Electronic Colloquium on Computational Complexity, Report No. 103 (2007)

Selected Results in Additive Combinatorics: An Exposition

Emanuele Viola*

September 27, 2007

Abstract

We give a self-contained exposition of selected results in additive combinatorics over the group $GF(2)^n = \{0,1\}^n$. In particular, we prove the celebrated theorems known as the Balog-Szemeredi-Gowers theorem ('94 and '98) and the Freiman-Ruzsa theorem ('73 and '99), leading to the remarkable result by Samorodnitsky ('07) that linear transformations are efficiently testable.

No new result is proved here. However, we strip down the available proofs to the bare minimum needed to derive the efficient testability of linear transformations over $\{0,1\}^n$, thus hoping to provide a computer science-friendly introduction to the marvelous field of additive combinatorics.

1 Introduction

Additive combinatorics is a fascinating area of mathematics that has recently found several applications in computer science, for example in the areas of extractors [BIW], property testing [Sam], PCP's [ST], hardness amplification [Vio, VW], and pseudorandomness [BV]. A growing number of beautiful expositions of additive combinatorics is available, including the book by Tao and Vu [TV] and the notes by Green [Gre]. However, these expositions are somewhat targeted to mathematicians, and their breadth may be disorienting to the uninitiated. On the other hand, in these notes we aim to provide a computer science-friendly introduction to additive combinatorics. We hope to achieve this by giving a self-contained exposition of selected results in additive combinatorics, stripped-down to the bare minimum needed to obtain the following remarkable result by Samorodnitsky [Sam] that linear transformations are efficiently testable.

*viola@cs.columbia.edu. Work done while the author was at the School of Mathematics, Institute for Advanced Study, Princeton, NJ, 08540, supported by NSF grant CCR-0324906.

ISSN 1433-8092

Theorem 1.1 ([Sam]; Testing linear transformations)**.** *Let $f : \{0,1\}^n \to \{0,1\}^n$ be any function, and let '+' denote bit-wise XOR. If $\Pr_{x,x' \in \{0,1\}^n}[f(x)+f(x') = f(x+x')] \geq \epsilon$, then there is an $n \times n$ matrix M such that $\Pr_{x \in \{0,1\}^n}[f(x) = Mx] \geq \epsilon'$, where ϵ' depends on ϵ only.*

The same machinery that goes in the proof of Theorem 1.1 constitutes the core of the proof of the "Inverse theorem for the Gowers U_3 norm" [GT1, Sam] which in turn can be used to obtain correlation bounds for quadratic polynomials [Vio, VW] and pseudorandom generators for cubic polynomials [BV].

To introduce the subject and motivate the following sections, let us now informally see how the property-testing result in Theorem 1.1 follows from some results in additive combinatorics, which will be presented along the way.

Proof idea for Theorem 1.1 We are interested in the additive combinatorics of the *graph* A of the function f:

$$A := \{(x, f(x)) : x \in \{0,1\}^n\} \subseteq \{0,1\}^{2n}.$$

The approach to prove the theorem is to show that A is *approximately* a linear space. This approach is motivated by the observation that if A' was *exactly* a linear space, then f must be a linear transformation, because in this case $(x, f(x))+(x', f(x')) = (x+x', f(x)+f(x')) \in A' \subseteq A = \{(x, f(x)) : x \in \{0,1\}^n\}$, and so $f(x) + f(x')$ must equal $f(x + x')$.

We start by noting that our assumption can be written as

$$\Pr_{a,a' \in A}[a + a' \in A] \geq \epsilon. \tag{1}$$

Here we apply our first result in additive combinatorics, namely the *Balog-Szemeredi-Gowers (BSG) theorem* [BS, Gow]. This theorem states that if a set A satisfies (1) then it contains a large subset that is nearly closed under addition. More formally, defining $2S := \{a + a' : a, a' \in S\}$, the BSG theorem says that there is a set $A' \subseteq A$ of large size $|A'| \approx |A|$ such that

$$|2A'| \approx |A'|. \tag{2}$$

(From Equation (2) we cannot in general conclude that $|2A| \approx |A|$, which motivates considering the subset $A' \subseteq A$.)

At this point we apply our second result in additive combinatorics, namely *Ruzsa's theorem* [Ruz], which is a finite-field analogue of an older theorem by Freiman [Fre]. This theorem says that if a set A' satisfies (2) then it is approximately a linear space. Specifically, denoting by $span(A')$ the vector space spanned by elements of A', Ruzsa's theorem states that

$$|span(A')| \approx |A'|. \tag{3}$$

In other words, Ruzsa's theorem says that if linear combinations of length 2 (i.e. $2A'$) do not buy much size, then neither do linear combinations of arbitrary length (i.e. $span(A')$).

Finally, even though A' may not be a linear space, from (3) one can still draw the conclusion that f is close to a linear transformation, thus concluding the proof of the theorem. □

Before discussing how this exposition is organized, we stress that it focuses on additive combinatorics over the group $GF(2)^n = \{0,1\}^n$. This choice is motivated by the importance of this group to computer science, and the fact that the proofs of the relevant results in additive combinatorics appear to be cleanest over $\{0,1\}^n$: More work is needed to switch between '+' and '−' (see, e.g., [TV]). We also would like to mention that, recently, Green and Tao have given a new direct proof of the combination of the BSG and Ruzsa theorems over $\{0,1\}^n$ [GT3], using Fourier analysis. Going back to the proof of Theorem 1.1, their result goes directly from (1) to (3).

Organization. After some preliminaries in Section 2, we prove the BSG theorem in Section 3. In Section 4 we prove Ruzsa's theorem. In Section 5 we conclude the proof of the testability of linear transformations (Theorem 1.1). Our presentation of the BSG theorem in Section 3 follows one by Sudakov, Szemeredi, and Vu [SSV], which relies on a graph-theoretic lemma regarding certain paths in dense graphs. In Section 6 we also present the proof of the optimality of the path length of this lemma, due to Kostochka and Sudakov [KS].

2 Preliminaries

In this work we are concerned with subsets of the group $GF(2)^n = \{0,1\}^n$, whose operation is the component-wise addition (a.k.a. bit-wise XOR) denoted by '$+$'. Throughout these notes, A denotes a subset of $\{0,1\}^n$. For an integer l and a set $A \subseteq \{0,1\}^n$ we denote by lA the set of all sums of length l with elements in A, i.e. $lA := \{\sum_{i \le l} a_i : a_i \in A\}$. We also write $A + A$ for $2A$, $A + A + A$ for $3A$, and so on. We generalize this notation to allow summands from different sets, as in $A + B$; finally, we denote by $span(A)$ the span of the elements of A, i.e. $span(A) = \bigcup_l lA$.

We will use several times the following basic counting argument, whose proof is straightforward.

Proposition 2.1 (Double counting)**.** *Let $f : D \to S$ be a function. Suppose that for every $s \in S$ there are t distinct $d_i \in D$ such that $f(d_i) = s$. Then $|S| \le |D|/t$.*

3 The Balog-Szemeredi-Gowers (BSG) theorem

In this section we prove the Balog-Szemeredi-Gowers (BSG) theorem, which is stated next.

Theorem 3.1 ([BS, Gow])**.** *Suppose that $\Pr_{a,a' \in A}[a + a' \in A] \ge \epsilon$, where $A \subseteq \{0,1\}^n$. Then there is $A' \subseteq A$, $|A'| \ge (\epsilon/3) \cdot |A|$, such that $|2A'| \le (6/\epsilon)^8 \cdot |A|$.*

One way to think of the BSG theorem is the following. For a subset E of the cartesian product $A \times A$, let us denote its set of sums by $\sum E := \{a + b : (a,b) \in E\}$. Then the BSG theorem says that from a dense $E \subseteq A \times A$ such that $|\sum E|$ is small, we can obtain a dense $A' \subseteq A$ such that $|\sum(A' \times A')| = |2A'|$ is small.

The proof that we present of the above Theorem 3.1 follows one by Sudakov, Szemeredi, and Vu [SSV]. It relies on the following graph-theoretical statement, which does not use any property of addition and only relies on the density of the graph.

Lemma 3.2 ([SSV])**.** *Let $G = (A, E)$ be an undirected graph with $|A| = N$ nodes, $|E| = \epsilon \cdot N^2$ edges, and no self-loops. Then there is a set $A' \subseteq A$, $|A'| \geq \epsilon \cdot N$ such that for every $a, b \in A'$ there are at least $(\epsilon/2)^8 \cdot N^3$ paths of length 4 in G from a to b, i.e. there are at least $(\epsilon/2)^8 \cdot N^3$ choices of triples (c_1, c_2, c_3) such that $\{(a, c_1), (c_1, c_2), (c_2, c_3), (c_3, b)\} \subseteq E$.*

Proof of Theorem 3.1 assuming Lemma 3.2 Define E to be the set of edges $\{a, b \neq a\}$ such that $a + b \in A$. By assumption, $|E| \geq (\epsilon/3) \cdot |A|^2$ (the factor $1/3$ grossly accounts for the translation between the hypothesis, which talks about pairs, and Lemma 3.2, which talks about edges). Now let A' be the subset of A given by Lemma 3.2. Consider any two $a, b \in A'$. By Lemma 3.2 there are $\epsilon' \cdot |A|^3$ paths (a, c_1, c_2, c_3, b) with edges in E, where $\epsilon' = (\epsilon/6)^8$. By definition of E, the sum of two consecutive nodes in any path lies in A. Thus, considering the function $f(x, x', x'', x''') := x + x' + x'' + x'''$, we have that, for every $a + b \in 2A'$,

$$f(x := a + c_1, x' := c_1 + c_2, x'' := c_2 + c_3, x''' := c_3 + b) \\ = (a + c_1) + (c_1 + c_2) + (c_2 + c_3) + (c_3 + b) = a + b,$$

for at least $\epsilon' \cdot |A|^3$ inputs $(x, x', x'', x''') \in A^4$. Note that we are using that distinct triples (c_1, c_2, c_3) give rise to distinct inputs (x, x', x'', x'''), which is immediate to see after we recall that a and b are fixed. By double counting (Proposition 2.1) we obtain

$$|2A'| \leq |A|^4/(\epsilon' \cdot |A|^3),$$

concluding the proof. □

Proof of Lemma 3.2 The idea is to exhibit a set $A' \subseteq A$ such that every $a \in A'$ shares many ($\Omega(N)$) neighbors with most ($(1 - 0.1)$ fraction) nodes in A'. From this we have that, for every two nodes $a, b \in A'$, most ($(1 - 0.2)$ fraction) nodes c_2 in A' share many ($\Omega(N)$) neighbors c_1 with a and also share many neighbors c_3 with b, which gives the result.

For a node $v \in G$ let us denote by $N(v) \subseteq A$ the neighborhood of v. A' will be a subset of $N(v')$ for some v' given by a probabilistic argument. For this argument, let $V \in G$ be a random node in G, and call a pair $\{u, w \neq u\}$ *bad* if $|N(u) \cap N(w)| \leq \epsilon^3 \cdot N$.

We are interested in the number of bad pairs inside $N(V)$. Let $B_{\{u,w\}}$ be the 0/1 indicator variable which is 1 when $\{u, w\}$ is a bad pair in $N(V)$, i.e. such that $\{u, w\} \subseteq N(V)$. For every bad pair $\{u, w\}$ (not necessarily such that $\{u, w\} \subseteq N(V)$) we see that $\{u, w\} \subseteq N(V)$ when V is a common neighbor of u and w, which by the definition of bad pair happens with probability at most ϵ^3. Consequently, by linearity of expectation, we have

$$E_{V \in A}[\text{number of bad pairs in } N(V)] \leq \epsilon^3 \cdot \binom{N}{2} \leq \epsilon^3 \cdot N^2/2. \tag{4}$$

Let us now denote by $S(V)$ the set of nodes $u \in N(V)$ that form a bad pair with at least $\epsilon^2 \cdot N$ other nodes $w \in N(V)$. Since there are always $|S(V)| \cdot \epsilon^2 \cdot N/2$ bad pairs in $N(V)$, where the factor $1/2$ comes from the fact that each bad pair $\{u, w\}$ is counted once for u and another time for w, Equation (4) implies that

$$E_{V \in A}[|S(V)|] \leq (\epsilon^3 \cdot N^2/2)/(\epsilon^2 \cdot N/2) = \epsilon \cdot N. \tag{5}$$

Therefore, using the fact that $E[|N(V)|] = 2 \cdot \epsilon \cdot N$ because the graph has $\epsilon \cdot N^2$ edges and no self-loops, we have

$$E_{V \in A}[|N(V) - S(V)|] = E[|N(V)|] - E[|S(V)|] \geq 2 \cdot \epsilon \cdot N - \epsilon \cdot N = \epsilon \cdot N.$$

We now fix a $v' = V$ that maximizes the above expectation and let $A' := N(\bar{v}) - S(\bar{v})$ be the corresponding set with $|A'| \geq \epsilon \cdot N$.

To see that A' satisfies the conclusion of the lemma, consider any $a, b \in A'$. Since we removed the nodes in $S(\bar{v})$, i.e. those that form a bad pair with at least $\epsilon^2 \cdot N$ other nodes $w \in N(\bar{v})$, both a and b form a good pair with all but $\epsilon^2 \cdot N$ nodes of A'. So there are at least $|A'| - 2 \cdot \epsilon^2 \cdot N \geq \epsilon \cdot N - 2 \cdot \epsilon^2 \cdot N \geq \epsilon^2 \cdot N$ nodes $c_2 \in A'$ that form a good pair with both a and b, where the last inequality holds if we assume that $\epsilon \leq 1/3$. For every such c_2 we have, by definition of good pair, $\epsilon^3 \cdot N$ choices for c_1 and as many for c_3 such that (a, c_1, c_2, c_3, b) is a path in G. In total, we have at least $(\epsilon^3 \cdot N)(\epsilon^2 \cdot N)(\epsilon^3 \cdot N) = \epsilon^8 \cdot N$ such paths in G. This proves the theorem except for the assumption that $\epsilon \leq 1/3$. If $\epsilon > 1/3$ the same proofs works replacing ϵ with $\epsilon/2$, which is at most $1/3$ because any undirected graph trivially has at most $N^2/2 \geq \epsilon \cdot N^2$ edges. □

4 Ruzsa's theorem

In this section we prove Ruzsa's theorem, which is stated next.

Theorem 4.1 ([Ruz]; The span of A does not expand if $2A$ does not)**.** *Suppose that* $|2A| \leq c \cdot |A|$*, where* $A \subseteq \{0,1\}^n$*. Then* $|span(A)| \leq c' \cdot |A|$*, where* c' *depends on* c *only.*

The core of the proof of Theorem 4.1 is the following lemma.

Lemma 4.2 ($4A$ does not expand if $2A$ does not)**.** $|4A| \leq 16 \cdot (|2A|/|A|)^4 \cdot |2A|$.

Proof of Theorem 4.1 assuming Lemma 4.2 We start with the following *covering claim*: There is a set $X \subseteq 3A$ whose size depends only on c such that for every $b \in 3A$ we have

$$|(X + A) \cap (b + A)| \geq 1. \qquad (\star)$$

To prove the covering claim, initialize X to the empty set, and as long as there is some $b \in 3A$ violating $(\star)$, add b to X. The resulting X satisfies the intersection requirement by construction. To verify the bound on the size of X, note that at each iteration the set $X + A$

grows in size by $|A|$, but $X + A \subseteq 4A$ always, and so at the end of the process $|X|$ is at most $|4A|/|A|$, a quantity which, by Lemma 4.2 and our assumption that $|2A| \leq c \cdot |A|$, depends only on c.

Now we show by induction that, for every $\ell \geq 3$, $\ell A \subseteq (\ell - 2)X + 2A$. This will conclude the proof as the size of X depends only on c. Specifically we obtain that $span(A) \subseteq span(X) + 2A$, and so $|span(A)| \leq |span(X)| \cdot |2A| \leq 2^{|X|} \cdot c \cdot |A|$.

For the base case $\ell = 3$ of the induction, take any $x \in 3A$. By $(\star)$, $X + A$ intersects $b + A$, and therefore $b \in X + 2A$.

For the inductive step, write $\ell A = (\ell - 1)A + A \subseteq (\ell - 3)X + 2A + A \subseteq (\ell - 2)X + 2A$, where we apply the inductive hypothesis and then the base case. $\square$

Proof of Lemma 4.2 We start with the following *covering claim*, whose statement and proof are very similar to those of the covering claim in the proof of Theorem 4.1 from the current lemma. There is a set $X \subseteq A$ of size $|X| \leq 2 \cdot |2A|/|A|$ such that for every $b \in A$ we have

$$|(X + A) \cap (b + A)| \geq |A|/2.$$

As a consequence of the covering claim, we have that for every $b \in A$ there are at least $|A|/2$ triples $(a_0 \in A, a_1 \in A, x \in X)$ such that $b = x + a_0 + a_1$; This is because each element $y \in (X + A) \cap (b + A)$ gives rise to, say, one such triple with $a_1 := b + y$ (this last requirement makes all the triples distinct).

Now we use the above consequence to prove the lemma. Fix an arbitrary $z = b_0 + b_1 \in 2A$. By the above, there are at least $|A|/2$ triples $(a_0 \in A, a_1 \in A, x \in X)$ such that $z = b_0 + a_0 + a_1 + x$. Since $b_0 + a_0 \in 2A$, there are at least $|A|/2$ triples $(c \in 2A, a_1 \in A, x \in X)$ such that $z = c + a_1 + x$. By repeating the argument for another arbitrary $z' = b'_0 + b'_1$, we obtain that for any two arbitrary $z, z' \in 2A$ there are at least $(|A|/2)^2$ sixtuples $(c \in 2A, c' \in 2A, a_1 \in A, a'_1 \in A, x \in X, x' \in X)$ such that

$$\begin{cases} z &= c + a_1 + x, \\ z' &= c' + a'_1 + x'. \end{cases}$$

Note that, in any solution to the above system, a_1 and a'_1 are uniquely determined once c, x, c', x' are (recall that z and z' are fixed). Consequently, the map that takes a solution $(c \in 2A, c' \in 2A, a_1 \in A, a'_1 \in A, x \in X, x' \in X)$ to the quintuple $(c \in 2A, c' \in 2A, a_1 + a'_1 \in 2A, x \in X, x' \in X)$ is one-to-one (i.e., injective). Moreover, such a quintuple sums up to $z+z'$. Therefore, similarly to the proof of Theorem 3.1, we have a function $f(x, x', x'', x''', x'''') := x + x' + x'' + x''' + x''''$ such that for every element $z + z' \in 4A$ there are at least $(|A|/2)^2$ distinct inputs y such that $f(y) = z$. By Proposition 2.1 we have

$$|4A| \leq |2A|^3 \cdot |X|^2 \cdot 4/|A|^2 \leq 16 \cdot (|2A|/|A|)^4 \cdot |2A|,$$

where we are using the fact, established at the beginning of this proof, that $|X| \leq 2 \cdot |2A|/|A|$. $\square$

Remark 4.3 (On the loss in parameters)**.** *We remark that one can eliminate the factor* 16 *in Lemma 4.2 by applying the lemma to the set* $A \times A \times \ldots \times A$*, see e.g. [TV, Corollary 2.18]. Turning back to the main result of this section, i.e. Theorem 4.1, we note that the current best upper bound on* c' *is obtained by Green and Tao [GT2] who prove* $c' \leq 2^{2 \cdot c}$ *modulo lower order factors. It can be verified that* $c' \leq 2^{2 \cdot c}$ *is the best possible, and in particular that* c' *in general must be exponential in* c*. However, if one is willing to settle for the span of a large subset* A' *of* A*, rather than all of* A*, in the same spirit as the BSG theorem, then it is conjectured that* c' *can be made polynomial in* c*; cf. the supplement "The polynomial Freiman-Ruzsa conjecture" to [Gre].*

5 Obtaining a linear transformation

In this section we conclude the proof of the property testing result in Theorem 1.1. The last component of the proof is the following linear-algebraic fact that states that if the span of (a large subset of) $\{(x, f(x))\}$ does not grow much, then f is approximately a linear transformation.

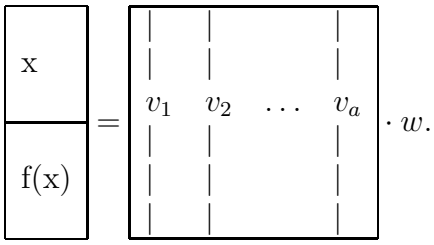
Lemma 5.1. *Let* $f : \{0,1\}^n \to \{0,1\}$ *be a function, and* $A \subseteq \{(x, f(x))\} \subseteq \{0,1\}^{2 \cdot n}$*. Suppose that*

$$\epsilon \cdot 2^n \leq |A| \leq |span(A)| \leq 2^n/\epsilon.$$

Then there is a linear transformation (i.e., an $n \times n$ *matrix* M*) such that*

$$\Pr_{x \in \{0,1\}^n}[f(x) = Mx] \geq (\epsilon/2)^2.$$

Proof. We start by finding an affine transformation $Tx+u$, then we observe how this implies a linear transformation. Let $v_1, v_2, \ldots, v_a$ be basis vectors for $span(A)$. By definition, every vector $(x, f(x)) \in A$ is a linear combination of the v_i's, i.e. for every $(x, f(x)) \in A$ there exists $w \in \{0,1\}^a$ such that



Let us now add to our collection new vectors $v_{a+1}, v_{a+2}, \ldots, v_k$ so that the projection onto the first n coordinates of $span(\{v_1, \ldots, v_k\})$ is all of $\{0,1\}^n$. Since each new vector multiplies the size of the span by 2, and the first a vectors by assumption span a set of size at least $\epsilon \cdot 2^n$, we have $k - a \leq \log(1/\epsilon)$. Also note that $a \leq n + \log(1/\epsilon)$ by assumption, hence $k \leq n + 2\log(1/\epsilon)$. Let V_k be the resulting matrix of the v_i's. By performing Gaussian elimination,

we can find an invertible transformation that turns V_k into the following canonical form:

I	0
T	U

,

where I is the identity $n \times n$ matrix, T is also $n \times n$, and T is $n \times (k-n)$. This is possible because the projection of the vectors $v_1, \ldots, v_k$ onto the first n coordinates spans $\{0,1\}^n$. Since this transformation is invertible, we still have the property that for every vector $(x, f(x)) \in A$ there exists $w \in \{0,1\}^k$ such that

x		I	0
	=			$\cdot\, w.$
f(x)		T	U

This means that the first n coordinates of w must equal x. Consequently, every value $f(x)$ equals $Tx + Uz$ for some $z \in \{0,1\}^{k-n}$. Therefore, by an averaging argument, there exists a fixed $u = Uz$ so that

$$\Pr_{x \in \{0,1\}^n}[f(x) = Tx + u] \geq 2^{-(k-n)} \geq \epsilon^2.$$

This gives us an affine transformation, and in what follows we show how in fact one can get a linear transformation, i.e. get rid of the 'u' above, with only a slight loss in probability. We claim that

$$\text{there is an } i \leq n \text{ such that } \Pr_{x \in \{0,1\}^n}[f(x) = Tx + u | x_i = 1] \geq \epsilon^2/2. \qquad (\star)$$

Such a claim lets us construct a linear transformation M by summing u to the i-th column of T, concluding the proof of the lemma (the extra factor of $1/2$ in the conclusion of the lemma accounts for the probability that $x_i = 1$).

It remains to prove $(\star)$. For this, let Y be the distribution on $\{0,1\}^n$ that is obtained by selecting a random $i \leq n$, setting to 1 the i-th bit, and choosing uniformly at random for the other bits. The statistical distance between Y and the uniform distribution on $\{0,1\}^n$ tends to 0 with n, as can be verified by realizing that the distance is maximized by the set of strings of weight at most $n/2$, and using Stirling's approximation for the binomial coefficients. Therefore,

$$\Pr_{x \in Y}[f(x) = Tx + u] \geq \epsilon^2/2,$$

and the claim follows by fixing the selection, in the definition of Y, of the bit to be set to 1. $\square$

We can now paste everything together to quickly conclude the proof of the Theorem 1.1 about testability of linear transformations.

Proof of Theorem 1.1 The proof is a straightforward composition of Theorems 3.1 and 4.1 and of Lemma 5.1. □

We remark that the dependence of ϵ' on ϵ in Theorem 1.1 is exponential, and this is due to the analgous loss in Ruzsa's theorem, see Remark 4.3.

6 Optimality of the graph-theoretic lemma in the BSG theorem

In this section we discuss the optimality of the path length in the graph-theoretic Lemma 3.2 that is the core of the proof of the BSG theorem. Recall that the lemma establishes that every dense graph contains a large subset such that every two nodes in the subset are connected by many paths of length 4. It is natural to ask if the path length can be reduced from 4, and one can quickly see that it cannot be set to 3, because the graph could be bipartite, and any large set would have two nodes on the same side which cannot be connected by a path of odd length 3. We now state and prove a result by Kostochka and Sudakov that also rules out path length 2. Thus, path length 4 is optimal in Lemma 3.2.

Theorem 6.1 ([KS])**.** *For every $\epsilon > 0$ there is a graph on N vertices with $N^2/3$ edges such that in every set of $\epsilon \cdot N$ nodes there are two nodes with less than $\epsilon \cdot N$ common neighbors.*

Proof of Theorem 6.1.
Let $n := \log_2 N$, and identify the set of N nodes with the binary strings of length n. Let $\Delta(u, v)$ denote the absolute Hamming distance between nodes u and v, i.e. the number of positions i such that $u_i \neq v_i$, and connect two nodes u and v if and only if $\Delta(u, v) \leq n/2$. Crudely, this graph has at least $N^2/3$ edges; we now show that it also has the desired property, if N is sufficiently large. The main idea is that any set of $\epsilon \cdot N$ nodes must contain two nodes at Hamming distance at least $n - O(\sqrt{n})$, but such two nodes have less than $\epsilon \cdot N$ common neighbors.

We now present the formal proof, starting with the next claim that gives us two distant nodes.

Claim 6.2. *Let S be any set of $\epsilon \cdot N$ nodes. Then S contains two nodes at Hamming distance at least $n - d \cdot \sqrt{n}$, where d is a constant that depends on ϵ only.*

Claim 6.2 follows from the isoperimetric fact that, among sets of the same cardinality, a Hamming ball has the smallest diameter. The following different proof was communicated to us by B. Sudakov, and relies on the following standard concentration bound (see, e.g., [DP, Theorem 5.18]).

Fact 6.3. *Let $f : \{0,1\}^n \to \Re$ be a 1-Lipschitz function, i.e. a function such that $|f(x) - f(y)| \leq 1$ if x and y differ in at most one coordinate. Then, for every $t \geq 0$, we have*

$$\Pr_{X \in \{0,1\}^n} \left[\left| f(X) - E_Y[f(Y)] \right| > t \right] \leq \alpha^{t^2/n},$$

for an absolute constant $\alpha < 1$.

Proof of Claim 6.2 First, let us note that if $\epsilon > 1/2$ then the claim is easily proved with maximal Hamming distance n, that is we can find two nodes u and v such that $\Delta(u, v) \geq n$. This is because we can pair off each node with its complement at distance n, and a set S of size $|S| \geq \epsilon \cdot N > N/2$ must take both nodes from some pair. To handle the case $\epsilon \leq 1/2$, we apply the argument to the set S' of nodes that are "close" to some node in S, which by a concentration bound has measure bigger than $1/2$.

For $x \in \{0,1\}^n$, let $f(x)$ denote the minimum Hamming distance of x from a node in S. Note that with probability ϵ a random node x falls in S, in which case $f(x) = 0$ and certainly $|f(x) - E_Y[f(Y)]| \geq E_Y[f(Y)]/2 =: t$. Since f is a 1-Lipschitz function, by Fact 6.3, we have

$$\epsilon \leq \alpha^{t^2/n},$$

from which we conclude that $E[f] \leq c \cdot \sqrt{n}$ for a constant $c = c(\epsilon)$. By choosing a larger constant $c' = c'(\epsilon)$ and applying the same Fact 6.3, we obtain that the probability that $f(X) \geq c' \cdot \sqrt{n}$ is strictly smaller than $1/2$. Therefore, the set S' of nodes at Hamming distance at most $c' \cdot \sqrt{n}$ from S has measure bigger than $1/2$. By the above argument, S' contains two nodes at distance n. Since each of these nodes is at distance at most $c' \cdot \sqrt{n}$ from some node in S, the set S contains two nodes at distance at least $n - 2 \cdot c' \cdot \sqrt{n}$. $\square$

Now that we have these two nodes at distance $n - 2 \cdot c' \cdot \sqrt{n}$, we conclude the proof by showing that the number of their common neighbors is less than $\epsilon \cdot N$. Without loss of generality, let these two nodes, which we denote by u_1 and u_2, respectively consist of the all-zero vector and of the vector which is 0 exactly in the first $k := 2 \cdot c' \cdot \sqrt{n}$ coordinates. Let us now see what nodes are common neighbors of u_1 and u_2. Let $X \in \{0,1\}^n$ be a node, and let $P = P(X)$ be its number of 1's in the first k coordinates, and $Q = Q(X)$ in the other $n - k$. The node X is a common neighbor to u_1 and u_2 precisely when $P + Q \leq n/2$ and $P + (n - k - Q) \leq n/2$. By combining the inequalities, we obtain that

$$n/2 - k + P \leq Q \leq n/2 - P, \qquad (\star)$$

i.e., for a given P, if X is a neighbor of both u_1 and u_2 then Q has to lie in a set of $2 \cdot P - k$ integers.

The intuition for the rest of the proof is as follows. A typical P is within $O(\sqrt{k})$ of $k/2$, and by $(\star)$ such a P constricts Q to lie in a set of $2 \cdot P - k = O(\sqrt{k})$ integers. As is well known, by Stirling's approximation the probability that Q is equal any particular integer is $O(1/\sqrt{n-k})$ (which is achieved for the integer $(n-k)/2$). Since $k = O(\sqrt{n})$ and n is large, we have $O(1/\sqrt{n-k}) = o(1/\sqrt{k})$, and so by a union bound Q falls in the set of $O(\sqrt{k})$ integers with probability tending to 0, and this proves the theorem.

More formally, let $d = d(\epsilon)$ be a sufficiently large constant to be determined later. Let us choose a random node X, and let $P = P(X)$ and $Q = Q(X)$ respectively denote the Hamming weight of its first k and last $n-k$ bits. We have:

$$\begin{aligned}
\Pr_{P,Q}[(\star)] &\leq \Pr_{Q}\left[(\star)\middle||P-k/2| \leq d\cdot\sqrt{k}\right] + \Pr_{P}\left[|P-k/2| > d\cdot\sqrt{k}\right] \\
&\leq 2\cdot d\cdot\sqrt{k}\cdot\Pr_{Q}\left[Q=(n-k)/2\right] + \epsilon/2 \\
&\leq O(d\cdot\sqrt{k}/\sqrt{n-k}) + \epsilon/2 \\
&< \epsilon.
\end{aligned}$$

Above, to bound the term $\Pr_Q\left[(\star)\middle||P-k/2| \leq d\cdot\sqrt{k}\right]$ we use a union bound and the fact that $\Pr_Q\left[Q=(n-k)/2\right] \geq \Pr_Q\left[Q=(n-k)/2+r\right]$ for any $r \in \Re$. To bound the term $\Pr_P\left[|P-k/2| > d\cdot\sqrt{k}\right]$ we apply Fact 6.3, choosing $d = d(\epsilon)$ to be sufficiently large. Later, we use Stirling's approximation to bound $\Pr_Q\left[Q=(n-k)/2\right]$. Finally, the last inequality holds for sufficiently large n recalling that $k = 2\cdot c'\cdot\sqrt{n}$.
End of the proof of Theorem 6.1.

Acknowledgment We thank Andrej Bogdanov, Vladimir Trifonov, and Avi Wigderson for helpful discussions. We are also grateful to Benny Sudakov for pointing out and explaining to us the paper [KS]. Finally, we thank the Columbia theory reading group for the opportunity to present this material and for helpful comments.

References

[BS] A. Balog and E. Szemerédi. A statistical theorem of set addition. *Combinatorica*, 14(3):263–268, 1994.

[BIW] B. Barak, R. Impagliazzo, and A. Wigderson. Extracting randomness using few independent sources. *SIAM J. Comput.*, 36(4):1095–1118 (electronic), 2006.

[BV] A. Bogdanov and E. Viola. Pseudorandom bits for polynomials. In *48th Annual Symposium on Foundations of Computer Science*. IEEE, Oct. 2007.

[DP] D. Dubhashi and A. Panconesi. Concentration of Measure for the Analysis of Randomised Algorithms, 2005. Manuscript. Available from `http://http://www.dsi.uniroma1.it/~ale/papers.html`.

[Fre] G. A. Freĭman. *Foundations of a structural theory of set addition*. American Mathematical Society, Providence, R. I., 1973. Translated from the Russian, Translations of Mathematical Monographs, Vol 37.

[Gow] W. T. Gowers. A new proof of Szemerédi's theorem for arithmetic progressions of length four. *Geom. Funct. Anal.*, 8(3):529–551, 1998.

[Gre] B. Green. Finite field models in additive combinatorics, 2004. arXiv:math/0409420v1; also http://www.dpmms.cam.ac.uk/ bjg23/.

[GT1] B. Green and T. Tao. An inverse theorem for the Gowers U^3 norm, 2005. arXiv.org:math/0503014.

[GT2] B. Green and T. Tao. Freiman's theorem in finite fields via extremal set theory, 2007. arXiv:math/0703668v1.

[GT3] B. Green and T. Tao. A note on the Freiman and Balog-Szemerdi-Gowers theorems in finite fields, 2007. arXiv:math/0701585v1.

[KS] A. Kostochka and B. Sudakov. On Ramsey numbers of sparse graphs. *Combin. Probab. Comput.*, 12(5-6):627–641, 2003. Special issue on Ramsey theory.

[Ruz] I. Z. Ruzsa. An analog of Freiman's theorem in groups. *Astérisque*, (258):xv, 323–326, 1999.

[Sam] A. Samorodnitsky. Low-degree tests at large distances. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, CA USA*, 2007.

[ST] A. Samorodnitsky and L. Trevisan. Gowers uniformity, influence of variables, and PCPs. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, WA, USA*, pages 11–20, 2006.

[SSV] B. Sudakov, E. Szemerédi, and V. H. Vu. On a question of Erdős and Moser. *Duke Math. J.*, 129(1):129–155, 2005.

[TV] T. Tao and V. Vu. *Additive combinatorics*, volume 105 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.

[Vio] E. Viola. New correlation bounds for GF(2) polynomials using Gowers uniformity. *Electronic Colloquium on Computational Complexity*, Technical Report TR06-097, 2006. `http://www.eccc.uni-trier.de/eccc`.

[VW] E. Viola and A. Wigderson. Norms, XOR lemmas, and lower bounds for GF(2) polynomials and multiparty protocols. In *Proceedings of the 22nd Annual Conference on Computational Complexity*. IEEE, June 13–16 2007.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/