MARCINKIEWICZ CENTENARY VOLUME
BANACH CENTER PUBLICATIONS, VOLUME 95
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSZAWA 2011

ON FAMILIES OF WEAKLY DEPENDENT RANDOM VARIABLES

TOMASZ ŁUCZAK

Adam Mickiewicz University, Faculty of Mathematics and Computer Science
61-614 Poznań, Poland
E-mail: tomasz@amu.edu.pl

Abstract. Let $\mathcal{G}_n^{(k)}$ be a family of random independent k-element subsets of $[n] = \{1, 2, \ldots, n\}$ and let $\mathcal{H}(\mathcal{G}_n^{(k)}, \ell) = \mathcal{H}_n^{(k)}(\ell)$ denote a family of ℓ-element subsets of $[n]$ such that the event that S belongs to $\mathcal{H}_n^{(k)}(\ell)$ depends only on the edges of $\mathcal{G}_n^{(k)}$ contained in S. Then, the edges of $\mathcal{H}_n^{(k)}(\ell)$ are 'weakly dependent', say, the events that two given subsets S and T are in $\mathcal{H}_n^{(k)}(\ell)$ are independent for vast majority of pairs S and T. In the paper we present some results on the structure of weakly dependent families of subsets obtained in this way. We also list some questions which, despite the progress which has been made for the last few years, remain to puzzle researchers who work in the area of probabilistic combinatorics.

1. Introduction. On October 11th 1938 Józef Marcinkiewicz gave a talk at Poznań University entitled *The development of the probability theory for the last 25 years*. Soon after he was offered a position at the University starting September 1st, 1939. This date could be an important date for the history of mathematics in Poznań as well as for Marcinkiewicz's academic career. Unfortunately, instead it is remembered as the beginning of the Second World War which, let us remind, took also Marcinkiewicz's life.

Nowadays in Poznań there are several research groups active in the areas of mathematics which profit from Marcinkiewicz's immense and diverse scientific legacy. In this paper we report on some old and new developments concerning certain aspects of the theory of random discrete structures, a relatively new part of probability theory which has been extensively studied by probabilists, combinatorists, and computer scientists for the last thirty years. As we can see, although Marcinkiewicz was not directly involved in studies of such objects (the systematic studies of random graphs started only in the sixties) we can still link some recent results to theorems which most likely he mentioned in his lecture over seventy years ago.

2010 *Mathematics Subject Classification*: Primary 05C80; Secondary 11B25, 05C35.
Key words and phrases: random graph, hypergraph, arithmetic progression, limit theorem, extremal properties, large deviation, dependence.
The paper is in final form and no version of it will be published elsewhere.

DOI: 10.4064/bc95-0-9

2. Random hypergraphs. A k-uniform hypergraph H, or briefly k-graph, is a pair (V, E), where V is a set of vertices and E consists of k-element subsets of V, called edges. The main mathematical object we shall be interested in is $G^{(k)}(n,p)$ defined as the (random) k-graph with the vertex set $[n] = \{1, 2, \ldots, n\}$ such that each S, $S \subseteq [n]$, $|S| = k$, belongs to $G^{(k)}(n,p)$ with probability p, independently for each of $\binom{n}{k}$ k-element subsets of $[n]$. Note that each property of hypergraphs (such as, say, that it contains at least $n/2$ subsets) holds for $G^{(k)}(n,p)$ with some probability, and each characteristic of hypergraphs (such as the number of sets in a hypergraph) becomes in $G^{(k)}(n,p)$ a random variable (e.g. the number of edges of $G^{(k)}(n,p)$ is binomially distributed with parameters $\binom{n}{k}$ and p). $G^{(k)}(n,p)$ is called the binomial model of random hypergraph. Another, closely related model of a random hypergraph is $G^{(k)}(n,M)$, where we select a family of k-element sets uniformly at random from all families of M k-element subsets of $[n]$.

Thus, $G^{(k)}(n,p)$ as well as $G^{(k)}(n,M)$ are probabilistic spaces whose elements are k-graphs which are finite. However, we shall be interested only in its asymptotic behavior when $n \to \infty$. In particular, we often allow $p = p(n)$ or $M = M(n)$ to be a function of n and say that some property of $G^{(k)}(n,p)$ holds a.a.s. if the probability of this property tends to 1 as $n \to \infty$. Then, in many cases properties of both $G^{(k)}(n,p)$ and $G^{(k)}(n,M)$ are similar, i.e. these two models are equivalent in some well defined way (for a precise statement of this fact see [11]).

Random k-graphs have been extensively studied for the last twenty years. They can be used as an effective tool to show existence of k-graphs with some special properties (such as expanders), or serve as models of some real-world networks (e.g. internet graphs). One can also apply them to study the average case behavior of some algorithms, and to model and investigate the phase transition phenomena. In this paper we consider yet another application of random graphs of more probabilistic flavor – we shall treat them as models of weakly dependent families of random variables.

Let us suppose that $\mathcal{H}$ is a family of some (usually small) k-graphs whose vertex set is contained in $[n]$. Then, by $\mathcal{H}(n,p)$ $[\mathcal{H}(n,M)]$ we denote the family which consists of k-graphs from $\mathcal{H}$ with all edges contained in $G^{(k)}(n,p)$ $[G^{(k)}(n,M)]$. In this paper we concentrate on two particular examples of $\mathcal{H}$, although most of results remain true in a more general setting. The first one is the family $\mathcal{K}_\ell^{(k)}$ of all complete k-graphs on ℓ vertices contained in $[n]$; the second one of more algebraic flavor is the family $\mathcal{AP}_\ell$ of ℓ-element arithmetic progressions contained in $[n]$. Note that now we can treat $\mathcal{K}_\ell^{(k)}(n,p)$ and $\mathcal{AP}_\ell(n,p)$ as random ℓ-graphs whose edges are not independent but, in a way, weakly dependent. In order to get some feeling of the notion of weak dependence we shall freely use in a non-rigorous manner, let us look at the family $\mathcal{K}_3^{(2)}(n,p)$ of all triangles contained in $G^{(2)}(n,p)$. One can view $\mathcal{K}_3^{(2)}(n,p)$ as the family of the indicator variables X_{ijk}, where

$$X_{ijk} = \begin{cases} 1 & \text{if } \{i,j\}, \{j,k\}, \{i,k\} \text{ are edges of } G^{(2)}(n,p) \\ 0 & \text{otherwise.} \end{cases}$$

Since typically $p = p(n)$ is small and tends to 0 as $n \to \infty$, X_{ijk} is positively correlated with $3n$ other variables $X_{i'j'k'}$ such that $|\{i,j,k\} \cap \{i',j',k'\}| = 2$, and independent of the

values of $\binom{n}{3}-3n$ variables $X_{i''j''k''}$ for which $|\{i,j,k\}\cap\{i',j',k'\}|\le 1$. In $G^{(2)}(n,M)$ the situation is slightly more complex – the suitably defined random variable $\hat{X}_{ijk}$ is strongly positively correlated to $3n$ variables $\hat{X}_{i'j'k'}$ and weakly negatively correlated to all the others. Nonetheless, in both cases we are dealing with families of random triangles whose appearance in a given place is either independent or weakly correlated to the presence of most of remaining triangles in the family. The main goal of this paper is to describe and comment on some results on the structure of random hypergraphs $\mathcal{H}(n,p)$ of such a type, in particular random ℓ-graphs $\mathcal{K}_\ell^{(k)}(n,p)$ and $\mathcal{AP}_\ell(n,p)$.

3. Limit theorems. The most basic question on ℓ-graphs $\mathcal{K}_\ell^{(k)}(n,p)$ is the one about the distribution number of its edges. Equivalently, we want to know the number $X_\ell(n,p)$ of complete graphs on ℓ vertices contained in $G^{(k)}(n,p)$. This particular case is not very hard and well understood. Clearly, since $X_\ell(n,p)$ is a sum of weakly dependent identically distributed random indicator variables one can expect that it has asymptotically normal distribution and in most cases it is indeed so (provided both the expectation and variance of $X_\ell(n,p)$ tend to infinity as $n\to\infty$). One can prove for $X_\ell(n,p)$ the central limit theorem as well as its local version, and get good bounds on the rate with which the distribution tends to the normal one, using either the convergence of moments or more advanced techniques such as the Stein–Chen method or the orthogonal projection technique (see [11]). Here we describe yet another way of dealing with such problems which is relatively less known but has been proved useful to get limit theorems for other cases of a sum of weakly dependent random variables.

Let us first recall that for the random variable X with $\mathbb{E}|X|^j<\infty$ the jth semiinvariant (or the jth cumulant) $\kappa_j(X)$ of X is defined as

$$\kappa_j(X)=(-i)^j\frac{d^j}{(dt)^j}\log\phi_X(0)\,,$$

i.e. semiinvariants are coefficients of Maclaurin expansion of the characteristic function of X.

In 1939 Marcinkiewicz proved ([18], Théorème 2^{bis}) that if all high enough semiinvariants vanish, then the random variable is normal or degenerate.

THEOREM 3.1. *If for some m and all $j\ge m$ we have $\kappa_j(X)=0$, then $\kappa_j(X)=0$ for all $j\ge 2$, i.e. X is either degenerate, or has the normal distribution.*

Almost fifty years later Janson [9] showed the following asymptotic version of this result.

THEOREM 3.2. *Let $X_1,X_2,\dots$ be a sequence of random variables such that for some m*

$$\begin{aligned}&\kappa_1(X_n)=\mathbb{E}X_n\to\mu,\\&\kappa_2(X_n)=\mathbb{V}X_n\to\sigma^2>0,\\&\kappa_j(X_n)\to 0\quad\text{for all}\quad j\ge m.\end{aligned}$$

Then, $X_n\xrightarrow{D}N(\mu,\sigma^2)$ and for each $j\ge 1$ the moments $\mathbb{E}X_n^j$ converge to the jth moment of $N(\mu,\sigma^2)$.

More importantly, Janson noticed that from the above theorem one can immediately deduce the normal convergence for a sum of certain families of weakly dependent random variables.

THEOREM 3.3. *For each n let $\{Z_{n,i}\}_{i=1}^{N_n}$ be a family of uniformly bounded random variables, say $|Z_{n,i}| \le 1$. Let us also assume that each random variable $Z_{n,i}$ is independent of all but at most M_n variables from the family $\{Z_{n,i}\}_{i=1}^{N_n}$.*

Let $X_n = \sum_{i=1}^{N_n} Z_{n,i}$ and $\sigma^2(X_n) \to \sigma^2$. If for some m we have

$$\Big(\frac{N_n}{M_n}\Big)^{1/m} \frac{M_n}{\sigma_n} \longrightarrow 0 \quad \text{as} \quad n \to \infty,$$

then

$$\frac{X_n - \mathbb{E}X_n}{\sigma_n^2} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \to \infty.$$

Thus, for instance, for the case of triangles in $G^{(2)}(n,p)$ where $p < 1/2$ and $p^3 n \ge n^{\varepsilon}$ for some $\varepsilon > 0$, we have $N_n = \binom{n}{3}$, $M_n = 3n$ and, since all indicator variables for triangles are positively correlated, $\sigma_n^2 \ge \binom{n}{3} p^3(1-p^3) \ge 0.1 n^{2+\varepsilon}$. Then, setting $m = \lceil 6/\varepsilon \rceil$, we get

$$\Big(\frac{N_n}{M_n}\Big)^{1/m} \frac{M_n}{\sigma_n} \le 10 n^{2/m - \varepsilon/2} \le 10 n^{-\varepsilon/10} \to 0,$$

so the number of triangles after standardization tends to the normal distribution. As we have already mentioned for triangles this convergence can be shown more directly. However, in many other cases, where the structure of weakly dependent family of variables is more complex so that the estimating of the third moment is already difficult, Theorem 3.3 based on the asymptotic version of Marcinkiewicz's result proved to be extremely useful.

4. Large deviation theorems. Once we know that for the sum of weakly dependent variables the central limit theorem holds, we can ask if it is possible to prove a large deviation results similar to Chernoff's bounds for the tails of Bernoulli distributions. In many cases we can answer this question in the affirmative. Here we present only one such result by Janson, Łuczak, Ruciński [10] which is proved using Laplace transforms, for other ones we refer the reader to [11].

In order to do that in a general case of a sum of weakly dependent random variables we need to introduce some notation. Suppose that $\{J_i : i \in I\}$ is a finite family of independent 0–1 random variables. Let A be the family of subsets of I and for each $\alpha \in A$ let $X_\alpha = \prod_{i \in \alpha} J_i$. Finally, let $X_A = \sum_{\alpha \in A} X_\alpha$. Thus, for instance, if I is the family of all 2-element subsets of $[n]$, $P(J_i = 1) = 1 - P(J_i = 0) = p$, for each $i \in I$, and A is the family of triples which form a triangle, then X_A counts the number of triangles in $G^{(2)}(n,p)$.

Clearly, if all random variables X_A were independent and $P(X_\alpha = 1)$ was small for each $\alpha \in A$ we would have

$$P(X_A = 0) = \prod_{\alpha \in A} (1 - P(X_\alpha = 1)) \lesssim \exp\Big(-\sum_{\alpha \in A} P(X_\alpha = 1)\Big) = \exp(-\mathbb{E}X_A).$$

It turns out (see [10]) that a similar inequality holds for weakly dependent families of random variables provided we supplement the power in the left hand side by an additional

term

$$\sum\sum\nolimits_{\alpha\neq\beta,\alpha\cap\beta\neq\emptyset}\mathbb{E}X_\alpha X_\beta$$

which clearly measures how dependent is the family $\{X_\alpha\}_{\alpha\in A}$.

THEOREM 4.1.

$$\Pr(X_A=0)\le\exp\Big(-\frac{(\mathbb{E}X_A)^2}{\sum\sum_{\alpha\cap\beta\neq\emptyset}\mathbb{E}X_\alpha X_\beta}\Big)\le\exp\Big(-\mathbb{E}X_A+\sum\sum\nolimits_{\alpha\neq\beta,\alpha\cap\beta\neq\emptyset}\mathbb{E}X_\alpha X_\beta\Big).$$

It turns out that the above upper estimate gives the correct value of $\log\Pr(X_A=0)$ up to a constant factor and similar estimates can be obtained also for the random variables counting edges of $\mathcal{K}_\ell^{(k)}(n,M)$ (see [10] or [11]). In particular, the probability that $G^{(2)}(n,p)$ contains no triangles is $\exp(-\Theta(n^3p^3))$ for $np^2\le 1$ and $\exp(-\Theta(n^2p))$ whenever $np^2\ge 1$. Similarly, the probability that there are no triangles in $G^{(2)}(n,M)$ is $\exp(-\Theta(M^3/n^3))$ if $M\le n^{3/2}$, and $\exp(-\Theta(M))$ if $n^{3/2}\le M\le n/3$. We also mention that if, say, $p=cn^{-1/2}$ for some constant $c>0$, then we expect

$$\lim_{n\to\infty}\log P(X_A=0)n^{3/2}=-a(c)\,,$$

for some constant $a(c)>0$, but at this moment we do not know the correct value of $a(c)$.

5. Beyond large deviation results. So far we have just counted the number of edges in $\mathcal{K}_\ell^{(k)}(n,M)$; now we try to say something about the structure of these graphs. Thus, let us ask the 'extremal type' question: what is the minimum number $Y=Y(n,M)$ of edges one should delete from $G^{(2)}(n,M)$, to get rid of all triangles contained in it? Clearly, Y is bounded from below by the maximum number of edge-disjoint triangles contained in $G^{(2)}(n,M)$, and bounded from above by the number of all triangles in $G^{(2)}(n,M)$. If $M=o(n^{2/3})$ a.a.s. these two random variables are not far from each other so we can determine Y quite precisely. The question becomes much more interesting when $M\gg n^{2/3}$, i.e. when the expected number of triangles becomes much larger than the number of edges. Then, a natural upper bound for Y is $M/2$, since one can destroy all triangles in $G^{(2)}(n,M)$ by splitting its vertices into two equal parts and removing all the edges inside each of the parts (note that in this way we delete roughly half of the edges). Frankl and Rödl [3] and Haxell, Kohayakawa and Łuczak [8] showed that if $Mn^{-2/3}\to\infty$, then this upper bound gives the asymptotically correct value of Y, i.e. for $M\gg n^{3/2}$ we have a.a.s. $Y=(1+o(1))M/2$.

THEOREM 5.1. *If $Mn^{-2/3}\to\infty$ as $n\to\infty$, and $\eta>0$, then a.a.s. each subgraph of $G^{(2)}(n,M)$ with more than $(1/2+\eta)M$ edges contains a triangle.*

The proof of the above result was based on the following idea which is an example of the *pseudorandom paradigm* (see Łuczak [17] and Tao [26]). Basically we shall argue that a.a.s. subgraph of $G^{(2)}(n,M)$ contains a large 'pseudorandom', 'essentially non-bipartite' subgraph (this part of the argument is mainly deterministic), and then use a probabilistic result (Lemma 5.2 below) to infer that such a subgraph must contain a triangle.

In order to be slightly more precise let us assume that we count all subgraphs H of $G^{(2)}(n,M)$ which are triangle-free and have at least $(1/2+\varepsilon)M$ edges. The number of such subgraphs is bounded from above by 2^M and the probability that H contains no triangles

is the same as the probability that $G^{(2)}(n,(1/2+\varepsilon))M$ contains no triangles. Hence, if we could only show that the last probability is much smaller than 2^{-M} we would be done. However it is plainly wrong – the probability that each edge of $G^{(2)}(n,(1/2+\varepsilon))M$ has one end in $\{1,2,\ldots,\lfloor n/2\rfloor\}$ and the other in $\{\lfloor n/2\rfloor+1,\ldots,n\}$ is roughly $2^{-(1/2+\varepsilon)M} \gg 2^{-M}$.

Now we use pseudorandomness for the first time. Although it is not true that the probability that $G^{(2)}(n,(1/2+\varepsilon))M$ contains no triangles is smaller that 2^{-M} we may still hope to get a much better bound if we condition on the event that $G^{(2)}(n,(1/2+\varepsilon))M$ is 'pseudorandom'. And, indeed, it turns out that there exist properties $\mathcal{B}(\varepsilon)$ and $\mathcal{T}(\varepsilon)$, such that conditioned on $\mathcal{B}(\varepsilon)\cap\mathcal{T}(\varepsilon)$ the probability that $G^{(2)}(n,M)$ contains no triangles significantly decreases, i.e. the following holds.

LEMMA 5.2. *For each $\varepsilon>0$ there exists $c>0$ such that for every $M\ge cn^{3/2}$ we have*

$$\Pr\big(G^{(2)}(n,M) \text{ contains no triangles} \mid \mathcal{B}(\varepsilon)\cap\mathcal{T}(\varepsilon)\big) \le \varepsilon^M. \tag{1}$$

Precise definitions of $\mathcal{B}(\varepsilon)$ and $\mathcal{T}(\varepsilon)$ are somewhat technical but roughly $\mathcal{T}(\varepsilon)$ states that one cannot remove from a graph an ε fraction of edges and make it bipartite (i.e. it states that a graph is 'essentially non-bipartite'), while $\mathcal{B}(\varepsilon)$ means that the edges of graphs are uniformly distributed, where this uniformity is measured by a parameter $\varepsilon>0$ (i.e. a graph is 'pseudorandom'). In order to use Lemma 5.2 one should match it with so called Regularity Lemma, one of the most efficient tools of modern random graph theory. The Regularity Lemma was first proved in Szemerédi's celebrated paper on the Density Theorem [25] and was adapted for sparse graphs independently by Kohayakawa and Rödl. In this setting it states that each dense subgraph H of a graph G whose edges are, in some way, uniformly distributed (as is the case, for instance, in $G^{(2)}(n,k)$) contains a relatively large and dense subgraph F which fulfils the uniformity condition $\mathcal{B}(\varepsilon)$, i.e. each dense subgraph contains a large 'pseudorandom subgraph'. Now Theorem 5.1 can be proved as follows. We choose $\varepsilon>0$ much smaller than η. From Lemma 5.2 we deduce that a.a.s. each large subgraph F of $G^{(2)}(n,M)$ for which both $\mathcal{T}(\varepsilon)$ and $\mathcal{B}(\varepsilon)$ hold contains a triangle. Now we take a subgraph H of $G^{(2)}(n,p)$ with more than $(1/2+\eta)M$ edges. We apply to it Regularity Lemma and deduce that it contains a large subgraph $F'\subseteq H$ which has property $\mathcal{B}(\varepsilon)$. Furthermore, since H contains more than half of all edges one can argue that there is a large subgraph $F\subseteq F'$ which has both properties $\mathcal{T}(\varepsilon)$ and $\mathcal{B}(\varepsilon)$ and, as we have already mentioned, a.a.s. each such F (and thus H) must contain a triangle.

Can one repeat this argument to prove a result analogous to Theorem 5.1 for $\mathcal{K}^{(k)}_\ell(n,M)$, i.e. for ℓ-cliques in random k-graphs? It should be possible but at this moment we do not know how to do that. For graphs the main obstacle is that we do not know if the statement analogous to Lemma 5.2 holds not only for triangles but for all graphs. This so called KŁR Conjecture (for its rigorous statement see Kohayakawa, Łuczak and Rödl [13]) has been verified only for special cases of graphs such as cycles, cliques of size four and five, and certain classes of bipartite graphs (see Gerke, Schickinger, Steger [5] and the references therein). For k-graphs with $k\ge 3$ the problem is much more difficult – the statement of Regularity Lemma for hypergraphs is so complicated it is not even clear what should be a rigorous statement of 'uniformity' condition $\mathcal{B}_k(\varepsilon)$.

6. Arithmetic progressions in random subsets. In the previous section we discussed the structure of $\mathcal{K}_\ell^{(k)}(n,M)$, now we look at the random hypergraph $\mathcal{AP}_\ell(n,M)$. Let us first state the problem analogous to that we considered in the previous section for k-graphs. We say that a subset $A \subseteq [n]$ has property $\mathcal{P}(\eta,\ell)$ if every B, $B \subseteq A$, such that $|B| \geq \eta|A|$, contains a nontrivial arithmetic progression of length ℓ. We want to find the threshold function $M_{\eta,\ell}(n)$ such that if $M/M_{\eta,\ell} \to 0$, then a.a.s. the random set $G^{(1)}(n,M)$ has no property $\mathcal{P}(\eta,\ell)$, while for $M/M_{\eta,\ell} \to \infty$ the property $\mathcal{P}(\eta,\ell)$ a.a.s. holds for $G^{(1)}(n,M)$. It is easy to check that if $Mn^{-(\ell-2)/(\ell-1)} \to 0$, then the number of non-trivial arithmetic progressions of length ℓ is much smaller than M, i.e. $\mathcal{P}(\eta,\ell)$ does not hold. Thus, it is natural to conjecture that $M_{\eta,\ell} = n^{(\ell-2)/(\ell-1)}$ is the threshold for $\mathcal{P}(\eta,\ell)$, i.e. if $Mn^{-(\ell-2)/(\ell-1)} \to \infty$, then a.a.s. $\mathcal{P}(\eta,\ell)$ holds for $G^{(1)}(n,M)$.

This problem is hard even in the simplest possible 'deterministic' case when $M = n$, i.e. when we ask if the set $[n]$ has property $\mathcal{P}(\eta,\ell)$. For $\ell = 3$ it was proved by Roth [22] in 1953, while for general $\ell \geq 3$ it was settled by Szemerédi [25] in 1975. Soon after an entirely different argument, based on ergodic theory, was given by Furstenberg [4]. The subject was revived by Gowers, who provided another proof of this fact using so called Gowers' norms [6]. Yet another proof, this time based on a version of Regularity Lemma for hypergraphs, was given by Nagle, Rödl, Skokan and Schacht [19], [21] (see also [20]), and independently by Gowers [7]. In the above results the constant $\varepsilon > 0$ can be replaced by a function $r_\ell(n)$ which tends slowly to 0 as $n \to \infty$; the question of the correct rate of convergence of this function is not known even for $r_3(n)$ (the best estimate for $r_3(n)$ are given by Bourgain [1]) and remains a major open question in combinatorial number theory.

It should be pointed out that although nowadays we know at least four different proofs of Szemerédi's Density Theorem each of them is either hard or invokes deep results from graphs theory or ergodic theory. Thus, it seems hopeless to expect that we can prove similar theorems for much harder 'random set' case. However, somewhat surprisingly, it can be done. First such attempt has been made by Kohayakawa, Łuczak and Rödl [12], who used an idea of Ruzsa and Szemerédi [21] and successfully treated the case $\ell = 3$. For a solution of the case $\ell \geq 4$ we waited much longer; the breakthrough has come only recently with the papers of Conlon and Gowers [2] and Schacht [24] who described a fairly general method to deal with this and similar cases. In particular their techniques provide solutions for most of the problems which can be solved using KŁR conjecture described in the previous section, although they do not solve the conjecture itself.

The main idea behind the method of Conlon and Gowers [2] and Schacht [24] is to rigorously justify the paradigm which has been known for a long time to experts in random graph theory: if $M = M(n)$ is large enough then the random graph $G^{(k)}(n,M)$ can be treated as a sparse version of the complete graph $G^{(k)}(n,\binom{n}{k})$. Thus, for instance, suppose that B is a subset of some 'random enough' set A, $|A| = pn$, such that $|B| \geq |A|/100$. From that and the fact that each subset $\hat{B}$ of $[n]$ of $n/100$ elements contains at $\Omega(|\hat{B}|^2)$ arithmetic progressions of length 3 we want to deduce that B contains $\Omega(p|B|^2)$ arithmetic progressions of length 3. The idea of Conlon and Gowers can be roughly described as follows. They view the statement of subsets of $[n]$ as the statement about the indicator

function of $\hat{B}$ and show that the analogous fact remains true not only for indicator functions of the sets but also for any function which is close to an indicator function in some special norm (designed especially for the arithmetic progressions of length three). It turns out that the rescaled indicator function of $B \subseteq A$, provided A is random enough, is close enough to true indicator function of dense subset in $[n]$ so the assertion follows. The argument of Schacht is purely combinatorial. Suppose that we select vertices from $B \subseteq A$ and $\hat{B} \subseteq [n]$ one by one so that we would like to minimize the final number of arithmetic progressions of length 3. It turns out that for random enough A selecting kp elements from B would create roughly the same number of potential candidates for arithmetic progressions in $[n]$ as we would expect if we were to select k elements of some $\hat{B}$; furthermore, one can argue that, since A is random-like, roughly p-fraction of all these candidates are in A. Hence, if it is impossible to avoid $\Omega(|\hat{B}|^2)$ arithmetic progressions of length in $\hat{B}$, it is also impossible to have fewer than $\Omega(p|B|^2)$ arithmetic progressions in B. The proofs of the above arguments are technically very involved but the main theorems in both papers are stated in such a general way that one can use these results as in a variety of situations by verifying a few technical conditions. The importance and significance of these two papers for random structures theory can hardly be overestimated.

7. Future directions. At first sight it seems that nowadays we have got a fair understanding of property $\mathcal{P}(\eta, \ell)$ and extremal properties of k-graphs, and after the papers of Conlon and Gowers [2] and Schecht [24] we also know how to study the structure of $\mathcal{AP}_\ell(n, M)$. We conclude this note with a few remarks suggesting this is not quite the case.

The most intriguing (and annoying) issue which bothers experts working in this area is of somewhat 'philosophical' nature (but, as we shall see shortly, solving it can have some well defined mathematical consequences). In order to explain the problem let us look once again at the Density Theorem. Besides Szemerédi's original argument there are at least three other proofs of this result: Furstenberg's ergodic argument, the proof based on Gowers' norm, and one which deduces the Density Theorem from extremal properties of k-graphs (such as the Regularity Lemma or one of its consequences, the Removal Lemma). All these methods are based on a 'pseudorandom paradigm' which states, in a most general form, that each dense structure contains a large pseudorandom substructure. Are these methods essentially different, or do they use just different settings to present the same argument? We do not know. They certainly give different estimates, say, for $r_3(n)$: the ergodic argument gives $o(n)$, using a 'naive' hypergraph approach of Ruzsa and Szemerédi [21] one gets $r_3(n) = O(n/\log^* n)$, and Roth's original argument gives $r_3(n) = n/\log\log n$. Two last estimates can be improved but one cannot hope to do much better that Roth's estimate by purely combinatorial means. However, it is not at all clear if the reason for that lies in the additional symmetries of the problem (for the proof of the Density Theorem it is enough to know properties of very symmetric k-graphs) or, perhaps, these methods are intrinsically different. The same question can be asked about the techniques developed to deal with random structures. Is there any direct connection between analytic approach of Conlon and Gowers and combinatorial argument of Schacht?

It is not clear. This is quite unfortunate since understanding this relation could possibly shed some new light on many related questions. For instance, analysts may want to have a combinatorial version of Gowers' norm. On the other hand combinatorists are struggling with finding better measures for pseudorandomness of sparse k-graphs. Analytic tools have already been known to be very useful in this case (e.g. for investigating graphons, cf. Lovász and Szegedy [15]) but it seems that we see just the tip of the iceberg.

As we have mentioned in the previous section the general approach of Conlon and Gowers and Schacht can be used to show many consequences of KŁR conjecture. However, some of them cannot be deduced in this way. For instance, suppose that we choose a graph from the family of all triangle-free graphs with n vertices and M edges and ask for which $M = M(n)$ it is a.a.s. bipartite, i.e. we want to find $M = M(n)$ such that

$$P\big(G^{(2)}(n,M) \text{ is bipartite} \mid G^{(2)}(n,M) \text{ is triangle-free}\big) = 1 - o(1).$$

From KŁR conjecture (which holds for triangles) it follows that the threshold function M is $\Theta(n^{3/2})$, and if the conjecture holds it would give thresholds for analogous properties of this type (for details see Łuczak [16]).

We conclude with one of several problems in this area which are natural, important, and easy to state but which seem to be out of our reach at this moment. Let us call a k-graph linear if each pair of its edges shares at most one vertex. Note that since every pair of vertices of a linear k-graph on n edges is contained in at most one edge, a k-graph has got at most $\binom{n}{2}$ edges.

Conjecture. For each $\alpha > 0$ and $k \geq 3$ there exists n_0 such that each linear k-graph with $n \geq n_0$ vertices and αn^2 edges contains k edges of type $\{a_0, a_1^1, a_2^1, \ldots, a_{k-1}^1\}$, $\{a_0, a_1^2, a_2^2, \ldots, a_{k-1}^2\}, \ldots, \{a_0, a_1^{k-1}, a_2^{k-1}, \ldots, a_{k-1}^{k-1}\}$, and $\{a_0', a_1^1, a_2^2, \ldots, a_{k-1}^{k-1}\}$.

The above conjecture holds for $k = 3$. Moreover, it is easy to see that for a given k it implies the Density Theorem for arithmetic progressions of length k, so its proof is probably hard. On the other hand, there is a general believe that we can show it using pseudorandom paradigm; unfortunately, we do not have a slightest idea how the definition of pseudorandom linear k-graph should look like!

References

[1] J. Bourgain, *Roth's theorem on progressions revisited*, J. Anal. Math. 104 (2008), 155–192.

[2] D. Conlon, W. T. Gowers, *Combinatorial theorems in sparse random sets*, `arXiv:1011.4310v1`

[3] P. Frankl, V. Rödl, *Large triangle-free subgraphs in graphs without K_4*, Graphs Combin. 2 (1986), 135–144.

[4] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. 31 (1977), 204–256.

[5] S. Gerke, T. Schickinger, A. Steger, *K_5-free subgraphs of random graphs*, Random Structures Algorithms 24 (2004), 194–232.

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

[7] W. T. Gowers, *Quasirandomness, counting and regularity for 3-uniform hypergraphs*, Combin. Probab. Comput. 15 (2006), 143–184.

[8] P. E. Haxell, Y. Kohayakawa, T. Łuczak, *Turán's extremal problem graphs: forbidding odd cycles*, Combinatorica 16 (1996), 107–122.

[9] S. Janson, *Normal convergence by higher semi-invariants with applications to sums of dependent random variables and random graphs*, Ann. Probab. 16 (1988), 305–312.

[10] S. Janson, T. Łuczak, A. Ruciński, *An exponential bound for the probability of nonexistence of a specified subgraph in a random graph.* In: Random Graphs (Poznań, 1987), Wiley, Chichester, 1990, 73–87.

[11] S. Janson, T. Łuczak, A. Ruciński, *Random Graphs*, Wiley, New York, 2000.

[12] Y. Kohayakawa, T. Łuczak, V. Rödl, *Arithmetic progressions of length three in subsets of a random set*, Acta Arith. 75 (1996), 133–163.

[13] Y. Kohayakawa, T. Łuczak, V. Rödl, *On K_4-free subgraphs of random graphs*, Combinatorica 17 (1997), 173–213.

[14] Y. Kohayakawa, V. Rödl, *Regular pairs in sparse random graphs*, Random Structures Algorithms 22 (2003), 359–434.

[15] L. Lovász, B. Szegedy, *Szemerédi's Lemma for the analyst*, Geom. Funct. Anal. 17 (2007), 252–270.

[16] T. Łuczak, *On triangle-free random graphs*, Random Structures Algorithms 16 (2000), 260–276.

[17] T. Łuczak, *Randomness and regularity*, in: International Congress of Mathematicians, Vol. III, Eur. Math. Soc., Zürich, 2006, 899–909.

[18] J. Marcinkiewicz, *Sur une propriété de la loi de Gauss*, Math. Z. 44 (1939), 612–618.

[19] B. Nagle, V. Rödl, M. Schacht, *The counting lemma for regular k-uniform hypergraphs*, Random Structures Algorithms 28 (2006), 113–179.

[20] V. Rödl, B. Nagle, J. Skokan, M. Schacht, Y. Kohayakawa, *The hypergraph regularity method and its applications*, Proc. Natl. Acad. Sci. USA 102 (2005), 8109–8113.

[21] V. Rödl, J. Skokan, *Regularity lemma for k-uniform hypergraphs*, Random Structures Algorithms 25 (2004), 1–42.

[22] K. F. Roth, *On certain sets of integers*, J. London Math. Soc. 28 (1953), 104–109.

[23] I. Ruzsa, E. Szemerédi, *Triple systems with no six points carrying three triangles*, in: Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II, Colloq. Math. Soc. János Bolyai 18 (1978), North-Holland, Amsterdam, 939–945.

[24] M. Schacht, *Extremal results for discrete random structures*, preprint, `http://www.math.uni-hamburg.de/home/schacht/publikationen.html.en`

[25] E. Szemerédi, *On sets of integers containing no k elements in arithmetic progression*, Acta Arith. 27 (1975), 199–245.

[26] T. Tao, *The dichotomy between structure and randomness, arithmetic progressions, and the primes*, in: International Congress of Mathematicians, Vol. I, Eur. Math. Soc., Zürich, 2007, 581–608.