COMBINATORICA
Akadémiai Kiadó – Springer-Verlag

COMBINATORICA 14 (3) (1994) 263–268

A STATISTICAL THEOREM OF SET ADDITION

ANTAL BALOG[1] and ENDRE SZEMERÉDI

Received April 21, 1992

1. Introduction

Let $A=\{a_1,\ldots,a_n\}$ be a finite set of integers and $A+A=\{a_i+a_j\}$ as usual. G. Freiman discovered in the late sixties that if $A+A$ is small compared to A then A has a certain structure. There are a few different ways (not all equivalent) of stating his result. The next one with a very elegant proof can be found in the recent work of I. Z. Ruzsa [2].

Freiman's theorem. *Let $A=\{a_1,\ldots,a_n\}$ be a finite set of integers. If $|A+A|\le c_1 n$ then there are integers d, $q_0,\ldots,q_d$, $X_1>0$, $\ldots$, $X_d>0$, and $c_2>0$ such that*

(1) $A\subset\{q_0+q_1x_1+\ldots+q_dx_d \mid 0\le x_1<X_1,\ldots,0\le x_d<X_d\}, X_1\ldots X_d\le c_2n,$

furthermore d and c_2 depend at most on c_1.

In other words if $A+A$ is small then A can be covered by a not much bigger multidimensional arithmetical progression. (Here and in the sequel all statements are understood for sufficiently large n compared to the actual constants $c_1,c_2,\ldots$. In many instances constants with higher indices can be calculated from constants with lower indices.) This theorem gives information about A whenever we have control over *all* sums a_i+a_j. In many cases, however, we have only control over *some* of the sums a_i+a_j. One example is a problem of P. Erdős. Let $A=\{a_1,\ldots,a_n\}$ be a finite set of integers. Suppose that A contains at least c_3n^2 three-term arithmetical progressions. Does it follow that A contains a k-term arithmetical progression for any k if n is large enough? The condition can be expressed as the equation

(2) $$a_i+a_j=2a_m$$

[1] Research supported by Hungarian NFSR grant 1901.

AMS subject classification code (1991): 11 B 05, 05 B 10, 11 B 75 has at least c_3n^2 solutions in A and this implies that at least c_3n^2 of the sums a_i+a_j fall into a small set, namely into $2A=\{2a_m\,|\,a_m\in A\}$. Actually, any linear equation in three variables serves equally well in place of (2), or any linear equation in $\ell\geq 3$ variables if it has at least $c_4n^{\ell-1}$ solutions. Our main theorem reduces this type of situation to the *all* case by showing that a large subset $A'\subset A$ satisfies the condition of Freiman's theorem.

The next Proposition expresses our main condition in three equivalent ways.

Proposition. *Let $A=\{a_1,\dots,a_n\}$ be a finite set of integers and $s(x)=\#\{x=a_i+a_j\}$. Consider the next three statements.*

(C1) $\sum\limits_x(s(x))^2\geq c_5n^3$.

(C2) *There exists $\mathcal{X}\subset\mathbb{Z}$, $|\mathcal{X}|\geq c_6n$ such that $s(x)\geq c_7n$ for $x\in\mathcal{X}$.*

(C3) *There exists $\mathcal{J}\subset[1,n]\times[1,n]$, $|\mathcal{J}|\geq c_8n^2$ such that $\#\{a_i+a_j\,|\,(i,j)\in\mathcal{J}\}\leq c_9n$.*

If any of these conditions holds for some positive constants c_i and for all sufficiently large n then so do the other two for some positive constants c_j calculable from the c_i and for all sufficiently large n.

The proof of this proposition is a straightforward use of the trivial facts.

$$s(x)\leq n;\quad \sum_x s(x)=n^2;\quad \sum_x(s(x))^2\leq n^3, \tag{3}$$

we leave the details to the reader.

Theorem. *Let $A=\{a_1,\dots,a_n\}$ be a finite set of integers. If any of the conditions (C1), (C2), or (C3) is satisfied then there are positive constants c_{10} and c_{11} and a subset $A'\subset A$ such that $|A'|\geq c_{10}n$ and $|A'+A'|\leq c_{11}n$.*

Combining this result with Freiman's theorem we get the next corollary.

Corollary. *Let $A=\{a_1,\dots,a_n\}$ be a finite set of integers. If any of the conditions (C1), (C2), or (C3) is satisfied then there are a positive constant c_{12} and a multi-dimensional arithmetical progression of the form (1) containing more than $c_{12}n$ elements of A.*

Now the affirmative answer for Erdős problem follows the above Corollary via Szemerédi theorem [3].

Szemerédi theorem. *Let $A\subset[1,n]$ be a set of integers with $|A|\geq c_{13}n$. For any integer $k\geq 3$ there is a k-term arithmetical progression in A if $n>n_0(k,c_{13})$.*

Actually, in this line of argument, Freiman's theorem can be substituted by something weaker, see Ruzsa [1] about an effective connection between arithmetical progressions and the number of sums. Note that the result can also be obtained directly, see [5].

Acknowledgement. The first named author is thankful to Prof. I. Z. Ruzsa and Prof. A. Hildebrand for useful conversations in the topic.

2. The Regularity Lemma

The proof is based on a graph theoretical lemma of E. Szemerédi, see [4]. We introduce some notations. Let $G=G(V,E)$ be a graph with vertex set V and edge set E. Given two subsets U, W of V we denote by $E(U,W)$ the set of edges joining U to W and we put $\Delta(U,W)=\frac{|E(U,W)|}{|U||W|}$. This is the density of edges joining U to W.

Regularity Lemma. *For any $\varepsilon>0$ there exists an integer $K(\varepsilon)$ such that for any graph $G=G(V,E)$ the set of vertices V can be divided into disjoint classes $V_0,V_1,\dots,V_K$ for some $K\le K(\varepsilon)$ with the properties $|V_i|\le\varepsilon|V|$ if $i=0,\dots,K$, $|V_i|=|V_j|$ if $i=1,\dots,K$, $j=1,\dots,K$ and for all but εK^2 pairs (i,j), $i<j$ the following condition holds. Whenever $U_i\subset V_i$, $|U_i|\ge\varepsilon|V_i|$, $U_j\subset V_j$, $|U_j|\ge\varepsilon|V_j|$ we have $|\Delta(U_i,U_j)-\Delta(V_i,V_j)|\le\varepsilon$.*

This lemma says that after omitting not too many edges any graph can be cut into very regular bipartity graphs. Note that an earlier version of the Regularity Lemma plays a crucial role in the proof of Szemerédi's theorem.

3. Proof of the Theorem

Let $A=\{a_1,\dots,a_n\}$ be a finite set of integers and $s(x)=\#\{x=a_i+a_j\}$. We assume that (C2) holds (and then so do (C1) and (C3) by the Proposition), i.e. there exists $\mathcal{X}\subset\mathbb{Z}$, $|\mathcal{X}|\ge c_6n$ such that $s(x)\ge c_7n$ for $x\in\mathcal{X}$. We have that

$$|\mathcal{X}|c_7n\le\sum_{x\in\mathcal{X}}s(x)\le\sum_x s(x)=n^2,$$

which implies

$$c_6n\le|\mathcal{X}|\le\frac{1}{c_7}n. \tag{4}$$

We consider the graph $G=G(V,E)$, where $V=A$ and $E=\bigcup_{x\in\mathcal{X}}E(x)$, $E(x)=\{\{a_i,a_j\}\,|\,a_i+a_j=x\}$. Note that $E(x)$ is a set of at least $\frac{1}{2}s(x)$ vertex independent edges (in the definition of $s(x)$ we did care the order of summands) and the edge sets $E(x)$ are pairwise disjoint. We say that an edge has "color" x if it belongs to $E(x)$. Thus we have

$$|E|\ge|\mathcal{X}|\frac{c_7}{2}n\ge\frac{c_6c_7}{2}n^2. \tag{5}$$

Put $\varepsilon=\frac{c_6c_7^2}{10}<\frac{1}{10}$ and apply the Regularity Lemma to G. We get a partition $V=V_0\cup V_1\cup\ldots\cup V_K$. Put $|V_i|=M\le\varepsilon n$ for $i=1,\dots,K$ and note that $(1-\varepsilon)\frac{n}{K}\le M\le\frac{n}{K}$. We define for $x\in\mathcal{X}$, $i=1,\dots,K$, $j=1,\dots,K$.

$$V_i(x)=\{a\in V_i\mid \text{there is a } b\in V \text{ such that } \{a,b\}\in E(x)\},$$

$$f(x,i,j)=\begin{cases}|E(x)\cap E(V_i,V_j)|, & \text{if } |V_i(x)|>\varepsilon M \text{ and } |V_j(x)|>\varepsilon M;\\ 0, & \text{otherwise.}\end{cases}$$

In other words $V_i(x)$ is the set of vertices of $E(x)$ in V_i and $f(x,i,j)$ is the number of edges of "color" x joining V_i to V_j whenever there are many vertices of $E(x)$ in both V_i and V_j. Finally let $\mathcal{J}$ be the set of those pairs (i,j), $1\le i<j\le K$ which are not among the at most εK^2 exceptional pairs given by the Regularity Lemma. Using the trivial bounds $|E(U,W)|\le|U||W|$ and $|E(x)\cap E(V_i,V)|\le|V_i(x)|$ we have

$$\begin{aligned}|E|\le&\sum_{x\in\mathcal{X}}\sum_{(i,j)\in\mathcal{J}}f(x,i,j)+\sum_{x\in\mathcal{X}}\sum_{\substack{1\le i\le K\\ |V_i(x)|\le\varepsilon M}}|E(x)\cap E(V_i,V)|+|E(V_0,V)|+\\&+\sum_{1\le i\le K}|E(V_i,V_i)|+\sum_{\substack{1\le i<j\le K\\(i,j)\notin\mathcal{J}}}|E(V_i,V_j)|\\ \le&\sum_{x\in\mathcal{X}}\sum_{(i,j)\in\mathcal{J}}f(x,i,j)+|\mathcal{X}|\varepsilon n+\varepsilon n^2+KM^2+\varepsilon K^2M^2,\end{aligned}$$

which implies by (4) and (5) that

$$\sum_{x\in\mathcal{X}}\sum_{(i,j)\in\mathcal{J}}f(x,i,j)\ge\frac{c_6c_7}{10}n^2. \tag{6}$$

We can fix a pair $(i,j)\in\mathcal{J}$ such that

$$\sum_{x\in\mathcal{X}}f(x,i,j)\ge\frac{c_6c_7}{5}\left(\frac{n}{K}\right)^2. \tag{7}$$

There are two important conclusions of (7). On one hand we have

$$|E(V_i,V_j)|\ge\frac{c_6c_7}{5}\left(\frac{n}{K}\right)^2,\quad\text{i.e. }\Delta(V_i,V_j)\ge\frac{c_6c_7}{5}, \tag{8}$$

on the other hand, as $f(x,i,j)\le|V_i|\le\frac{n}{K}$, we have

$$|\mathcal{X}'|\ge\frac{c_6c_7}{5K}n,\quad\text{where }\mathcal{X}'=\{x\in\mathcal{X}\mid f(x,i,j)>0\}. \tag{9}$$

By the definition of $f(x,i,j)$ we have for any $x\in\mathcal{X}'$ that $|V_i(x)|>\varepsilon M\ge\varepsilon(1-\varepsilon)\frac{n}{K}$. For any (not necessarily different) $x_1\in\mathcal{X}'$ and $x_2\in\mathcal{X}'$ the Regularity Lemma implies that $|\Delta(V_i(x_1),V_j(x_2))-\Delta(V_i,V_j)|\le\varepsilon$ and then from (8) we have

$$|E(V_i(x_1),V_j(x_2))|\ge\left(\frac{c_6c_7}{5}-\varepsilon\right)\left(\varepsilon(1-\varepsilon)\frac{n}{K}\right)^2\ge\frac{c_6^3c_7^5}{160K^2}n^2. \tag{10}$$

Next we will show that $\mathfrak{Z}=\mathcal{X}'+\mathcal{X}'$ is small, namely $|\mathfrak{Z}|\le\frac{160K^2}{c_6^3c_7^6}n$. For any $z\in\mathfrak{Z}$ we fix a representation $z=x_1+x_2$, where $x_1\in\mathcal{X}'$ and $x_2\in\mathcal{X}'$. For any $z\in\mathfrak{Z}$ we associate a set $\mathfrak{N}_z\subset\mathbb{Z}^3$ in the following way. Let $z=x_1+x_2$ be the fixed representation of z, let $\{b_1, b_2\}$ be an edge joining $V_i(x_1)$ to $V_j(x_2)$, there is exactly one edge in $E(x_1)$ (resp. $E(x_2)$) with vertex b_1 (resp. b_2), and let a_1 (resp. a_2) be the other vertex of this edge. We set $(b_1+b_2, a_1, a_2) \in \mathfrak{N}_z$. More formally we set

$$\mathfrak{N}_z = \{(x, a_1, a_2) \mid x \in \mathcal{X},\ a_1 \in A,\ a_2 \in A,\ \text{there is } \{b_1, b_2\} \in E(V_i(x_1), V_j(x_2)),\\ b_1 \in V_i(x_1),\ b_2 \in V_j(x_2),\ x = b_1 + b_2,\ a_1 = x_1 - b_1,\ a_2 = x_2 - b_2\}.$$

We have by (10)

$$|\mathfrak{N}_z| = |E(V_i(x_1), V_j(x_2))| \geq \frac{c_6^3 c_7^5}{160K^2} n^2, \tag{11}$$

and the sets $\mathfrak{N}_z$ are pairwise disjoint since $(x, a_1, a_2) \in \mathfrak{N}_z$ implies $x + a_1 + a_2 = z$. On the other hand the total number of such triplets cannot exceed $|\mathcal{X}|\,|A|^2$. Thus we have by (4) and (11) that

$$|\mathfrak{Z}| \frac{c_6^3 c_7^5}{160K^2} n^2 \leq \sum_{z \in \mathfrak{Z}} |\mathfrak{N}_z| = |\bigcup_{z \in \mathfrak{Z}} \mathfrak{N}_z| \leq \frac{1}{c_7} n^3,$$

which means (remember (9) and that $\mathfrak{Z} = \mathcal{X}' + \mathcal{X}'$) that we have found an $\mathcal{X}' \subset \mathcal{X}$ with

$$|\mathcal{X}'| \geq \frac{c_6 c_7}{5K} n, \quad |\mathcal{X}' + \mathcal{X}'| \leq \frac{160K^2}{c_6^3 c_7^6} n. \tag{12}$$

Finally let $E' = \bigcup_{x \in \mathcal{X}'} E(x)$ and $G' = G(V, E') \subset G$. We have $|E'| \geq \frac{c_7}{2} n |\mathcal{X}'| \geq \frac{c_6 c_7^2}{10K} n^2$ and then there is an $a_0 \in A$ such that at least $\frac{c_6 c_7^2}{10K} n$ edges of "color" $\in \mathcal{X}'$ start from it. The set of the other vertices of these edges will be suitable as A'. Indeed, $|A'| \geq \frac{c_6 c_7^2}{10K} n$ and if $a_i \in A'$, $a_j \in A'$ then $a_0 + a_i \in \mathcal{X}'$, $a_0 + a_j \in \mathcal{X}'$ and thus $2a_0 + a_i + a_j \in \mathcal{X}' + \mathcal{X}'$. (The indices i and j are arbitrary here, independent of their formerly fixed values.) This proves that $|A' + A'| \leq |\mathcal{X}' + \mathcal{X}'| \leq \frac{160K^2}{c_6^3 c_7^6} n$.

References

[1] I. Z. Ruzsa: Arithmetical progressions and the number of sums, to appear in *Periodica Math. Hung.*

[2] I. Z. Ruzsa: Generalized arithmetical progressions and sum sets, in preparation.

[3] E. Szemerédi: On sets of integers containing no k elements in arithmetic progression, *Acta Arithmetica* **27** (1975), 299-345.

[4] E. Szemerédi: Regular partitions of graphs, *Problèmes Combinatories at Theorie des Graphes*, (Ed. J-C. Bermond, et al.), CNRS aris, (1978) 399–401.

[5] E. SZEMERÉDI: no title, in preparation.

Antal Balog

Mathematical Institute of the
Hungarian Academy of Sciences
PO Box 127
Budapest, 1368–Hungary
H1165BAL@ELLA.HU

Endre Szemerédi

Rutgers University, New Brunswick
Department of Computer Science

Mathematical Institute of the
Hungarian Academy of Sciences
PO Box 127
Budapest, 1368–Hungary
SZEMERED@CS.RUTGERS.EDU