many cases.

Theorem 9 *Let F be any infinite field, or $F = GF(q)$, the finite field with $q \geq tn$ elements. Then for any n and t we have $F_t(n,n) = n - \left\lceil \frac{n}{t} \right\rceil$.*

Proof. By (5) it suffices to bound $F_t(n,n)$ from below. Let

$$M = \begin{pmatrix} 1 & \alpha_1 & \ldots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \ldots & \alpha_2^{n-1} \\ \ldots & & & \\ 1 & \alpha_{tn} & \ldots & \alpha_{tn}^{n-1} \end{pmatrix}.$$

be a matrix of size $tn \times n$, where $\alpha_i, i = 1, \ldots, tn$ are pairwise different elements from the field F. As M is a Vandermonde matrix, any n rows of M are linearly independent. To construct a desired collection $\mathcal{A} = \{X_1, ..., X_n\}$ we split the rows of M in an arbitrary way into sets A_i of size t. Let $n = st + k$, where $0 \leq k < t$. For any permutation π and $i \leq s$

$$\mathbf{rank}(X_{\pi(1)}, \ldots, X_{\pi(i)}) - i = i(t-1).$$

For $i \geq s+1$

$$\mathbf{rank}(X_{\pi(1)}, \ldots, X_{\pi(i)}) - i = n - i$$

Therefore, if $k = 0$, then the maximum difference is achieved at $i = s = \lceil \frac{n}{t} \rceil$. If $k > 0$, then the maximum difference is achieved at $i = s + 1 = \lceil \frac{n}{t} \rceil$. So

$$\max_{\pi} \min_{i} \left(\mathbf{rank}(X_{\pi(1)}, \ldots, X_{\pi(i)}) - i \right) = n - \lceil \frac{n}{t} \rceil.$$

The theorem follows. ■

4 Lower and Upper Bounds for Systems over $GF(2)$

In this section we focus on the binary field, the field most important for cryptographic applications. We start with an upper bound.

Theorem 10 *For n sufficiently large, $F_2(n,n) \leq \frac{n}{2} - \frac{1}{8}\log_2 n$.*

Proof. Let n be a fixed natural number. We choose t such that

$$4^t 2t + 2 \leq n < 4^{t+1}2(t+1) + 2$$

Then, for n sufficiently large, $t > \frac{1}{4}\log_2 n$. Let $\mathcal{A} = \{X_1, ..., X_n\}$ be a collection of sets $X_i = \{\mathbf{v}_i, \mathbf{w}_i\}$ comprising two binary vectors of length n.

Let k be the largest number such that there exists sets $X_{i_1}, ..., X_{i_k}$ with $\mathbf{rank}((\bigcup_{j=1}^{k} X_{i_j}) = 2k$. Clearly, $k \leq \frac{n}{2}$, and

$$\mathbf{rank}((\bigcup_{j=1}^{k} X_{i_j} \cup X_s) < 2k + 2 \tag{6}$$

for all $s \in \{1, ..., n\} - \{i_1, ..., i_k\}$. For simplicity, assume that $\{i_1, ..., i_k\} =: \{1, ..., k\}$.

We will consider two cases. First, let $k < \frac{n}{2} - \frac{1}{8}\log_2 n$. Then, for each $1 \leq s \leq k$, it is

$$\max_s \mathbf{rank}(\bigcup_{j=1}^{s} X_j) - s = 2s - s \leq k.$$

By definition of k we have that for all $s > k$, there is in X_i at least one vector that is a linear combination of vectors in $\bigcup_{i=1}^{k} X_i$. Thus we get

$$\max_s \mathbf{rank}(\bigcup_{j=1}^{s} X_j) - s \leq 2k + (s-k) - s = k.$$

Hence in this case $\Delta(\mathcal{X}) =: \min_\pi \max_s(\mathbf{rank}(\bigcup_{j=1}^{s} X_j) - s) \leq k < \frac{n}{2} - \frac{1}{8}\log_2 n$.

Now let $k \geq \frac{n}{2} - \frac{1}{8}\log_2 n$. As mentioned above, by definition of k, for each $i > k$, there is in X_i at least one vector that is a linear combination of vectors in $\bigcup_{i=1}^{k} X_i$. Let D be a multiset that contains one such vector from each of sets $X_i, i = k+1, ..., n$. We note that some vectors might occur in D more than once. Hence, $|D| \geq n - k \geq \frac{n}{2}$, and each vector in D is a linear combination of vectors in $\bigcup_{i=1}^{k} X_i = \bigcup_{j=1}^{k} \{\mathbf{v}_i, \mathbf{w}_i\}$. To each vector $\mathbf{u}$ in D we assign a k-tuple $T_{\mathbf{u}} = (x_1, ..., x_k)$, where $x_i \in \{\alpha, \beta, \gamma, \delta\}$, such that $x_i = \alpha$ if neither of $\mathbf{v}_i, \mathbf{w}_i$ occurs in the expression of $\mathbf{u}$ as a linear combination of elements from $\bigcup_{i=1}^{k} X_i$,

$x_i = \beta$ if $\mathbf{v}_i$ is there but $\mathbf{w}_i$ is not, $x_i = \gamma$ if $\mathbf{w}_i$ is there but $\mathbf{v}_i$ is not, and finally, $x_i = \delta$ if both $\mathbf{v}_i, \mathbf{w}_i$ occur there.

As $n \geq 4^t 2t + 2$ there are in D at least $4^t t + 1$ vectors. Of these vectors at least $4^{t-1}t+1$ have the same first component, and of the latter $4^{t-1}t+1$ vectors at least $4^{t-2}t + 1$ coincide in the first two components, etc. Thus, there are at least $t+1$ vectors, say $\mathbf{y}_0, \mathbf{y}_1, ..., \mathbf{y}_t$, that coincide in the first t components. For each $i = 1, ..., t$, we get $\mathbf{y}_i = \mathbf{y}_0 + \mathbf{x}$, where $\mathbf{x}$ is a linear combination of vectors in $X_{t+1}, ..., X_k$. Without loss of generality, assume $\mathbf{y}_i \in X_{k+1+i}$ for $i = 0, 1, ..., t$. Let π be a permutation $(t+1, ..., k+t+1, 1, ..., t, k+t+2, ..., n)$. We are going to show that, for all $1 \leq s \leq n$,

$$\mathbf{rank}(\bigcup_{i=1}^{s} X_{\pi(i)}) - s \leq \frac{n}{2} - \frac{t}{2}.$$

Consider a family $\mathcal{B} = \{B_1, ..., B_n\}$ that satisfies (6), i.e., for all $s > k$

$$\mathbf{rank}(\bigcup_{j=1}^{k} B_j) = 2k, \text{ and } \mathbf{rank}(\bigcup_{j=1}^{k} B_j \cup B_s) < 2k + 2.$$

Further, there is $z_i \in B_{k+1+i}$ for $i = 0, ..., t$, such that, for all $i = 1, ..., n$ it is $z_i = z_0 + x_i$, where x_i is a linear combination of vectors $\bigcup_{j=t+1}^{k} B_j$, such that

$$\mathbf{rank}(\bigcup_{i=t+1}^{k+t+1} B_i) = 2k - 2t + 2 + t = 2k - t + 2. \tag{7}$$

In addition, the collection $\mathcal{B}$ satisfies, for all $1 \leq s \leq t$,

$$\mathbf{rank}(\bigcup_{i=t+1}^{k+t+1} B_i \cup \bigcup_{i=1}^{s} B_i) = \min\{n, \mathbf{rank}(\bigcup_{i=t+1}^{k+t+1} B_i) + 2s\}. \tag{8}$$

We note that such a collection $\mathcal{B}$ exists to each collection $\mathcal{X}$ and it is not difficult to construct it. For each $1 \leq s \leq n$, we have

$$\mathbf{rank}(\bigcup_{i=1}^{s} X_{\pi(i)}) \leq \mathbf{rank}(\bigcup_{i=1}^{s} B_{\pi(i)}),$$

which in turn implies

$$\mathbf{rank}(\bigcup_{i=1}^{s} X_{\pi(i)}) - s \leq \mathbf{rank}(\bigcup_{i=1}^{s} B_{\pi(i)}) - s.$$

Thus, $\Delta(\mathrm{X}) \leq \Delta(\mathcal{B})$. To finish the proof we show that, for all $1 \leq s \leq n$, it is $\mathbf{rank}(\bigcup_{i=1}^{s} B_{\pi(i)}) - s \leq \frac{n}{2} - \frac{t}{2}$. For $1 \leq s \leq k + t + 1$, by (7) and (8),

$$\mathbf{rank}(\bigcup_{i=1}^{s+1} B_{\pi(i)}) \geq \min\{n, \mathbf{rank}(\bigcup_{i=1}^{s} B_{\pi(i)}) + 1\}.$$

In addition, as $2(k+t+1) \geq 2(\frac{n}{2} - \frac{1}{8}\log_2 n + \frac{1}{4}\log_2 n + 1) \geq n$, it is

$$\mathbf{rank}(\bigcup_{i=1}^{k+t+1} B_{\pi(i)}) = n.$$

Therefore, for each $s \leq n-1$, we have

$$\mathbf{rank}(\bigcup_{i=1}^{s+1} B_{\pi(i)}) \geq \min\{n, \mathbf{rank}(\bigcup_{i=1}^{s} B_{\pi(i)}) + 1\}.$$

Therefore, $\mathbf{rank}(\bigcup_{i=1}^{s} B_{\pi(i)}) - s$ is a non-decreasing function in s and it achieves its maximum at the smallest value of s where $\mathbf{rank}(\bigcup_{i=1}^{s} B_{\pi(i)}) = n$. By (7) and (8) we get that the maximum value is achieved at $s = k + 1 + s_0$ where

$$s_0 = \left\lceil \frac{1}{2}(n - 2k - 2 + t) \right\rceil \leq t$$

as we assume in this case that $k \geq \frac{n}{2} - \frac{1}{8}\log_2 n$.

This in turn implies

$$\Delta(\mathcal{B}) = \mathbf{rank}(\bigcup_{i=1}^{s_0} B_i \cup \bigcup_{i=t+1}^{k+t+1} B_i) - (s_0 + k + 1) = n - (s_0 + k + 1) = \frac{n}{2} - \frac{t}{2}.$$

Therefore $\Delta(\mathrm{X}) \leq \Delta(\mathcal{B}) \leq \frac{n}{2} - \frac{\log_2 n}{8}$ as $t > \frac{1}{4}\log_2 n$; i.e., $F_2(n,n) \leq \frac{n}{2} - \frac{1}{8}\log_2 n$. ■

Now we state two linear lower bounds on $F_2(n,n)$.

Theorem 11 *For all n sufficiently large, $F_2(n,n) \geq \frac{n}{9.0886}$.*

Proof. First we note that it is easy to check that $c = 0.2200557288$ satisfies the inequality

$$H_2(c/2) < 1/2, \tag{9}$$

where H_2 is the binary entropy function.

To prove the statement we will show that for all n sufficiently large there exists a matrix H over $GF(2)$ of size $2n \times n$ with the property that any $d-1$ of its rows are linearly independent, where $d = \lceil cn \rceil$. The matrix H will be constructed from a parity-check matrix of a suitable linear code.

Let $N = 2n$ and let r be the natural number such that

$$2^{r-1} \leq \sum_{i=1}^{d-2} \binom{N-1}{i} < 2^r,$$

It is well known, see e.g. Corollary 9, Chapter 10 in [4], that for $0 < \lambda < \frac{1}{2}$, it is

$$\sum_{k=0}^{\lambda n} \binom{n}{k} \leq 2^{nH_2(\lambda)}. \tag{10}$$

By definition of c we have $d-2 \leq \frac{c}{2}(N-1)$. Applying (10) and using $\frac{d-2}{N-1} \leq \frac{c}{2}$ we arrive at

$$2^{r-1} \leq \sum_{i=1}^{d-2} \binom{N-1}{i} \leq 2^{(N-1)H_2\left(\frac{d-2}{N-1}\right)} \leq 2^{(N-1)H_2\left(\frac{c}{2}\right)} \leq 2^{n-1}$$

as $(N-1)H_2\left(\frac{c}{2}\right) \leq n-1$ for n sufficiently large. This in turn implies $r \leq n$ for those n. Therefore, by the Gilbert-Varshamov bound, Theorem 12, Chapter 1 in [4], there is a binary linear code of length N with at most r parity checks, and the minimum distance at least d. The theorem is proved by an explicit construction of an $N \times r$ binary matrix P such that no $d-1$ rows are linearly dependent. To get a desired matrix H we expand P by arbitrary $n-r$ columns, and split rows of H into a collection X= $\{X_1, .., X_n\}$ of n pairs of vectors. Any

$$\lfloor (d-1)/2 \rfloor$$

of such pairs constitute a set of $2\lfloor (d-1)/2 \rfloor$ linearly independent vectors. For n sufficiently large we get

$$\min_{\pi} \max_{k} (\mathbf{rank}\left(\bigcup_{i=1}^{k} X_{\pi(i)}\right) - k) \geq \lfloor (d-1)/2 \rfloor \geq \frac{cn-2}{2} \geq \frac{n}{9.08861}.$$

The proof is complete. ■

At the moment we do not have a conjecture about the asymptotic rate of growth of the function $F_2(n,n)$. To indicate the difficulty of the problem we present a family exhibiting that a linear lower bound on $F_2(n,n)$ can be obtained even by a very special system.

Theorem 12 *For sufficiently large n , there is a positive constant c and a family $\mathcal{X}$ =$\{X_1, ..., X_n\}$ of pairs of binary vectors, where for all $i \in [n]$, $|X_i| = 2$, and X_i contains a unit vector and a vector with exactly two non-zero coordinates, such that*

$$\min_{\pi} \max_{1 \leq k \leq n} (\mathbf{rank}\left(\bigcup_{i=1}^{k} X_{\pi(i)}\right) - k) \geq cn,$$

where the minimum runs over all permutations on $[n]$.

Proof. To simplify notation, we construct a family of $3n$ sets of vectors, rather than just n of them. That is, the vector space V has dimension $3n$ and $\mathcal{X} = \{X_1, ..., X_{3n}\}$. We partition the set of $3n$ linearly independent unit vectors into three sets:

$$A = \{a_1, \ldots, a_n\}, \quad B = \{b_1, \ldots, b_n\}, \quad C = \{c_1, \ldots, c_n\}.$$

As in the proof of Theorem 14 in [2], we select two permutations σ and τ on $[n]$ at random, uniformly, and independently; that is, any permutation on $[n]$

coincides with each of σ and τ with probability 1/n!, and any ordered pair of permutations of $[n]$ coincides with (σ, τ) with probability $(1/n!)^2$. The following fact was proved in [2], with explicit values[1] of the constants c', c'', and q:

($*$) There exist positive constants c', c'', q such that the following property holds with probability at least q: in every ordering of the family of 3-element (multi)sets

$$\{\{i, \sigma(i), \tau(i)\} \mid i = 1, \ldots, n\}$$

the union of the first $\lfloor c'n \rfloor$ sets has cardinality at least $\lfloor c'n \rfloor + c''n$.

We now consider the collection of sets of vectors $\mathcal{X} = \{X_1, \ldots, X_{3n}\}$ over $A \cup B \cup C$ where, for $i = 1, \ldots, n$,

$X_{3i-2} = \{a_i, b_i + c_i\}$, $X_{3i-1} = \{a_i, b_i + c_{\sigma(i)}\}$, $X_{3i} = \{a_i, b_i + c_{\tau(i)}\}$.
We will prove that this $\mathcal{X}$ satisfies the stated inequality of the theorem (approximately for $c = \frac{1}{2}c''$) with positive probability; this immediately implies that there exists a suitable choice of $\mathcal{X}$.

First we observe some properties of the auxiliary random bipartite (multi)graph H, constructed from the pair (σ, τ), which has the vertex set $V(H) = B \cup C$ and the edge set

$$E(H) = E_0 \cup E_\sigma \cup E_\tau$$

where

$$E_0 = \{b_i c_i \mid i = 1, \ldots, n\},\ E_\sigma = \{b_i c_{\sigma(i)} \mid i = 1, \ldots, n\},\ E_\tau = \{b_i c_{\tau(i)} \mid i = 1, \ldots, n\}.$$

We call an edge of H a 0-edge, or σ-edge, or τ-edge, if it is in E_0, or E_σ, or E_τ, respectively.

Claim 1. For every $\ell \in \mathbb{N}$ there exists a constant d_ℓ such that the expected number of cycles of length 2ℓ in H is at most d_ℓ.

Proof. Consider any cycle, say C^*, of length 2ℓ in H. Quantitatively it can be associated with a triplet $(\ell_0, \ell_\sigma, \ell_\tau)$ which represents that C^* has ℓ_0 0-edges, ℓ_σ σ-edges, and ℓ_τ τ-edges. For every fixed ℓ there exist a bounded number of cycle types with respect to the positions of the three different kinds of edges — the number of possibilities for prescribing the edge types around C^* is clearly smaller than $3 \cdot 2^{2\ell}$ as each of $0, \sigma, \tau$ must correspond to a matching, but the actual exact number is unimportant for our purpose. There are $\binom{n}{\ell} = O(n^\ell)$ ways to specify the set $V(C^*) \cap B$, and those ℓ vertices can occur in $\ell!/2$ different orders along C^*. Once an order and a reference vertex of C^* are fixed, they specify the positions of 0-edges, and so they also determine ℓ_0 vertices of C^* in the vertex class C. Thus, the independent and uniform random selection of σ and τ implies that there are $\binom{n-\ell_0}{\ell-\ell_0} = O(n^{\ell-\ell_0})$ ways to choose the other $\ell - \ell_0$ vertices of C^* (and $(\ell - \ell_0)!$ ways to permute them). Consequently the

[1] One of the goals in [2] was to find as good estimates on c'' as possible, but in the present paper we only aim at proving that a positive constant exists as lower bound.

number of choices for $V(C^*)$ is $O(n^{2\ell-\ell_0})$. For any fixed selection of those 2ℓ vertices with their fixed order and specified positions of the edge types around the cycle, the probability that the prescribed edges are present in H is equal to $\left(\prod_{j=0}^{\ell_\sigma-1}(n-j)\cdot\prod_{k=0}^{\ell_\tau-1}(n-k)\right)^{-1} = O((n^{\ell_\sigma+\ell_\tau})^{-1}) = O(n^{-(2\ell-\ell_0)})$ as n gets large (where the 'hidden coefficient' in 'O' depends on ℓ). The expected number of cycles of a given length is not larger than the product of the probability and the number of possible choices, thus it does not exceed a suitably chosen constant d_ℓ. $\Diamond$

Let us fix now a positive constant q as in $(*)$.

Claim 2. For every $\ell \in \mathbb{N}$ it has probability at least $q/2$ that $(*)$ holds simultaneously with the property that, for every $k \le \ell$, the number of cycles of length $2k$ in H is at most $2\ell d_k/q$.

Proof. It is an elementary fact in probability theory that every positive-valued random variable with expectation $\mathbb{E}(\xi)$ satisfies the inequality $\mathbb{P}(\xi > s\cdot\mathbb{E}(\xi)) < 1/s$ for every $s > 1$. Applying this for $s = 2\ell/q$ and $k = 1,\dots,\ell$, the assertion follows. $\Diamond$

Returning to the proof of the theorem, we combine $(*)$ with Claim 2 and observe that the following event has probability at least $q/2$:

$(**)$ If ℓ is fixed (arbitrarily) and n is sufficiently large (with respect to ℓ) then the number of cycles of length at most 2ℓ in H is bounded from above by some constant $f(\ell)$; moreover, each $Y \subset B$ with $|Y| = \lfloor c'n\rfloor$ has at least $\lfloor c'n\rfloor + c''n$ neighbors in C.

So, we choose $\mathcal{X}$ (determined by a suitable choice of σ and τ) accordingly, and assume from now on that H satisfies $(**)$.

Let π be any permutation of $\{1,\dots,3n\}$, say $\pi = (i_1,\dots,i_{3n})$, and consider the sequence $X_{i_1},\dots,X_{i_{3n}}$ of sets of vectors. For any $k = 1,\dots,3n$, in the rest of this proof, with a slight abuse of notation let $b_{\pi(k)}$ mean the vertex $b_{\lceil i_k/3\rceil}$, that corresponds to the vector in the B-component of X_k. In order to define $c_{\pi(k)}$ with an analogous meaning, we write k in the form $k = 3i-3+j$ where $j \in \{1,2,3\}$. Then let $c_{\pi(k)} = c_i$ if $j = 1$, $c_{\pi(k)} = c_{\sigma(i)}$ if $j = 2$, and $c_{\pi(k)} = c_{\tau(i)}$ if $j = 3$. Now we set

$$B^{\le k} = \{b_{\pi(j)} \mid 1 \le j \le k\}.$$

That is, $B^{\le k}$ represents those unit vectors in B which have a contribution to the first k sets of vectors.

This also identifies the pairs $b_{\pi(j)}c_{\pi(j)}$ for $j = 1,\dots,k$, which we may view as the edges of a bipartite (multi)graph H_k. Denoting by k_1, k_2, and k_3 the number of vertices of H_k in B which have degree 1, 2, and 3, respectively, the equality $k = k_1 + 2k_2 + 3k_3$ is valid. If $k = 1$ then $k_2 = k_3 = 0$, and if $k = 3n$ then $k_3 = n$; moreover, the sum $k_2 + k_3$ either remains unchanged or increases by exactly 1 when we increase k by 1. Thus, we can and will fix a value k for

which $k_2 + k_3 = \lfloor c'n \rfloor$ holds in the graph H_k, where the positive constant c' is the one from $(*)$ and $(**)$.

We consider two modifications of H_k. The first one, denoted by H_k^-, is the subgraph of H_k that belongs to the subsequence obtained by removing those k_1 sets X_{i_j} which correspond to the degree-1 vertices of H_k. The second one, denoted by H_k^+, is the supergraph of H_k^- obtained by inserting those k_2 edges of H which correspond to later sets X_{i_j} in which a degree-2 vertex of H_k occurs. Hence both H_k^- and H_k^+ have $k_2 + k_3$ vertices in B; moreover, the number of edges in H_k^- is $2k_2 + 3k_3$, and the number of edges in H_k^+ is $3k_2 + 3k_3$.

Due to the partition $(A, B \cup C)$ concerning $\mathcal{X}$, the rank is a sum of two numbers, namely of the rank in A and in $B \cup C$. In A it is just $k_1 + k_2 + k_3$, independently of the situation in $B \cup C$. Moreover, linear independence for the vectors $b_{\pi(j)} + c_{\pi(j)}$ inside $B \cup C$ equivalently means that the corresponding subgraph is cycle-free. Thus, writing $comp(G)$ to denote the number of connected components in a graph G, we have

$$\mathbf{rank}\left(\bigcup_{i=1}^{k} X_{\pi(i)}\right) = (k_1 + k_2 + k_3) + (|V(H_k)| - comp(H_k)). \tag{11}$$

It will simplify the computation if we modify the formula from H_k to H_k^+ as

$$\begin{aligned} |V(H_k)| - comp(H_k) &= k_1 + |V(H_k^-)| - comp(H_k^-) \\ &\geq k_1 + |V(H_k^+)| - comp(H_k^+) - k_2 \end{aligned} \tag{12}$$

which is valid because the k_1 pendant edges are not involved in any cycles, and each of the k_2 additional edges of $E(H_k^+) \setminus E(H_k^-)$ may or may not increase the rank by 1 (at most), but they can never decrease it. The combination of (11) and (12) yields

$$\mathbf{rank}\left(\bigcup_{i=1}^{k} X_{\pi(i)}\right) \geq 2k_1 + k_3 + |V(H_k^+)| - comp(H_k^+), \tag{13}$$

so the problem is reduced to finding an appropriate lower bound on the difference $|V(H_k^+)| - comp(H_k^+)$.

We are going to give two kinds of lower bounds, their combination will prove the theorem. First, recall that there are $k_2 + k_3$ vertices in $B^{\leq k}$ which have degree at least 2 in H_k. All of them have degree 3 in H_k^+, therefore we can apply $(**)$ and obtain

$$|V(H_k^+)| \geq 2k_2 + 2k_3 + c''n. \tag{14}$$

For a different lower bound, we partition H_k^+ into the following three parts:

- H^1 – tree components with at most 2ℓ vertices for an integer ℓ to be defined later,
- H^2 – components containing at least one cycle and having at most 2ℓ vertices,

- H^3 – components with more than 2ℓ vertices.

For $j = 1, 2, 3$ let us denote $k_{\langle j \rangle} := |B \cap V(H^j)|$. Then, in particular, we have

$$k_{\langle 1 \rangle} + k_{\langle 2 \rangle} + k_{\langle 3 \rangle} = k_2 + k_3.$$

If a component F of H_k^+ is a tree, then it satisfies the equality $|C \cap V(F)| = 2|B \cap V(F)| + 1$; we shall use this fact for the components in H^1. Moreover, due to the presence of 0-edges $b_i c_i$, the subgraph $H^2 \cup H^3$ has at least as many vertices in C as in B. Consequently, since $k_{\langle 1 \rangle} \geq comp(H^1)$, we obtain:

$$\begin{aligned} |V(H_k^+)| &\geq 3k_{\langle 1 \rangle} + comp(H^1) + 2(k_{\langle 2 \rangle} + k_{\langle 3 \rangle}) \\ &\geq 2k_2 + 2k_3 + 2comp(H^1). \end{aligned} \tag{15}$$

Taking the arithmetic means on both sides of the equations (14) and (15) yields:

$$|V(H_k^+)| \geq 2k_2 + 2k_3 + \tfrac{1}{2}c''^1).$$

Substituting this formula into (13), moreover recalling that $k = k_1 + 2k_2 + 3k_3$ and noting that $comp(H_k^+) = comp(H^1) + comp(H^2) + comp(H^3)$, the following lower bound is derived:

$$\mathbf{rank}\left(\bigcup_{i=1}^{k} X_{\pi(i)}\right) \geq k + k_1 + \tfrac{1}{2}c''^2) - comp(H^3). \tag{16}$$

The property concerning short cycles, as described in $(**)$, immediately implies

$$comp(H^2) \leq f(\ell)$$

and, since the components of H^3 are large by definition, due to the choice $k_2 + k_3 = \lfloor c'n \rfloor$ we also have

$$comp(H^3) < \frac{c'n + n}{2\ell} < \frac{n}{\ell}.$$

Choosing ℓ large (e.g., $\ell = \lceil 3/c'' \rceil$), the latter two inequalities together with (16) complete the proof. ■

Acknowledgement. The authors are indebted to Noga Alon for discussions on expanders and on probabilistic methods, which lead to an improvement of the lower bound in Theorem 1, and to Øyvind Ytrehus for a discussion on the Gilbert-Varshamov bound.

References

[1] N. T. Courtois and J. Pieprzyk, Cryptanalysis of block ciphers with overdefined systems of equations in Advances of Cryptology, Asiacrypt 2002, LNCS 2501, Springer, 2002, 267-287.

[2] P. Horak and Zs. Tuza, Speeding up deciphering by hypergraph ordering, Designs, Codes and Cryptography 75(2015), 175-185.

[3] K. Iwama, Worst-Case Upper Bounds for kSAT, The Bulletin of the EATCS, 82(2004), 61-71.

[4] F. J. MacWilliams and N. J. A. Sloan, *The theory of error correcting codes*, North Holland Publishing Company, 1981.

[5] H. Raddum, MRHS equation systems, in: Selected Areas in Cryptography-SAC '07, 14th International Workshop (C. Adams et al., eds.), Lecture Notes in Comput. Sci., Vol. 4876, Springer-Verlag, Berlin, 2007, pp. 232–245.

[6] H. Raddum and I. Semaev, Solving multiple right-hand side equations, Designs, Codes and Cryptography 49(2008), 147-160.

[7] I. Semaev, On solving sparse algebraic equations over finite fields, Designs, Codes and Cryptography 49(2008), 47-60.

[8] I. Semaev, MaxMinMax problem and sparse equations over finite fields, Designs, Codes and Cryptography, DOI 10.1007/s10623-015-0058-6, also available at Cryptology ePrint Archive, report 2014/004.

[9] I. Semaev, New results in the linear cryptanalysis of DES, Cryptology ePrint Archive, report 2014/361.