COMBINATORICA
Bolyai Society – Springer-Verlag

COMBINATORICA 28 (2) (2008) 163–185
DOI: 10.1007/s00493-008-2195-2

PARITY CHECK MATRICES AND PRODUCT REPRESENTATIONS OF SQUARES

ASSAF NAOR, JACQUES VERSTRAËTE

Received March 8, 2005

Let $N_{\mathbb{F}}(n,k,r)$ denote the maximum number of columns in an n-row matrix with entries in a finite field $\mathbb{F}$ in which each column has at most r nonzero entries and every k columns are linearly independent over $\mathbb{F}$. We obtain near-optimal upper bounds for $N_{\mathbb{F}}(n,k,r)$ in the case $k>r$. Namely, we show that $N_{\mathbb{F}}(n,k,r) \ll n^{\frac{r}{2}+\frac{cr}{k}}$ where $c \approx \frac{4}{3}$ for large k. Our method is based on a novel reduction of the problem to the extremal problem for cycles in graphs, and yields a fast algorithm for finding short linear dependencies. We present additional applications of this method to a problem on hypergraphs and a problem in combinatorial number theory.

> *"It is certainly odd to have an instruction in an algorithm asking you to play with some numbers to find a subset with product a square ... Why should we expect to find such a subsequence, and, if it exists, how can we find it efficiently?"*
>
> Carl Pomerance [31].

1. Introduction

Since the mid-1930s it has been well established that there is a tight connection between combinatorial number theory and extremal combinatorics (see [32,12] and the references therein). The basic paradigm is that for certain number theoretic problems, one can construct a combinatorial object (e.g. a graph or a hypergraph), and prove that it cannot contain certain configurations (e.g. cycles). Thus, in many cases one can use theorems on

Mathematics Subject Classification (2000): 05Cxx, 05Dxx, 94B65 excluded configurations in extremal combinatorics to bound the size of the combinatorial structure that was constructed, and this translates back to give number theoretic consequences. The present paper develops a novel reduction of this type, and applies it to two problems in algebraic combinatorics, namely to coding theory and combinatorial number theory. Additionally, our proofs are constructive, and thus yield the best known algorithms for several natural computational problems.

Our original motivation comes from a problem in coding theory. Low density parity check codes were introduced by Gallager in the 1960s, and have since found numerous theoretical and practical applications in engineering and computer science (see [21,28,29] for an account of this theory. We also refer to [11] for a nice introduction to the geometry of codes). Given a linear code $C \subseteq \mathbb{F}_2^m$ of dimension ℓ and minimum Hamming weight t, an $(m-\ell)\times m$ matrix H is called a *parity check matrix* of the code C if $C=\{v\in\mathbb{F}_2^m : Hv=0\}$. The *Syndrome Decoding Algorithm* for such codes works as follows: given a corrupted signal z one computes the vector x of minimal weight satisfying $Hz=Hx$, and decodes z to $z-x$. This algorithm corrects at most $\frac{1}{2}t$ errors. As such computations are faster if the sparseness of H is exploited, it is desirable to obtain codes with sparse parity check matrices. Indeed, sparse parity check matrices occur in many of the known constructions of codes, e.g. codes based on bounded degree graphs such as expander codes [34,35], and we also refer to [27] for theoretical and experimental coding theory applications of very sparse matrices (we stress here that the present paper deals with a different range of parameters – our bounds will be for codes in which the minimal weight is not proportional to the dimension. Such codes occur in several contexts, e.g. certain BCH and Reed–Solomon codes [28], Turbo and Turbo-like codes [8,24,14,7]). Additionally, the above discussion makes sense for parity check matrices over arbitrary finite fields, which are also used in coding theory (see [28,29] for basic information on this topic, and [17] for empirical results on such codes). Finally, sparse parity check matrices are the key ingredient in the construction of small probability spaces and deterministic simulations of k-wise independent random variables, which are a key tool in derandomization [1,2, 4,10].

Somewhat surprisingly, in spite of their importance, there was a large gap between the known upper and lower bounds for the maximal number of columns of sparse parity check matrices. For a finite field $\mathbb{F}$ let $N_{\mathbb{F}}(n,k,r)$ be the maximal number of vectors in $\mathbb{F}^n$ with at most r non-zero coordinates such that no k of them are linearly dependent. Note that the linear independence condition corresponds to the fact that the kernel of the matrix whose rows are the given vectors is a code with minimal distance at least $k+1$. When $\mathbb{F}=\mathbb{F}_2$ we use the notation $N_{\mathbb{F}_2}(n,k,r)=N(n,k,r)$. The problem dealt with in this paper, namely that of estimating $N_{\mathbb{F}}(n,k,r)$, differs from the classical Gilbert–Varshamov bounds (see e.g. [28]), since the classical bounds on sizes of codes are geometric packing bounds which depend only on the minimum distance of the code. Here we introduce an additional algebraic restriction on the code – the existence of a sparse parity check matrix – which is motivated by computational issues. Thus, we are dealing with a mixture of a geometric and algebraic problem. In this paper we are primarily interested in the case that k and r are fixed and $n\to\infty$, although the results are valid for arbitrary k and r.

Throughout this paper we use the following notation: given two non-negative sequences $\{a_n\}_{n=1}^{\infty}$, $\{b_n\}_{n=1}^{\infty}$, we write $a_n \ll b_n$ if there exists a constant $C>0$ such that for all n, $a_n \le C\cdot b_n$.

1.1. Bounds on $N_{\mathbb{F}}(n,k,r)$

A probabilistic construction [26], using the first moment method, shows that

$$N(n,k,r) \gg n^{\frac{r}{2}+\frac{r}{2k-2}},$$

and this was generalized to arbitrary finite fields $\mathbb{F}$ in [25] to $N_{\mathbb{F}}(n,k,r)\gg n^{\frac{r}{2}+\frac{r}{2k-2}}$ for even k and $N_{\mathbb{F}}(n,k,r)\gg n^{\frac{r}{2}+\frac{r}{2k-4}}$ for odd k. When $k\ge 4$ is even, and $\gcd(k-1,r)=1$, the probabilistic lower bound above was improved in [9] to

$$N(n,k,r) \gg n^{\frac{r}{2}+\frac{r}{2k-2}} \cdot (\log n)^{\frac{1}{k-1}}.$$

This lower bound was generalized to arbitrary finite fields in [25], in which case the implicit constant also depends on the size of the field.

In [26] it was shown that when k is a power of 2, $N(n,k,r)\ll n^{\frac{1}{2}\left\lceil r+\frac{r}{k-1}\right\rceil}$, which coincides with the probabilistic lower bound (up to factors independent of n) when k–1 divides r. This upper bound was generalized to arbitrary finite fields in [25]. Observe that in the important case $k>r$, i.e., when the number of correctible errors is greater than the weight, this upper bound becomes: for k a power of 2, $N_{\mathbb{F}}(n,k,r)\ll n^{\frac{r}{2}+\frac{1}{2}}$. Thus the gap between the exponent of n in this bound and the probabilistic lower bounds deteriorates as k grows. Here we resolve this problem by proving the following theorem in Section 3:

Theorem 1.1. *For every integer $k\ge 8$ and every finite field $\mathbb{F}$*

$$N_{\mathbb{F}}(n,k,r) \ll n^{\frac{r}{2}+\frac{\lceil r/3\rceil}{2\lfloor k/8\rfloor}}$$

where the implied constant depends only on k, r *and* $|\mathbb{F}|$.

The exponent in the displayed inequality behaves roughly like $\frac{r}{2}+\frac{4r}{3k}$ when k is large. This should be compared with the exponent $\frac{r}{2}+\frac{r}{2k-2}$ in the probabilistic lower bound on $N_{\mathbb{F}}(n,k,r)$. In particular it follows from Theorem 1.1 that for any positive integer r,

$$\lim_{k\to\infty}\left[\liminf_{n\to\infty}\frac{\log N_{\mathbb{F}}(n,k,r)}{\log n}\right]=\lim_{k\to\infty}\left[\limsup_{n\to\infty}\frac{\log N_{\mathbb{F}}(n,k,r)}{\log n}\right]=\frac{r}{2}.$$

It is worthwhile to note here that the proof of Theorem 1.1 when r is even differs markedly from the proof in the case of odd r. In fact, it turns out that the case of odd r involves a substantially more subtle argument. The difference between these cases will be explained in Section 2. It is likely to be difficult to determine the exact values of the bracketed terms above for all k and r (the answer probably depends on arithmetic properties of k relative to r).

The proof of Theorem 1.1 is based on a novel reduction of the problem to the following Turán type problem: What is the maximum number of edges in an n-vertex graph which doesn't contain an even cycle of length $2k$? We then employ recent results on this problem [36, 3, 23, 30] to deduce bounds on $N_{\mathbb{F}}(n,k,r)$. Some of the previous results on $N_{\mathbb{F}}(n,k,r)$ reduced the problem to the study of certain Turán type questions on hypergraphs (see [9, 7]). Since very little is known on hypergraph Turán problems, our main contribution is the method of reducing such questions to a problem on graphs. We believe that this approach is of independent interest. Indeed, as an example we apply our result to a problem in combinatorial number theory, improving a theorem of Erdős, Sárközy and Sós [20] (this application relies heavily on the more difficult part of Theorem 1.1, namely the case of odd r). We also apply our result to the even cover problem for hypergraphs.

1.2. Even covers in hypergraphs

The methods used to prove Theorem 1.1 in the case $\mathbb{F}=\mathbb{F}_2$ extend the Even Cycle Theorem of Erdős [13] to hypergraphs. If we rephrase the theorem in the terminology of hypergraphs, then we define an *even cover* to be a non-empty collection of sets each point of which is in an even number of sets. For example, a graph contains an even cover of size at most k if and only if it has girth at most k. Applying Theorem 1.1 with $\mathbb{F}=\mathbb{F}_2$ to the incidence vectors of edges in a hypergraph, we obtain the following theorem:

Theorem 1.2. *Let $\mathcal{S}$ be a hypergraph on n points whose edges have size at most r each, and which does not contain an even cover of size k. Then*

$$|\mathcal{S}| \ll n^{\frac{r}{2}+\frac{\lceil r/3\rceil}{2\lfloor k/8\rfloor}}.$$

As an example of an application to extremal hypergraph theory, it is notoriously difficult to determine which configurations of triples are guaranteed to appear in every large enough Steiner triple system (for example, many of the questions discussed in [16] remain open). In the present context, it is known that there are infinitely many Steiner triple systems which do not contain an even cover of four triples (one is constructed in [26]). Using Theorem 1.2, we can at least guarantee small even covers in Steiner triple systems. For if $\mathcal{S}$ is a Steiner triple system on n points, then $\mathcal{S}$ has $\frac{1}{3}\binom{n}{2}$ edges [15], which is larger than the bound in Theorem 1.2 for $r=3$ and $k=16$, so every large enough Steiner triple system contains an even cover of size at most sixteen. We leave open the problem of determining if smaller even covers exist in every Steiner triple system.

1.3. Product representations of squares

Denote by $\mathrm{Rep}_k(n)$ the largest N such that there exists $A \subseteq \{1,\dots,n\}$ with $|A| = N$ such that for every $1 \le \ell \le k$ there are no distinct $a_1,\dots,a_\ell \in A$ and $x \in \mathbb{Z}$ satisfying $a_1 \cdot a_2 \cdots a_\ell = x^2$. The behavior of the sequence $\mathrm{Rep}_k(n)$ has been studied by several authors [20, 33, 22]. One of the motivations for studying this sequence is that the problem of finding product representations of squares is a key step in certain sub-exponential factoring algorithms (see the surveys [37, 31] for an account of this fascinating field, and [18] for the first rigorous analysis of a sub-exponential randomized factoring algorithm). In these algorithms the goal is to find efficiently a subset of a certain set of integers whose product is a square – the sets that are analyzed are carefully constructed so that such a product representation is guaranteed to exist, but it is of interest to ask how large such a set should be in order to ensure the existence of the required product representation. Moreover, once the cardinality of a set is above this threshold, one would like to efficiently find such a product representation. The best known results on this question follow from the work of Erdős [19] and Erdős, Sárközy and Sós [20], who showed that for every even $k \ge 6$,

$$\left(\frac{n}{(\log n)^2}\right)^{\frac{1}{2}+\frac{2}{3k}} \ll \mathrm{Rep}_k(n) - \pi(n) \ll \left(\frac{n}{(\log n)^2}\right)^{\frac{2}{3}},$$

where the implied constants are absolute and $\pi(n)$ is the number of primes less than or equal to n. Here we show that for each $\epsilon > 0$, if k is large enough, then $\mathrm{Rep}_k(n) - \pi(n)$ is of order at most $n^{\frac{1}{2}+\varepsilon}$. In Section 4, we prove the following theorem:

Theorem 1.3. *For every $k \geq 1$ and every integer n,*

$$\left(\frac{n}{(\log n)^2}\right)^{\frac{1}{2}+\frac{1}{12k}} \ll \mathrm{Rep}_{8k}(n) - \pi(n) \ll kn^{\frac{1}{2}+\frac{1}{2k}}(\log n)^2.$$

1.4. An algorithm for linear dependencies

Our proof of Theorem 1.1 yields a fast algorithm for finding a small linear dependency in large sets of vectors of weight r in $\mathbb{F}^n$. More precisely, the proof of Theorem 1.1 for r even gives a linear time (in $|X|$) algorithm for finding a linear dependency of size at most $2k$ in a set $X \subseteq \mathbb{F}^n$ such that

$$|X| > 2k\binom{n}{\frac{r}{2}}^{1+\frac{1}{k}} + 4k\binom{n}{\frac{r}{2}}.$$

For general r a quadratic time (in $|X|$) algorithm can be deduced from the proof of Theorem 1.1. These algorithms follow directly from our proof of Theorem 1.1, together with the Alon–Yuster–Zwick algorithm [5,6] for finding cycles in graphs, or alternatively the proof of the main theorem in [36], which gives better constants. Additionally, our proof yields an algorithm such that given a set $A \subseteq \{1,\ldots,n\}$, with $|A|$ much larger than the upper bound in Theorem 1.3 finds, in quadratic time in $|A|$, distinct $a_1,\ldots,a_j \in A$ with $1 \leq j \leq 8k$, such that $a_1 a_2 \cdots a_j$ is a square.

2. Forbidden cycles and sparse parity check matrices

Throughout this section, $\mathbb{F}$ is a finite field and $\mathbb{F}^* = \mathbb{F} \setminus \{0\}$ is the set of non-zero elements of $\mathbb{F}$. A set $X \subseteq \mathbb{F}^n$ is *k-wise independent* if no k vectors in X are linearly dependent. A vector $v \in \mathbb{F}^n$ is said to have *weight* r if it has exactly r non-zero coordinates. Then $N_{\mathbb{F}}(n,k,r)$ is the maximum size of a set of k-wise independent vectors of weight at most r in $\mathbb{F}^n$. The following result on forbidden cycles will be used in the proof of Theorem 1.1:

Theorem 2.1. *Let k be a positive integer and let $G = (V,E)$ be an N-vertex graph. If G contains no cycle of length exactly $2k$ then*

$$|E| < 2kN^{1+\frac{1}{k}}. \tag{1}$$

If G is an M by N bipartite graph containing no cycle of length $2k$, then

$$(2) \qquad |E| < 2k \cdot [M^{\frac{1}{2}} N^{\frac{1}{2}+\frac{1}{k}} + M + N].$$

If G has girth at least $2k+1$, then the factor $2k$ may be removed in each of the upper bounds.

Theorem 2.1 is proved in [30], the last statement is proved in [3], the bipartite case of which is proved in [23]. The first assertion (1), with the same dependence on n but a worse constant, is Erdős' Even Cycle Theorem – see [13].

2.1. Vectors of even weight

We will reduce the problem of estimating $N_{\mathbb{F}}(n,k,r)$ to the bounds in Theorem 2.1. The reduction we give is simple in the case that r is even, but substantially more involved when r is odd. The first theorem we prove is as follows:

Theorem 2.2. *Let n, k and r be positive integers, and define*

$$M = \sum_{\ell=0}^{\lfloor r/2 \rfloor} (|\mathbb{F}|-1)^{\ell} \binom{n}{\ell}, \qquad N = \sum_{\ell=0}^{\lceil r/2 \rceil} (|\mathbb{F}|-1)^{\ell} \binom{n}{\ell}.$$

Let $X \subseteq \mathbb{F}^n$ be a set of vectors of weight at most r such that

$$|X| > 2k \cdot \left[M^{\frac{1}{2}} N^{\frac{1}{2}+\frac{1}{k}} + M + N\right].$$

Then there are disjoint sets $A, B \subseteq X$, each of size k, such that $\sum_{a \in A} a = \sum_{b \in B} b$. In particular, there is a set of $2k$ linearly dependent vectors in X.

The bound in Theorem 2.2 implies in particular that $N_{\mathbb{F}}(n,2k,r) \ll n^{\frac{r}{2}+\frac{1}{2}}$ for all $k > r$. This generalizes the same bound which was proved for k a power of two in [26] to all values of $k > r$. The proof of Theorem 2.2 is short, and we present it here:

Proof of Theorem 2.2. For each vector $v \in X$ of weight $\omega \leq r$, fix a pair $e(v) = \{x, y\}$ of vectors in $\mathbb{F}^n$ of weights at most $\lfloor \omega/2 \rfloor$ and $\lceil \omega/2 \rceil$, respectively, satisfying $x + y = v$. If r is even, let G be the graph whose vertex set consists of all vectors in $\mathbb{F}^n$ of weight at most $r/2$ and whose edge set is $E = \{e(v) : v \in X\}$. Then $|E| = |X|$ and G has M vertices. By the first assertion (1) in Theorem 2.1, G contains a cycle of length $2k$. By the definition of G, there exist distinct vectors $v_1, v_2, \ldots, v_{2k} \in X$ such that the edge set of this cycle consist of the pairs $e(v_i)=\{x_i,x_{i+1}\}$ for $i=1,2,\dots,2k$, where $x_{2k+1}=x_1$. Then

$$\sum_{\ell=1}^{2k}(-1)^{\ell+1}v_i=(x_1+x_2)-(x_2+x_3)+\cdots-(x_{2k}+x_1)=0,$$

and the disjoint sets $A=\{v_1,v_3,\dots,v_{2k-1}\}$ and $B=\{v_2,v_4,\dots,v_{2k}\}$ have the same sum. If r is odd, then G is a bipartite graph whose parts have sizes M and N. By the second assertion (2) of Theorem 2.1, $C_{2k}\subseteq G$, and we conclude by the same argument as that presented above. ∎

Theorem 2.2 implies Theorem 1.1 in the case of even r: in fact in this case we obtain the stronger bound $N_{\mathbb{F}}(n,k,r)\ll n^{\frac{r}{2}+\frac{r}{k}}$ for even k. We also remark that the proof above gives a linear time algorithm – that is time linear in $|X|$ – for finding $2k$ linearly dependent vectors in a set $X\subseteq\mathbb{F}^n$ satisfying the requirements of the theorem – this follows from the linear time algorithm in [5,6] for finding a cycle of length $2k$ in a graph with appropriately many edges.

2.2. Vectors of odd weight

In the case of odd r, Theorem 2.2 is insufficient to deduce Theorem 1.1, since the rounding up of $\frac{r}{2}$ only yields an upper bound of order $n^{\frac{r}{2}+\frac{1}{2}}$. This difference between even and odd values of r leads to a more involved argument in the case of odd r, which nevertheless yields the bounds of Theorem 1.1 in this case as well. We deduce Theorem 1.1 by proving the following refinement of Theorem 2.2 when $k>r$ and r is odd. For convenience we define $[\frac{r}{3}]=r-\lceil\frac{r}{3}\rceil-\lfloor\frac{r}{3}\rfloor$ and $q=|\mathbb{F}|$:

Theorem 2.3. *Let n,k and r be positive integers, and define*

$$L=\sum_{\ell=0}^{\lfloor\frac{r}{3}\rfloor}(q-1)^\ell\binom{\lfloor n/3\rfloor}{\ell},\ M=\sum_{\ell=0}^{[\frac{r}{3}]}(q-1)^\ell\binom{[n/3]}{\ell},\ N=\sum_{\ell=0}^{\lceil\frac{r}{3}\rceil}(q-1)^\ell\binom{\lceil n/3\rceil}{\ell}.$$

Let $X\subseteq\mathbb{F}^n$ be an $8k$-wise independent set of vectors, each of weight at most r. Then

$$|X|<12kr\cdot\left[(LN)^{\frac{1}{2}}M^{\frac{1}{2}+\frac{1}{2k}}+N^{1+\frac{1}{2k}}\right].$$

In particular, it follows that $N_{\mathbb{F}}(n,8k,r)\ll n^{\frac{r}{2}+\frac{\lceil r/3\rceil}{2k}}$, which gives Theorem 1.1 when we replace k with $\lfloor k/8\rfloor$. The remainder of the paper is devoted to the proof of Theorem 2.3.

3. Proof of Theorem 1.1

We already proved Theorem 1.1 in the case that r is even. In this section we prove Theorem 2.3 which implies Theorem 1.1 when r is odd. Since the proof is quite involved, we begin with an outline of the proof for $q=2$ and $r=3$, and suppose $3 \mid n$. In that case, $L=M=N=\frac{1}{3}n$. For simplicity we omit all the multiplicative constants.

3.1. Proof Outline

We wish to find a linearly dependent set of $8k$ vectors in $X \subseteq \mathbb{F}^n$ whenever $|X| \gg N^{\frac{3}{2}+\frac{1}{2k}}$. Suppose, for a contradiction, that there is no such subset of X. By randomly partitioning the coordinates of $\mathbb{F}^n$, we find a set Z of vectors in X such that $|Z| \gg |X|$ and each vector in Z is the concatenation of three vectors in $\mathbb{F}^N$. For convenience, we view Z as a subset of $V = \mathbb{F}^N \oplus \mathbb{F}^N \oplus \mathbb{F}^N$. Write $V = U \oplus W$, where all vectors in $U = \mathbb{F}^N \oplus \{0\}^N \oplus \mathbb{F}^N$ have zero in the middle N coordinates and W is the orthogonal complement of U. The next step is to show (see §3.3) that there is a set $Y \subseteq Z$ where

$$|Y| \gg |Z| - N^{1+\frac{1}{2k}}$$

and such that the projection of Y onto U is an injection. In this case we say that Y is determined by projection. This allows us in §3.4 to count special four-element subsets of Y called partially dependent quadruples: these are special sets of four vectors in Y whose projection onto U is a linearly dependent set over $\mathbb{F}$. The key point in the proof is to prove that there are enough of these quadruples in Y to ensure that the projection onto W of all $8k$ vectors in some set of $2k$ quadruples is a linearly dependent set satisfying the same linear dependence as the $8k$ vectors in the quadruples (§3.5–§3.10). Since $V = U \oplus W$, this gives $8k$ linearly dependent vectors in Y, and therefore in X. This contradiction will complete the proof.

3.2. Calculations

To obtain a feeling for where the bound in Theorem 2.2 comes from, at least for $q=2$ and $r=3$, we give some of the bounds involved in the proof. Let $\mathcal{Q}(Y)$ denote the set of partially dependent quadruples in Y. We will show in §3.4 that

$$|\mathcal{Q}(Y)| \gg \frac{|Y|^4}{N^4}.$$

Now we use the lower bound on $|X|$ in the theorem, and the above lower bound for Y:

$$|Y| \gg |Z| - N^{1+\frac{1}{2k}} \gg |X| - N^{1+\frac{1}{k}} \gg N^{\frac{3}{2}+\frac{1}{2k}}$$

so the inequality for $|\mathcal{Q}(Y)|$ becomes $|\mathcal{Q}(Y)| \gg N^{2+\frac{2}{k}}$. Since Z is balanced and $Y \subseteq Z$, the projection of Y onto W consists of vectors of weight one. So the projection of each quadruple in $\mathcal{Q}(Y)$ onto W consists of four vectors of weight one. If we treat these projections of quadruples as vectors of weight four in $\mathbb{F}^N$, then there are $|\mathcal{Q}(Y)| \gg N^{2+\frac{2}{k}}$ of these vectors. Barring technical details, we could then apply Theorem 2.2 with $r=4$ to get a linear dependency of $2k$ of these vectors of weight four, and the $8k$ vectors in the corresponding quadruples in $Y \subseteq X$ satisfy a (carefully defined) linear dependency. Due to the possibility of trivial linear dependencies (where the coefficients are all zero modulo the characteristic of $\mathbb{F}$), we carefully modify the proof of Theorem 2.2 in §3.6–§3.10 to avoid the situation of trivial dependencies.

3.3. Sets determined by projection

We consider the vector space $\mathbb{F}^\alpha \oplus \mathbb{F}^\beta \oplus \mathbb{F}^\gamma$ where α, β, γ are arbitrary positive integers. At the end of the proof, we will put $\alpha = L$, $\beta = M$ and $\gamma = N$. For convenience, let $\lambda = |\mathbb{F}^*|\alpha$, $\mu = |\mathbb{F}^*|\beta$ and $\nu = |\mathbb{F}^*|\gamma$. For $u \in \mathbb{F}^\alpha$ and $v \in \mathbb{F}^\beta$, write $u \oplus v$ for the concatenation of u and v, which is a vector in $\mathbb{F}^\alpha \oplus \mathbb{F}^\beta$. Throughout the proof, a vector $w \in \mathbb{F}^\alpha \oplus \mathbb{F}^\beta \oplus \mathbb{F}^\gamma$ will be written canonically as $w_\alpha \oplus w_\beta \oplus w_\gamma$, where $w_\alpha \in \mathbb{F}^\alpha$, $w_\beta \in \mathbb{F}^\alpha$, and $w_\gamma \in \mathbb{F}^\alpha$. A set $Z \subseteq \mathbb{F}^\alpha \oplus \mathbb{F}^\beta \oplus \mathbb{F}^\gamma$ is *balanced* if $w_\alpha \oplus w_\beta \oplus w_\gamma \in Z$ implies w_α, w_β and w_γ all have exactly one non-zero co-ordinate i.e., they have weight one.

Projections. The *projection* of a set Z onto $\mathbb{F}^\alpha$ is $Z_\alpha = \{z_\alpha : z \in Z\}$. Similarly, $Z_{\alpha\beta} = \{z_\alpha \oplus z_\beta : z \in Z\}$ denotes the projection of Z onto $\mathbb{F}^\alpha \oplus \mathbb{F}^\beta$. The projections $Z_\beta, Z_\gamma, Z_{\alpha\gamma}$ and $Z_{\beta\gamma}$ are defined similarly. For $v \in \mathbb{F}^\alpha$, the *lift* of v into Z is $Z(v) := \{z \in Z : z_\alpha = v\}$.

Sets Determined by Projection. A vector $v \in \mathbb{F}^\alpha \oplus \mathbb{F}^\beta \oplus \mathbb{F}^\gamma$ is *determined by projection* if any two of $v_\alpha, v_\beta, v_\gamma$ uniquely determine v. A set Y is determined by projection if every $y \in Y$ is determined by projection.

The following lemma says that $8k$-wise independent sets contain large subsets which are determined by projection:

Lemma 3.1. *Let $Z \subseteq \mathbb{F}^\alpha \oplus \mathbb{F}^\beta \oplus \mathbb{F}^\gamma$ be a balanced $8k$-wise independent set. Then there exists a set $Y \subseteq Z$ such that Y is determined by projection and*

$$|Y| > |Z| - \lambda^{1+\frac{1}{2k}} - \mu^{1+\frac{1}{2k}} - \nu^{1+\frac{1}{2k}}. \tag{3}$$

Proof. For $v \in Z_{\alpha\beta}$, let $T(v)$ be a spanning tree of the complete graph on $Z(v)_\gamma$. So $|E(T(v))| = |Z(v)_\gamma| - 1$. We claim that the trees $\{T(v)\}_{v \in Z_{\alpha\beta}}$ can be chosen so that the *multigraph* $G_{\alpha\beta}$ consisting of all edges in all the trees $\{T(v)\}_{v \in Z_{\alpha\beta}}$ has girth greater than $4k$. To see this, choose the trees $\{T(v)\}_{v \in Z_{\alpha\beta}}$ so that the girth of $G_{\alpha\beta}$ is a minimum. This choice implies that if C is a shortest cycle in G, then $|E(C) \cap E(T(v))| \leq 1$ for all $v \in Z_{\alpha\beta}$. Here $E(C)$ and $E(T(v))$ denote the edge-sets of C and $T(v)$. We aim to show that $|C| > 4k$. Suppose the edge set of C is $\{\{w_1, w_2\}, \{w_2, w_3\}, \dots, \{w_\ell, w_1\}\}$. Then there are distinct vectors $v_j = v^j_\alpha \oplus v^j_\beta \in Z_{\alpha\beta}$ such that $\{w_j, w_{j+1}\} \in T(v_j)$ for all $j \leq \ell$. Let $x_j = v_j \oplus w_j \in Z$ and $y_j = v_j \oplus w_{j+1} \in Z$ for $j \leq \ell$. Then

$$\sum_{j=1}^{\ell} x_j - \sum_{j=1}^{\ell} y_j = 0.$$

Now x_i and y_j are distinct for all $i, j \leq \ell$. This means that $\{x_1, x_2, \dots, x_\ell, y_1, y_2, \dots, y_\ell\} \subseteq Z$ is linearly dependent. Since Z is $8k$-wise independent, it follows that $\ell > 4k$, as required, so G has girth greater than $4k$. The number of edges in $G_{\alpha\beta}$ is

$$|E(G)| = \sum_{v \in Z_{\alpha\beta}} (|Z(v)_\gamma| - 1). \tag{4}$$

On the other hand, since $G_{\alpha\beta}$ has at most ν vertices (since Z is balanced), we conclude from Theorem 2.1 that $|E(G_{\alpha\beta})| < \nu^{1+\frac{1}{2k}}$. We may define $G_{\beta\gamma}$ and $G_{\alpha\gamma}$ similarly, and by symmetry $|E(G_{\beta\gamma})| < \lambda^{1+\frac{1}{2k}}$. $|E(G_{\alpha\gamma})| < \mu^{1+\frac{1}{2k}}$. Using (4), these inequalities translate to

$$\sum_{v \in Z_{\alpha\beta}} (|Z(v)_\gamma| - 1) < \nu^{1+\frac{1}{2k}},$$
$$\sum_{v \in Z_{\alpha\gamma}} (|Z(v)_\beta| - 1) < \mu^{1+\frac{1}{2k}},$$
$$\sum_{v \in Z_{\beta\gamma}} (|Z(v)_\alpha| - 1) < \lambda^{1+\frac{1}{2k}}.$$

Finally, the number of vectors in Z which are *not* determined by projection is exactly the sum of the three terms on the left in the above inequalities.

So the number of vectors in Z which are determined by projection is greater than

$$|Z| - \lambda^{1+\frac{1}{2k}} - \mu^{1+\frac{1}{2k}} - \nu^{1+\frac{1}{2k}}.$$

Let Y be the set of all these vectors; then Y satisfies the requirements of the lemma. ∎

3.4. Partially dependent quadruples

For the remainder of the proof of Theorem 2.3, we restrict our attention to a set $Y \subseteq Z$ which is $8k$-wise independent and determined by projection, and satisfies (3). A *partially dependent quadruple* in Y is a quadruple $\{w, x, y, z\} \subseteq Y$ such that $(w_\alpha, x_\gamma, y_\alpha, z_\gamma) = (x_\alpha, y_\gamma, z_\alpha, w_\gamma)$. We write $\{w, x, y, z\}$ with the understanding that the non-zero co-ordinate of w_α precedes the non-zero co-ordinate of y_α and the non-zero co-ordinate of w_γ precedes the non-zero co-ordinate of y_γ in some prescribed ordering of the coordinates. A partially dependent quadruple is illustrated in Figure 1.

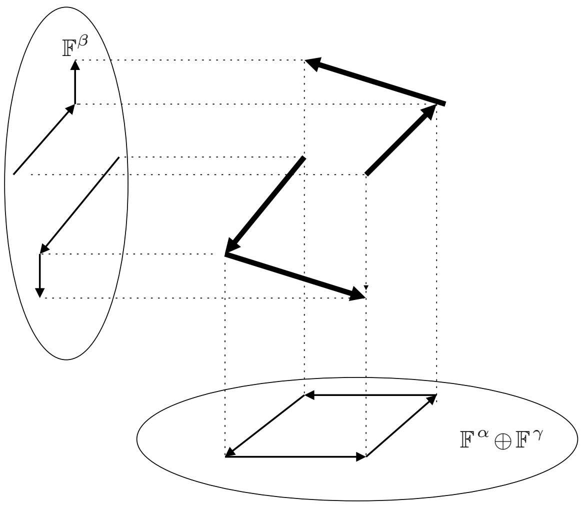

Figure 1. A partially dependent quadruple

The key point is that the projection $Q_{\alpha\gamma}$ of a partially dependent quadruple $Q \subset Y$ is a linearly dependent set in $\mathbb{F}^\alpha \oplus \mathbb{F}^\gamma$. This is because the projection $Q_{\alpha\gamma}$ consists of the vectors

$$w_\alpha \oplus w_\gamma \quad w_\alpha \oplus x_\gamma \quad y_\alpha \oplus x_\gamma \quad y_\alpha \oplus w_\gamma$$

and therefore we get a linear dependency in $\mathbb{F}^{\alpha} \oplus \mathbb{F}^{\gamma}$:

$$w_{\alpha} \oplus w_{\gamma} - w_{\alpha} \oplus x_{\gamma} + y_{\alpha} \oplus x_{\gamma} - y_{\alpha} \oplus w_{\gamma} = 0.$$

In the rest of the proof, our aim is to find a linearly dependent set of $2k$ projections of quadruples in Y onto $\mathbb{F}^{\beta}$. Since each quadruple consists of four vectors in Y, altogether this gives $8k$ linearly dependent vectors in Y. Let $\mathcal{Q}$ denote the set of partially dependent quadruples in Y.

Lemma 3.2. *Suppose that* $|Y| > \nu + 2\lambda\nu^{\frac{1}{2}}$. *Then*

$$|\mathcal{Q}| > \frac{|Y|^4}{4\lambda^2\nu^2}. \tag{5}$$

Proof. For $v \in \mathbb{F}^{\alpha}$, let Y_{γ}^{v} denote the projection of Y^{v} onto $\mathbb{F}^{\gamma}$. Recall that for $w \in \mathbb{F}^{\gamma}$, $Y^{w} = \{y \in Y : y_{\gamma} = w\}$. To prove (5), we use the following identity:

$$\sum_{\{v,x\} \subseteq \mathbb{F}^{\alpha}} \left|Y_{\gamma}^{v} \cap Y_{\gamma}^{x}\right| = \sum_{w \in \mathbb{F}^{\gamma}} \binom{|Y^{w}|}{2}. \tag{$*$}$$

To see this, draw a multigraph on $\mathbb{F}_{\alpha}$ by adding the edge $\{v, x\}$ with multiplicity $|Y_v \cap Y_x|$. Then the left hand side of $(*)$ is the number of edges in this multigraph. On the other hand, for each $w \in \mathbb{F}_{\gamma}$, we have drawn the complete graph on Y_{α}^{w}. Since Y is determined by projection, $|Y_{\alpha}^{w}| = |Y^{w}|$. This proves the identity. Let Γ be the left hand side and Δ the right hand side of the identity. Then

$$\begin{aligned} |\mathcal{Q}| &= \sum_{\{v,x\} \subseteq \mathbb{F}^{\alpha}} \binom{|Y_{\gamma}^{v} \cap Y_{\gamma}^{x}|}{2} \\ &\geq \binom{\lambda}{2} \binom{\frac{\Gamma}{\binom{\lambda}{2}}}{2} \\ &\overset{(*)}{=} \binom{\lambda}{2} \binom{\frac{\Delta}{\binom{\lambda}{2}}}{2} \\ &\geq \binom{\lambda}{2} \binom{\frac{\nu}{\binom{\lambda}{2}} \binom{\frac{|Y|}{\nu}}{2}}{2} \\ &> \frac{|Y|^4}{4\lambda^2\nu^2}. \end{aligned}$$

This is exactly (5). In each of the inequalities we used the convexity of the function $a \mapsto \binom{a}{2}$. In the last inequality we used the lower bound on $|Y|$ assumed in the lemma. ∎

3.5. Chains of quadruples

Let $Q_1, Q_2, \dots, Q_{2k}$ be partially dependent quadruples in Y, and $U = Q_1 \cup Q_2 \cup \cdots \cup Q_{2k}$. Then $U_{\alpha\gamma}$ is a linearly dependent set of at most $8k$ vectors in $\mathbb{F}^\alpha \oplus \mathbb{F}^\gamma$. We want to make sure that for some choice of U, the vectors in U_β satisfy the same linear dependence in $\mathbb{F}^\beta$ as the vectors in $U_{\alpha\gamma}$, so that U itself is a linearly dependent set. This will be done by ensuring that U_β is a closed walk in a certain multigraph G. Define the multigraph G on $\mathbb{F}^\beta \times \mathbb{F}^\beta$ by adding the edge $\pi(Q) = \{(w_\beta, y_\beta), (x_\beta, z_\beta)\}$ for each partially dependent quadruple $Q = \{w, x, y, z\} \in \mathcal{Q}$. Here we are using the standard ordering of the vectors in a partially dependent quadruple, namely that $(w_\alpha, x_\gamma, y_\alpha, z_\gamma) = (x_\alpha, y_\gamma, z_\alpha, w_\gamma)$. A non-returning walk of length k in G is a sequence $(\pi(Q_1), \pi(Q_2), \dots, \pi(Q_k))$ of edges of G such that $\pi(Q_i) \neq \pi(Q_{i+1})$ for all $i < k$. A *chain of length* k is a sequence $Q = (Q_1, Q_2, \dots, Q_k)$ where $Q_i \in \mathcal{Q}$ such that $(\pi(Q_1), \pi(Q_2), \dots, \pi(Q_k))$ is a non-returning walk of length k in G. Two chains are *concyclic* if the walks corresponding to them in G have the same endpoints.

3.6. Constructing linear dependencies

If $Q = (Q_1, Q_2, \dots, Q_k)$ and $R = (R_1, R_2, \dots, R_k)$ are concyclic chains, where $Q_i = \{s_i, t_i, u_i, v_i\}$ and $R_i = \{w_i, x_i, y_i, z_i\}$, then $\{(s_1)_\beta, (u_1)_\beta\} = \{(w_1)_\beta, (y_1)_\beta\}$ and $\{(t_k)_\beta, (v_k)_\beta\} = \{(x_k)_\beta, (z_k)_\beta\}$. It follows that

$$\sum_{i=1}^{k} (s_i + u_i - t_i - v_i) = \sum_{i=1}^{k} (w_i + y_i - x_i - z_i). \tag{6}$$

Let $U = \bigcup_{i=1}^{k} Q_i \cup R_i$. Then the vectors in U satisfy (6), so we get a linear dependency of size at most $8k$, unless every vector appears in the equation (6) with a coefficient which is zero modulo p, where p is the characteristic of $\mathbb{F}$. In that case, (6) is a trivial linear dependency, and we wish to avoid this to obtain a contradiction to the assumption that X is $8k$-wise independent. This technical scenario is what we deal with next.

3.7. Reduction of chains

The *reduction* of a chain $Q = (Q_1, Q_2, \dots, Q_k)$ where $Q_i = \{s_i, t_i, u_i, v_i\}$ is the set of vectors

$$\widehat{Q} := Q_k \triangle_p Q_{k-1} \triangle_p \cdots \triangle_p Q_1. \tag{7}$$

The expression above is read from right to left, and the symmetric difference operator $\triangle_p$ is defined as follows: we delete any vector once it appears with coefficient zero mod p in the sum:

$$\sum_{i=1}^{k}(s_i + u_i - t_i - v_i).$$

In order that (6) does not give a trivial linear dependency, it is sufficient to find concyclic chains Q and R such that $\widehat{Q} \neq \widehat{R}$.

Lemma 3.3. *Suppose Q, R are concyclic chains of length k in $\mathcal{Q}$. Then $\widehat{Q} = \widehat{R}$.*

Proof. If $\widehat{Q} \neq \widehat{R}$, then $\widehat{Q} \triangle \widehat{R}$ (here the symmetric difference is setwise) is a non-empty linearly dependent set of at most $8k$ vectors in Y, a contradiction. ∎

3.8. Degenerate chains

To find concyclic chains Q and R of length k with $\hat{Q} \neq \hat{R}$, we count so-called nondegenerate chains. A chain $Q = (Q_1, Q_2, \dots, Q_k)$ is *nondegenerate* if

$$|Q_i \cap (Q_1 \cup Q_2 \cup \cdots \cup Q_{i-1})| \le 1 \quad \text{for} \quad i = 1, 2, \dots, k$$

and Q is *degenerate* otherwise. The following crucial lemma says that if we grow a chain of length k by adding quadruples one by one, then at each stage only a few quadruples will cause the chain to become degenerate.

Lemma 3.4. *If $Q = (Q_1, Q_2, \dots, Q_\ell)$ is a chain, then at most $8\ell^2$ degenerate chains of length $\ell + 1$ contain Q.*

Proof. Let (w_β, y_β) be the endvertex of the non-returning walk $(\pi(Q_1), \pi(Q_2), \dots, \pi(Q_\ell))$ in G. We want to count the number of choices of a partially dependent quadruple

$$\begin{aligned} Q_{\ell+1} &= \{w, x, y, z\} \\ &= \{w_\alpha \oplus w_\beta \oplus w_\gamma, y_\alpha \oplus x_\beta \oplus w_\gamma, y_\alpha \oplus y_\beta \oplus y_\gamma, w_\alpha \oplus z_\beta \oplus y_\gamma\} \end{aligned} \tag{8}$$

such that $(Q_1, Q_2, \dots, Q_\ell, Q_{\ell+1})$ is degenerate. Since the chain is degenerate, we can find two vectors of $Q_{\ell+1}$ which are also in $U := Q_1 \cup Q_2 \cup \cdots \cup Q_\ell$.

We claim that $Q_{\ell+1}$ is uniquely determined once we fix two vectors of $Q_{\ell+1}$ in U. By symmetry we need only consider the cases (1) $\{w, y\} \subseteq$

$Q_{\ell+1}\cap U$ and (2) $\{x,z\}\subseteq Q_{\ell+1}\cap U$ and (3) $\{w,x\}\subseteq Q_{\ell+1}\cap U$. In case (1), $w_\alpha, w_\gamma, y_\alpha, y_\gamma$ are all fixed. But then two coordinates of $x=y_\alpha\oplus x_\beta\oplus w_\gamma$ and $z=w_\alpha\oplus z_\beta\oplus y_\gamma$ are fixed. Since Y is uniquely determined by projection, this means x and z are fixed, and so $Q_{\ell+1}$ is uniquely determined. In case (2), the choice of z fixes w_α and y_γ, and since w_β and y_β are fixed, this determines w and y. In case (3), $x_\alpha=y_\alpha$ is fixed. Since y_β is also fixed, two coordinates of y are fixed, which means $y=y_\alpha\oplus y_\beta\oplus y_\gamma$ is fixed. Now y_γ is fixed, and since w_α is fixed, this fixes $z=w_\alpha\oplus z_\beta\oplus y_\gamma$. So $Q_{\ell+1}$ is uniquely determined once we fix two vectors in $Q_{\ell+1}\cap U$.

Finally, there are at most $\binom{|U|}{2}\le\binom{4\ell}{2}<8\ell^2$ candidates for two vectors in $Q_{\ell+1}\cap U$, so there are less than $8\ell^2$ choices for $Q_{\ell+1}$ such that $(Q_1,Q_2,\dots,Q_\ell,Q_{\ell+1})$ is degenerate. ∎

3.9. Counting nondegenerate chains

Lemma 3.5. *Let $\mathcal{P}_k$ denote the set of nondegenerate chains of length k in $\mathcal{Q}$. Then*

$$|\mathcal{P}_k| > 4^{-k}m\big(d-32k^2\big)^k, \tag{9}$$

where $d>64k^2$ is the average degree in G and m is the number of vertices in G.

Proof. We claim that for all $\ell\le k$ and $d>64\ell^2$,

$$|\mathcal{P}_\ell| > 4^{-\ell}m\big(d-32\ell^2\big)^\ell. \tag{10}$$

The proof is by induction on $m+\ell$. If G contains a vertex of degree less than $d/4$, then we remove such a vertex to obtain a graph of average degree greater than $\widetilde{d}=d+d/(2m-2)$. By induction, the number of walks in $\mathcal{P}_\ell$ for this new graph is at least

$$4^{-\ell}(m-1)\big(\widetilde{d}-32\ell^2\big)^\ell > 4^{-\ell}m\big(d-32\ell^2\big)^\ell.$$

In particular, the number of non-returning walks of G which are in $\mathcal{P}_\ell$ is at least $4^{-\ell}m(d-32\ell^4)^\ell$, as required. Suppose every vertex of G has degree at least $d/4$, and

$$|\mathcal{P}_{\ell-1}| > 4^{-\ell+1}m\big(d-32(\ell-1)^2\big)^{\ell-1}.$$

Since every walk in $\mathcal{P}_{\ell-1}$ has at most $8(\ell-1)^2$ extensions to a degenerate walk of length ℓ, by Lemma 3.4, there are at least $d/4-8(\ell-1)^2-\ell>d/4-8\ell^2$ extensions of each walk in $\mathcal{P}_{\ell-1}$ to a walk in $\mathcal{P}_{\ell+1}$. This proves (10). Now the lemma follows from Lemma 3.4 with $\ell=k$. ∎

Lemma 3.6. *Let m be the number of vertices in G. Then*

$$|\mathcal{P}_k| < 4^k k^{4k} \binom{m}{2}. \tag{11}$$

Proof. By Lemma 3.3, if $Q = (Q_1, Q_2, \dots, Q_k)$ is a chain in $\mathcal{P}_k$, then the number of chains $R = (R_1, R_2, \dots, R_k)$ in $\mathcal{P}_k$ such that Q and R are concyclic is at most the number of choices of R such that $\widehat{R} = \widehat{Q}$. Let $U(Q) = \bigcup_{i=1}^k Q_i$ and $U(R) = \bigcup_{i=1}^k R_i$. The very important point now is that from the definition of non-degeneracy of Q and R, the projections of $U(Q)$ and of $U(R)$ onto $\mathbb{F}^\alpha$ and onto $\mathbb{F}^\gamma$ are invariant under reduction:

$$\begin{aligned} U(\widehat{Q})_\alpha &= U(Q)_\alpha \qquad & U(\widehat{Q})_\gamma &= U(Q)_\gamma \\ U(\widehat{R})_\alpha &= U(R)_\alpha \qquad & U(\widehat{R})_\gamma &= U(R)_\gamma. \end{aligned}$$

Therefore the number choices of $R \in \mathcal{P}_k$ concyclic with $Q \in \mathcal{P}_k$ is at most the number of choices of R such that $U(R)_\alpha = U(Q)_\alpha$ and $U(R)_\gamma = U(Q)_\gamma$. Since Y is determined by projection, any quadruple $R_i \in \mathcal{Q}$ is specified by its projection onto $\mathbb{F}^\alpha \oplus \mathbb{F}^\gamma$. The number of ways of choosing quadrilaterals $R_1, R_2, \dots, R_k$ so that $R = (R_1, R_2, \dots, R_k)$ is at most

$$\binom{|U(Q)_\alpha|}{2}^k \binom{|U(Q)_\gamma|}{2}^k \leq \binom{2k}{2}^k \binom{2k}{2}^k < 4^k k^{4k}$$

since $|U(Q)_\alpha| \leq 2k$ and $|U(Q)_\beta| \leq 2k$. This gives the required upper bound on $|\mathcal{P}_k|$. ∎

3.10. Linear Dependencies: Final Step

In this section we prove Theorem 2.3, using the lemmas we have developed in the last few sections. The following theorem combines all of these lemmas, and will also be used to prove Theorem 1.3.

Theorem 3.7. *Let $Z \subseteq \mathbb{F}^\alpha \oplus \mathbb{F}^\beta \oplus \mathbb{F}^\gamma$ be a balanced $8k$-wise independent set of vectors. Then*

$$|Z| < 4k(\lambda\mu\nu)^{\frac{1}{2}} \mu^{\frac{1}{2k}} + \lambda^{1+\frac{1}{2k}} + \mu^{1+\frac{1}{2k}} + \nu^{1+\frac{1}{2k}} + 2\lambda\nu^{\frac{1}{2}}. \tag{12}$$

Proof. Since Z is $8k$-wise independent, we may apply Lemma 3.1: there exists a set $Y \subseteq Z$ such that Y is determined by projection and

$$|Y| > \left(16k^2 \mu^{1+\frac{1}{k}} \lambda\nu\right)^{\frac{1}{2}} + \nu + 2\lambda\nu^{\frac{1}{2}}. \tag{13}$$

Let $\mathcal{P}_k$ denote the set of nondegenerate chains of length k in $\mathcal{Q}$, let d and m be the average degree and number of vertices in G, respectively. Combining (11) and (9) gives $d-32k^2<4k^4m^{\frac{1}{k}}$ and therefore $d<64k^4m^{\frac{1}{k}}$. It follows that since $|\mathcal{Q}|=|E(G)|=\frac{1}{2}dm$,

$$|\mathcal{Q}|<64k^4m^{1+\frac{1}{k}}. \tag{14}$$

For a contradiction, suppose that $|Z|$ is at least the expression claimed in (12). Now from (5) and (13),

$$|\mathcal{Q}|>\frac{|Y|^4}{4\lambda^2\nu^2}. \tag{15}$$

The vertices of G are ordered pairs of vectors of weight one in $\mathbb{F}^\beta$, so $m<\mu^2$. By (15) and (14),

$$|Y|^4<4\lambda^2\nu^2\cdot 64k^4m^{1+\frac{1}{k}}<256k^4\lambda^2\nu^2\mu^{2+\frac{2}{k}}=\left(16k^2\mu^{1+\frac{1}{k}}\lambda\nu\right)^2.$$

This contradicts (13), and proves the theorem. ∎

Remark. Theorem 3.7 can be used to derive a more precise version of Theorem 1.2: if $\mathcal{S}$ is an r-partite hypergraph with parts of sizes $N_1,N_2,\ldots,N_r$, and $\mathcal{S}$ contains no even cover of size at most $8k$, then the above theorem can be used to prove

$$|\mathcal{S}|\ll(N_1N_2\ldots N_r)^{\frac{1}{2}+\frac{1}{2k}}+(N_1+N_2+\cdots+N_r)^{\lceil\frac{r}{3}\rceil(1+\frac{1}{2k})}.$$

This result may be viewed as an extension of Theorem 2.1 from bipartite graphs to r-partite hypergraphs.

Proof of Theorem 1.1. Let $X\subseteq\mathbb{F}^n$ be a set of vectors of weight at most r. Let $\chi:\{1,2,\ldots,n\}\to\{1,2,3\}$ be a random three-coloring of the coordinates, where distinct coordinates are colored independently and each color is equiprobable. For a vector $x\in X$ of weight $\omega(x)=\omega$, the probability that x has exactly $\lfloor\frac{\omega}{3}\rfloor$ non-zero coordinates of color 1, exactly $\lceil\frac{\omega}{3}\rceil$ non-zero coordinates of color 2, and exactly $\lceil\frac{\omega}{3}\rceil$ non-zero coordinates of color three is exactly

$$\frac{1}{3^\omega}\binom{\omega}{\lfloor\frac{\omega}{3}\rfloor}\binom{\omega-\lfloor\frac{\omega}{3}\rfloor}{\lceil\frac{\omega}{3}\rceil}>\frac{1}{3\omega},$$

where we used the numerical Lemma 3.8 (see below) to obtain the inequality. In this case we say that x is equipartitioned by χ. Therefore the expected number of vectors $x\in X$ which are equipartitioned is greater than

$$\sum_{x\in X}\frac{1}{3\omega(x)}>\frac{|X|}{3r}.$$

This implies that there is a subset Z of X of size greater than $\frac{|X|}{3r}$ and a three-coloring χ such that every vector $z\in Z$ is equipartitioned by χ. Then Z may be regarded as a balanced subset of $\mathbb{F}^L\oplus\mathbb{F}^M\oplus\mathbb{F}^N$ where L, M and N are defined in Theorem 2.3. Applying Theorem 3.7 to Z with $\lambda=L$, $\mu=M$ and $\nu=N$, we obtain

$$\frac{|X|}{3r} < 4k(LN)^{\frac{1}{2}}M^{\frac{1}{2}+\frac{1}{2k}} + L^{1+\frac{1}{2k}} + M^{1+\frac{1}{2k}} + N^{1+\frac{1}{2k}} + 2LN^{\frac{1}{2}},$$

which implies that $|X|<12kr\cdot\left[(LN)^{\frac{1}{2}}M^{\frac{1}{2}+\frac{1}{2k}}+N^{1+\frac{1}{2k}}\right]$. This proves Theorem 2.3. ∎

Lemma 3.8. *Let ω be a positive integer. Then*

$$\binom{\omega}{\lfloor\frac{\omega}{3}\rfloor}\binom{\omega-\lfloor\frac{\omega}{3}\rfloor}{\lceil\frac{\omega}{3}\rceil} > \frac{3^{\omega-1}}{\omega}.$$

Proof. Let $f(\omega)$ denote the expression on the left in the inequality above. It is not hard to verify that the result is true for $\omega\in\{1,2,3\}$. Suppose $\omega>3$. Using the inequalities,

$$n^n e^{-n}(2\pi n)^{\frac{1}{2}} < n! < n^n e^{-n}(2\pi n)^{\frac{1}{2}} e^{\frac{1}{12n}},$$

which are valid for all positive integers n, we get that for all integers $s\geq 1$,

$$f(3s)=\binom{3s}{s}\binom{2s}{s}=\frac{(3s)!}{(s!)^3} > \frac{(3s)^{3s}e^{-3s}(6\pi s)^{1/2}}{s^{3s}e^{-3s}(2\pi s)^{3/2}e^{1/4s}} > \frac{3^{3s+\frac{1}{2}}}{2\pi s e^{1/4s}} > \frac{3^{3s}}{4s}.$$

This implies the required inequality when ω is a multiple of 3. We pass to general ω by noting that

$$f(3s+1)=\frac{3s+1}{s+1}f(3s) \quad \text{and} \quad f(3s+2)=\frac{(3s+2)(3s+1)}{(s+1)^2}f(3s).$$

In particular, for $s\geq 1$, $f(3s+1)\geq 2f(3s)$ and $f(3s+2)\geq 5f(3s)$, which implies the required inequality. ∎

4. Product representations of squares

In this section we prove Theorem 1.3. Before doing so, we require the following simple lemma:

Lemma 4.1. *Let $n>1$ be a positive integer. Then either n has a prime factor larger than $N^{\frac{1}{2}}$, or $n=xyz$ where $x,y,z\leq N^{\frac{1}{2}}$.*

Proof. Let $n = p_1 p_2 \cdots p_r$ denote the prime factorization of n into (not necessarily distinct) primes p_i where $p_1 \geq p_2 \geq \cdots \geq p_r$. Suppose $p_1 \leq N^{\frac{1}{2}}$. Then we can find a set X of p_is whose product x is at most $N^{\frac{1}{2}}$ but as close to $N^{\frac{1}{2}}$ as possible. Let y be a prime factor of n which isn't in X. Then $xy \geq N^{\frac{1}{2}}$, and we may take $z = n/(xy)$. ∎

In what follows we denote by $\Pi(n)$ the set of all primes in $\{1,\ldots,n\}$, and let $\pi(n) = |\Pi(n)|$ be the usual prime counting function.

Proof of Theorem 1.3. Let $A \subseteq \{1,2,\ldots,n\}$ be a set such that no product of at most $8k$ distinct elements of A is a square. Denote by $B \subseteq A$ the set of integers in A which have a prime factor larger than $N^{\frac{1}{2}}$, and write $C = A \backslash B$. By Lemma 4.1 we have that $C = \{a \in A\colon\ a = xyz,\ x,y,z \leq N^{\frac{1}{2}}\}$. Denote for $0 \leq i \leq \frac{1}{2}\log_2 n$,

$$P_i = \left\{p \in \Pi(n) :\ \frac{n}{2^{i+1}} < p \leq \frac{n}{2^i}\right\}.$$

Form a bipartite graph G_i with parts P_i and $\{1,\ldots,2^{i+1}\}$ such that $p \in P_i$ is joined to $q \in \{1,\ldots,2^{i+1}\}$ if $pq \in B$. Then G_i does not contain a cycle of length at most $8k$ since if for some $2 \leq \ell \leq 8k$, $p_1q_1, q_1p_2, p_2q_2, \ldots, q_{\ell-1}p_1$ is such a cycle then $p_jq_j, p_{j+1}q_j \in A$ are distinct and their product is a square. It was proved in [23] that an M by N bipartite graph of girth at least $2k+2$ has at most $(MN)^{\frac{1}{2}+\frac{1}{2k}} + M + N$ edges. Since G_i has girth at least $8k+2$, we deduce that

$$\begin{aligned}|E(G_i)| &\leq \left(2^{i+1}|P_i|\right)^{\frac{1}{2}+\frac{1}{8k}} + 2^{i+1} + |P_i| \\ &\leq \left(2^{i+1}\left[\pi\left(\frac{n}{2^i}\right) - \pi\left(\frac{n}{2^{i+1}}\right)\right]\right)^{\frac{1}{2}+\frac{1}{8k}} + 2^{i+1} + |P_i|.\end{aligned}$$

Adding these inequalities for $i = 0,\ldots,\left\lfloor\frac{1}{2}\log_2 n\right\rfloor$ gives

$$|B| = \sum_{i=0}^{\lfloor\frac{1}{2}\log_2 n\rfloor} |E(G_i)| \leq \pi(n) + O\big(n^{\frac{1}{2}+\frac{1}{8k}}\big).$$

We now estimate $|C|$. For each $t \in C$ fix a factorization $t = x_t y_t z_t$ with $x_t y_t, z_t \leq N^{\frac{1}{2}}$. We assume $x_t \geq y_t \geq z_t$, so in particular $z_t \leq n^{\frac{1}{3}}$. Denote

$$S = \Big\{(i,j) :\ 0 \leq i \leq j :\ i + j + 2 \leq \frac{1}{3}\log_2 n\Big\}.$$

For $(i,j) \in S$ let C_{ij} denote the set of all $t \in C$ such that

$$\frac{N^{\frac{1}{2}}}{2^{i+1}} < x_t \leq \frac{N^{\frac{1}{2}}}{2^i} \qquad \frac{N^{\frac{1}{2}}}{2^{j+1}} < y_t \leq \frac{N^{\frac{1}{2}}}{2^j} \qquad 1 \leq z_t \leq 2^{i+j+2}.$$

We now apply Theorem 3.7 with $\mathbb{F}=\mathbb{F}_2$ and $\lambda=2^{i+j+2}$, $\mu=N^{\frac{1}{2}}/2^{i+1}$ and $\nu=N^{\frac{1}{2}}/2^{j+1}$. Each $t\in C_{ij}$ may be considered as a vector of weight three in $\mathbb{F}^\lambda\oplus\mathbb{F}^\mu\oplus\mathbb{F}^\nu$: if $t=x_ty_tz_t$ is the prescribed factorization of t then the vector associated with t is the vector of weight three with a one in positions x_t, $\mu+y_t$ and $\mu+\nu+z_t$, and zeros elsewhere. Clearly no $8k$ of these vectors are linearly dependent, otherwise the product of the corresponding elements of C_{ij} is a square. By Theorem 3.7, we have that

$$\begin{aligned}|C_{ij}| &< 12kr(\lambda\mu\nu)^{\frac{1}{2}}\mu^{\frac{1}{2k}}+\lambda^{1+\frac{1}{2k}}+\mu^{1+\frac{1}{2k}}+\nu^{1+\frac{1}{2k}}+2\lambda\nu^{\frac{1}{2}}\\ &\ll k\left(\frac{N^{\frac{1}{2}}}{2^{i+1}}\cdot\frac{N^{\frac{1}{2}}}{2^{j+1}}\cdot 2^{i+j+2}\right)^{\frac{1}{2}}\left(\frac{N^{\frac{1}{2}}}{2^{i+1}}\right)^{\frac{1}{2k}}+\left(\frac{N^{\frac{1}{2}}}{2^{j+1}}\right)^{\frac{1}{2}+\frac{1}{2k}}+2^{i+j}\left(\frac{N^{\frac{1}{2}}}{2^{j+1}}\right)^{\frac{1}{2}}\\ &\ll kn^{\frac{1}{2}+\frac{1}{2k}}+2^{i+\frac{j}{2}}n^{\frac{1}{4}}.\end{aligned}$$

Finally, we sum this inequality over all $(i,j)\in S$. We chose λ,μ,ν carefully to ensure that the sum of the last term over $(i,j)\in S$ is of order at most $n^{\frac{1}{2}}$. Therefore we have

$$|C|\ll kn^{\frac{1}{2}+\frac{1}{2k}}\cdot|S|+n^{\frac{1}{2}}\ll kn^{\frac{1}{2}+\frac{1}{2k}}(\log n)^2.$$

This concludes the proof of Theorem 1.3. ∎

Acknowledgements

We are grateful to Noga Alon and Henry Cohn for helpful comments.

References

[1] N. Alon, L. Babai and A. Itai: A fast and simple randomized parallel algorithm for the maximal independent set problem, *J. Algorithms* **7(4)** (1986), 567–583.

[2] N. Alon, O. Goldreich, J. Håstad and R. Peralta: Simple constructions of almost k-wise independent random variables, *Random Structures Algorithms* **3(3)** (1992), 289–304.

[3] N. Alon, S. Hoory and N. Linial: The Moore bound for irregular graphs, *Graphs Combin.* **18(1)** (2002), 53–57.

[4] N. Alon and J. H. Spencer: *The probabilistic method*, Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2000. With an appendix on the life and work of Paul Erdős.

[5] N. Alon, R. Yuster and U. Zwick: Color-coding, *J. Assoc. Comput. Mach.* **42(4)** (1995), 844–856.

[6] N. Alon, R. Yuster and U. Zwick: Finding and counting given length cycles, *Algorithmica* **17(3)** (1997), 209–223.

[7] L. Bazzi, M. Mahdian and D. A. Spielman: The minimum distance of Turbo-like codes, preprint, 2003.

[8] C. Berrou, A. Glavieux and P. Thitimajshima: Near Shannon Limit Error Correcting Codes and Decoding: Turbo Codes; in *Proceedings of IEEE International Communications Conference*, pages 1064–1070, 1993.

[9] C. Bertram-Kretzberg, T. Hofmeister and H. Lefmann: Sparse 0-1 matrices and forbidden hypergraphs, *Combin. Probab. Comput.* **8(5)** (1999), 417–427.

[10] C. Bertram-Kretzberg and H. Lefmann: MOD_p-tests, almost independence and small probability spaces; *Random Structures Algorithms* **16(4)** (2000), 293–313.

[11] A. Beutelspacher and U. Rosenbaum: *Projective geometry: from foundations to applications*; Cambridge University Press, Cambridge, 1998.

[12] B. Bollobás: *Extremal graph theory*, Dover Publications Inc., Mineola, NY, 2004. Reprint of the 1978 original.

[13] J. Bondy and M. Simonovits: Cycles of even length in graphs, *J. Combinatorial Theory Ser. B* **16** (1974), 97–105.

[14] M. Breiling: A logarithmic upper bound on the minimum distance of Turbo codes, *IEEE Transactions on Information Theory* **50(8)** (2004), 1692–1710.

[15] A. E. Brouwer: Block designs, in: *Handbook of combinatorics*, Vol. **1**, **2**, pages 693–745, Elsevier, Amsterdam, 1995.

[16] W. G. Brown, P. Erdős and V. T. Sós: Some extremal problems on r-graphs, in: *New directions in the theory of graphs, Proc 3rd Ann. Arbor Conference on Graph Theory*, Academic Press, New York, 55–63, 1973.

[17] M. C. Davey and D. J. C. MacKay: Low-density parity check codes over $GF(q)$, *IEEE Communications Letters* **2(6)** (1998), 165–167.

[18] J. D. Dixon: Asymptotically fast factorization of integers, *Math. Comp.* **36(153)** (1981), 255–260.

[19] P. Erdős: On some applications of graph theory to number theoretic problems, *Publ. Ramanujan Inst.* **1** (1969), 131–136.

[20] P. Erdős, A. Sárközy and V. T. Sós: On product representations of powers, I; *European J. Combin.* **16(6)** (1995), 567–588.

[21] R. G. Gallager: *Low Density Parity Check Codes*, MIT Press, Cambridge MA, 1963. Research Monograph Series, no. **21**.

[22] E. Győri: C_6-free bipartite graphs and product representation of squares, *Discrete Math.* **165/166** (1997), 371–375. Graphs and combinatorics (Marseille, 1995).

[23] S. Hoory: The size of bipartite graphs with a given girth, *J. Combin. Theory Ser. B* **86(2)** (2002), 215–220.

[24] N. Kahale and R. Urbanke: On the minimum distance of parallel and serially concatenated codes, *IEEE Trans. Inform. Theory*, to appear.

[25] H. Lefmann: Sparse parity-check matrices over finite fields (extended abstract), in: *Computing and combinatorics*, volume **2697** of Lecture Notes in Comput. Sci., pages 112–121, Springer, Berlin, 2003.

[26] H. Lefmann, P. Pudlák and P. Savický: On sparse parity check matrices, *Des. Codes Cryptogr.* **12(2)** (1997), 107–130.

[27] D. J. C. MacKay: Good error-correcting codes based on very sparse matrices, *IEEE Trans. Inform. Theory* **45(2)** (1999), 399–431.

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

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

[30] A. Naor and J. Verstraëte: A note on bipartite graphs without $2k$-cycles, *Combin. Probab. Comput.* **14(5–6)** (2005), 845-849.

[31] C. Pomerance: A tale of two sieves, *Notices Amer. Math. Soc.* **43(12)** (1996), 1473–1485.

[32] C. Pomerance and A. Sárközy: Combinatorial number theory, in: *Handbook of combinatorics*, Vol. **1**, **2**, pages 967–1018, Elsevier, Amsterdam, 1995.

[33] G. N. Sárközy: Cycles in bipartite graphs and an application in number theory, *J. Graph Theory* **19(3)** (1995), 323–331.

[34] M. Sipser and D. A. Spielman: Expander codes, *IEEE Trans. Inform. Theory* **42(6/1)** (1996), 1710–1722. Codes and complexity.

[35] D. A. Spielman: Linear-time encodable and decodable error-correcting codes, *IEEE Trans. Inform. Theory* **42(6/1)** (1996), 1723–1731. Codes and complexity.

[36] J. Verstraëte: On arithmetic progressions of cycle lengths in graphs, *Combin. Probab. Comput.* **9(4)** (2000), 369–373.

[37] H. C. Williams and J. O. Shallit: Factoring integers before computers, in: *Mathematics of Computation 1943–1993: a half-century of computational mathematics (Vancouver, BC, 1993)*, volume **48** of Proc. Sympos. Appl. Math., pages 481–531, Amer. Math. Soc., Providence, RI, 1994.

Assaf Naor

Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012
USA
naor@cims.nyu.edu

Jacques Verstraëte

Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0112
USA
jverstraete@ucsd.edu