Concentration of Non-Lipschitz Functions and Applications

V. H. Vu*
Department of Mathematics, University of California at San Diego, La Jolla, California; e-mail: vanhavu@ucsd.edu; web: http://www.math.ucsd.edu/~vanvu

Received 27 December 2001; accepted 28 January 2002

DOI 10.1002/rsa.10032

ABSTRACT: Strong concentration results play a fundamental role in probabilistic combinatorics and theoretical computer science. In this paper, we present several new concentration results developed recently by the author and collaborators. To illustrate the power of these new results, we discuss applications in many different areas of mathematics, from combinatorial number theory to the theory of random graphs. © 2002 Wiley Periodicals, Inc. Random Struct. Alg., 20: 262–316, 2002

1. OVERVIEW

Strong concentration (or large deviation) inequalities are vital tools in probabilistic combinatorics and several related areas, such as theoretical computer science. The usual way to prove a strong concentration inequality is via either Azuma's (bounded martingale method) or Talagrand's inequalities. These inequalities work remarkably well when applied to functions with relative small (discrete) Lipschitz coefficients. However, they become less effective when the Lipschitz coefficient is large. Recently, J. H. Kim and the present author [40, 79] developed a new type of inequality which can be used to handle the latter case. These new inequalities have been used with considerable success in order to attack intricate problems in diverse areas, ranging from random graphs to finite geometry, leading to several improvements [41, 74, 75, 77, 78].

The present paper has two goals. The first is to give an up-to-date account about these new inequalities. Beside the results, we shall also describe the intuition behind our development and discuss the key ideas of the proofs. We believe that some of these ideas (especially Lemma 3.1) can be applied in situations far more general than those considered

*Most of this work was done while the author was with Microsoft Research.

in this paper and therefore could be of independent interest. Also in this part of the paper, we shall present the proof of our most recent and most general inequalities.

The second part of the paper is a survey on applications, which touch an amazingly large variety of basic concepts in combinatorics: random graphs, matching and coloring, finite projective planes, additive bases, to mention a few. A pleasant feature about these applications is that in many cases (see Sections 6.2, 6.3, and 8.2, for instance), our inequalities not only solve the problem at hand, but also lead to the rediscovery of important notions (such as the notion of balanced graphs in Section 6.2 or that of codegree in Section 8.2). This makes it possible to view these notions from a different aspect and provides a quantitative way to explain their necessity in the given problems. Another appealing fact is that our inequalities are easy to use; their applications do not require the sort of ingenuity one frequently needs when using, for instance, the martingale method.

Throughout the paper, we consider real-value functions depending on n variables $t_1, \ldots, t_n$, where the t_i's are independent random variables with arbitrary distributions in $[0, 1]$. We denote by Ω the product space spanned by the t_i's, equipped with the natural product measure. In most applications, the t_i's are binary (assuming only two values 0 and 1) and Ω is the n-dimensional unit hypercube. The lower case letters a, b, c, d denote positive constants, whose values may be different at each occurrence. $\mathbb{R}$ and $\mathbb{N}$ denote the set of real and natural numbers, respectively.

The paper is organized as follows. The first part consists of four sections, 2, 3, 4, and 5. Section 2 presents the general phenomenon of the theory of strong concentration and some classical concentration results such as Azuma's and Talagrand's inequalities, together with the obstacle these inequalities face when applied to functions with large Lipschitz coefficients. In the next section, we start by describing our intuition and main ideas, and follow by our general scheme for proving a concentration result for a certain class functions with large Lipschitz coefficient. We end this section with our key tool, a new martingale inequality (Lemma 3.1). In Section 4, we present our new concentration results and make a comparison with the classical ones on a well-known problem in the theory of random graphs. In this section, we also pose several questions which may stimulate further research and state the so-called polynomial method, which appears to be useful in many applications. Section 5 contains the proofs of the most recent results which nicely illustrate our general method.

While our method appears to be fairly general, our concentration results mainly focus on a special class of functions: polynomials with positive coefficients. This has two reasons. The first is that polynomials appear as the most natural candidate for our method. The second is that in all applications found so far, the objective function is a polynomial or can be very well approximated by one. On the other hand, our proofs do not rely too much on the properties of polynomials, so we can easily modify them to obtain results for a larger class of functions. Few examples of this kind are presented in Section 4.5, and we hope that the emergence of new applications will lead to new developments in this direction.

The second part of the paper consists of three sections: 6, 7, and 8. These sections are devoted to applications in diverse areas of mathematics and can be read independently. Section 6 discusses applications in random graph theory and similar random structures, focusing on the classical subgraph count problem, where our new inequalities provide breakthrough results on several questions. Section 7 contains applications in combinatorial number theory, including a solution to an open problem of Nathanson on the existence of thin Waring basis [53], posed in 1980. The final and richest section, Section 8 describes

a powerful combination of our results with the so-called "semirandom" method, a sophisticated method in probabilistic combinatorics. This combination is the key tool in the proofs of several highly non-trivial results [41, 79, 80]. These include the solution to a famous and long-standing open problem of Segre in finite geometry, dating back to the 1950s, and several improvements in two topics: nearly perfect matchings in hypergraphs and list coloring of locally sparse graphs. These improvements all have far-reaching and nontrivial consequences in several other problems, some of which will be mentioned briefly. Due to the complexity of the proofs, we shall only be able to present the sketch of the simplest proof, the proof of a recent theorem on nearly perfect matchings from [79]. On the other hand, as polynomials arise very naturally in applications of the semirandom method (as shown in this sketch) and as our approach is quite robust, we hope that this sketch not only convinces the reader about the power of our tools, but also makes the semirandom method more accessible for nonexperts.

2. CLASSICAL CONCENTRATION RESULTS

2.1. Chernoff, Azuma, Talagrand

Let us start by describing what we mean by "strong concentration." A typical strong concentration result is the following, due to Chernoff.

Theorem 2.1. *Let $Y = \sum_{i=1}^{n} t_i$, where t_i are i.i.d binary random variables with mean p. Then for any $\lambda > 0$*

$$Pr(|Y - \mathbb{E}(Y)| \geq \sqrt{\lambda n}) \leq 2e^{-\lambda/2}.$$

We say that a function is strongly concentrated if it satisfies an exponential deviation bound of Chernoff's type. Inequalities obtained by estimates of moments of small order (such as Chebysev's inequality) usually do not provide exponential bounds and are not discussed in this paper.

To know whether a function is strongly concentrated is an issue of fundamental importance in a number of areas in mathematics. This question has been investigated by great mathematicians for centuries and the excellent survey by M. Talagrand [67] is a good place to look for historical background and the most significant developments on the topic. Central to this survey as well as to the whole theory of concentration is the following phenomenon

If Y depends smoothly on the atom variables $t_1, \ldots, t_n$,
then Y is strongly concentrated. (1)

The traditional definition of smoothness uses the notion of Lipschitz coefficient. A function $Y = Y(t_1, \ldots, t_n)$ from Ω to $\mathbb{R}$ is called *r-Lipschitz* if

$$|Y(t) - Y(t')| \leq r,$$

whenever the vectors t and t' differ at exactly one coordinate. This definition of Lipschitz coefficient is motivated by applications of discrete nature, and is somewhat different from

the definition used in analysis. One then says that Y is smooth if r is relatively small. In the following we present two well-known results which nicely illustrate phenomenon (1). These results are not stated in the most general form, but in the form commonly used in combinatorics and theoretical computer science.

The first result is Azuma's inequality ([8], Chapter 7).

Theorem 2.2. *If Y has Lipschitz coefficient r, then for any $\lambda > 0$*

$$Pr(|Y - \mathbb{E}(Y)| \geq r\sqrt{\lambda n}) \leq 2e^{-\lambda/2}.$$

To state the second result, we first need to introduce the notion of "certificate." Let f be a function from $\mathbb{N}$ to $\mathbb{N}$. A function Y from Ω to $\mathbb{R}$ is *f-certifiable* if whenever $Y(t) \geq b$, there exists an index set $\mathcal{I}$ of at most $f(b)$ elements so that every $t' \in \Omega$ that agrees with t on the coordinates in $\mathcal{I}$ satisfies $Y(t') \geq b$. The following theorem is a consequence of a recent result of Talagrand [67] (see also Chapter 2 of [34]).

Theorem 2.3. *Assume that Y is r-Lipschitz, f-certifiable and let m be the median of Y. Then for any $T > 0$,*

$$Pr(Y \leq m - T) \leq 2e^{-T^2/4r^2f(m)},$$

and

$$Pr(Y \geq m + T) \leq 2e^{-T^2/4r^2f(m+T)}.$$

Talagrand's inequality provides a strong concentration result around m, the median of Y. In practice, we usually can, for the sake of convenience, think of m as the mean of Y. (One can make this rigorous by showing that once Y is strongly concentrated around its median, then its mean and its median are more or less the same.)

Azuma and Talagrand's inequalities are, perhaps, the most popular concentration results in combinatorics and theoretical computer science. Several beautiful applications can be found in various surveys and books [8, 50, 34, 64].

2.2. Difficulties with Large Lipschitz Coefficients

Once the Lipschitz coefficient r is small, Azuma's and Talgrand's inequalities are perfect tools. However, these inequalities become less effective as r increases. Consider, for instance, the typical case when the tail is $O(\mathbb{E}(Y))$. In this case, Theorem 2.2 fails to give a nontrivial bound if $r \gg \sqrt{\mathbb{E}(Y)}$ and Theorem 2.3 fails if $r \gg \mathbb{E}(Y)$. Unfortunately, functions with large Lipschitz coefficients do emerge in several natural problems. A typical example is the following well-known problem from the theory of random graphs.

Example. A random graph $G(N, p)$ on N vertices $1, 2, \ldots, N$ is defined by drawing an edge between each pair (i, j), $1 \leq i < j \leq N$, with probability p, independently. Here typically p can be a function in N. There are $n = \binom{N}{2}$ i.i.d. random variables t_{ij}, representing the choices; $t_{ij} = 1$ if the edge (i, j) is drawn and 0 otherwise. Random graphs are basic objects in combinatorics and for more information about the model, we

refer to two excellent books, by Bollobás [11] and Janson, Łuczak, and Ruciński [34], respectively.

We say that three vertices i, j, l form a triangle if there is an edge between any pair of them. Denote by Y the number of triangles in $G(N, p)$. We would like to have an exponential bound on the following probability

$$Pr(|Y - \mathbb{E}(Y)| \geq \epsilon \mathbb{E}(Y)), \tag{2}$$

where ϵ is a fixed positive constant and p is small.

Assume, for instance, that $p = \Theta(N^{-3/4})$. It is clear that in this range $\mathbb{E}(Y) = \binom{N}{3} p^3 = \Theta(N^{3/4})$. Moreover, T is chosen to be $\epsilon \mathbb{E}(Y) = \Theta(N^{3/4})$. Since an edge can be included in $N - 2$ triangles, deleting one edge can change Y by as much as $N - 2$. Thus, the Lipschitz coefficient of Y is at least $N - 2$ (in fact, it is equal to $N - 2$), which is much larger than both $\mathbb{E}(Y)$ and T. So neither Theorem 2.2 nor Theorem 2.3 yield a nontrivial bound.

It is clear that triangles do not play an essential role in the problem and the same difficulty occurs for any fixed graph. Although the question of estimating the probability in (2) has long been studied (see [34], Chapters 2, 3 and their references), due to the lack of knowledge about concentration of functions with large Lipschitz coefficients, no general exponential bounds (which works for an arbitrary fixed graph) were known prior to our study (see Section 6 for more details).

3. OUR MAIN IDEAS AND TOOLS

3.1. Intuition and Main Ideas

Let us take another look at the previous example. The Lipschitz coefficient of Y is indeed $N - 2$ and so Y is not at all smooth in the traditional sense. However, it is a very rare event that any particular edge spans $N - 2$ triangles. For an fixed edge e, the expectation of the number of triangles it spans is only $(N - 2)p^2 = O(1)$. This suggests that in order to obtain a strong concentration result, instead of the "worst case" Lipschitz coefficient, we might want to look at a sort of "average" or "typical" Lipschitz coefficient. In term of smoothness, it means we might want to consider some sort of "average smoothness," rather than the global smoothness defined in Section 2.1. A question of great importance is to find a reasonable definition of average smoothness such that for a large class of functions the following variant of (1) hold

If Y is smooth in average, then it is strongly concentrated. (3)

The attempt to understand the concept of average smoothness has been the driving force behind our development. At the highest level of generality, it seems that finding a proper definition for average smoothness might be impossible, or as hard as proving the concentration itself. On the other hand, we have discovered that for an important class of functions there is in fact a simple way to define average smoothness.

Another idea which underlines our study is to exploit the structural properties of the objective function Y. The proofs of both Azuma's and Talagrand's inequalities are based on an inductive argument on n, the number of atom variables (or the number of martingale

differences). These proofs are very general and require only a little knowledge about the objective function. On the other hand, in discrete problems, functions typically have a special structure which we may take advantage of. For instance, the structure may allow us to apply an extra inductive argument on another parameter.

To conclude this subsection, let us summarize the two leading ideas in our investigation

- Use "average smoothness" instead of "global smoothness."
- Exploit the structural properties of Y by, for instance, an extra induction.

3.2. A General Scheme

In order to use the above two ideas for certain class of functions, we have developed a proving scheme which basically consists of two steps. The first step is general and can be applied to any function. The heart of this step is a martingale lemma, which seems very useful and would be of independent interest. The second step works for functions with special structural properties such as polynomials. The special structure of the function provides the ground for the above-mentioned induction.

The first step: Approximation. Consider a function Y. We call a point $t \in \Omega$ *good,* if changing any coordinate of t does not (in a certain sense) influence Y by too much. All other points are *bad.* In this step, we want to approximate Y by a function Y' such that

(*) $\mathbb{E}(Y') = \mathbb{E}(Y)$,
(**) Y' is strongly concentrated,
(***) $Y(t) = Y'(t)$ for any good point t.

The idea of cutting out the bad domain (consisting of bad points with large Lipshitz coefficient) is fairly natural and was considered in various contexts by many authors, including Spencer-Shamir, Kahn, Kim, Grabble, Godbole, Kersten and others. The extra and critical point in our argument is that the set of bad points was defined in a delicate way that not only guarantees the strong concentration of Y', but also enables us to apply induction in the second step. It appears very convenient in the proof that we have $\mathbb{E}(Y) = \mathbb{E}(Y')$. In practice, it may be enough to guarantee that $|\mathbb{E}(Y) - \mathbb{E}(Y')|$ is considerably smaller than the deviation tail of Y.

Central to this step is a new martingale lemma discussed in Section 3.4 which (among others) provides the definition of bad points.

The second step: Induction. In this step, we need to make use of the structure of Y. Assume, for example, that Y is a polynomial of degree k. Once we have obtained the desired approximating function Y', we continue as follows. Due to the fact that Y and Y' have the same expectation

$$Pr(|Y - \mathbb{E}(Y)| \geq T) \leq Pr(|Y' - \mathbb{E}(Y')| \geq T) + Pr(Y \neq Y'). \tag{4}$$

By (**), Y' is already strongly concentrated, so we only need to bound $Pr(Y \neq Y')$. To do this, we make use of assumption (***), and simply bound the measure of the bad set. This set was defined with foresight in the previous step so that we can bound its

measure by a sum of large deviation probabilities involving other polynomials. These polynomials arise from the first-order partial derivatives of Y and will have degree at most $k - 1$. This crucial fact allows us to use induction on the degree to bound the large deviation probabilities in concern. The appearance of the partial derivatives should not be a big surprise, as the Lipschitz coefficient of Y can also be seen as the maximum value of a first order partial derivative.

Central to the whole scheme is the induction hypothesis, or the formulation of the theorem itself. Here we have discovered an amazing and essential fact that for certain class of functions, it is indeed possible to formalize a hypothesis using an "average Lipschitz coefficient," instead of the global Lipschitz coefficient. Finding this hypothesis was actually the *hardest* part of our work, requiring some insight and a numerous number of attempts. As the reader will see after reading Section 5, a proper induction hypothesis not only allows us to derive a powerful result, but also reduces the proof to a rather routine verification of few conditions.

3.3. The Bad Set and How To Bound It

The definition of bad points requires some other definitions. First, for every $x \in [0, 1]$, $1 \le i \le n$, and $t = (t_1, \ldots, t_n) \in \Omega$ define

$$C_i(x, t) = \left|\mathbb{E}(Y|t_1, \ldots, t_{i-1}, t_i = x) - \mathbb{E}(Y|t_1, \ldots, t_{i-1})\right|,$$

and $C(t) = \max_{i,x} C_i(x, t)$. Intuitively, $C(t)$ is the maximum effect of one coordinate. The crucial point here is that we do not consider a global bound on $C(t)$ (as done in the bounded martingale method), but think of $C(t)$ as a function of t.

To proceed, let us realize that the atom variables t_i's are not necessarily identically distributed, so they may generalize different measures. In order to avoid confusion, we use $\int_I f(x)\, d^i x$ to denote the integral of the function $f(x)$ over the unit interval I with the measure generalized by t_i

$$\int_I f(x)\, d^i x = \int_I f(t_i)\, dt_i.$$

Next, we set

$$V_i(t) = \int_I C_i^2(x, t)\, d^i x \qquad \text{and} \qquad V(t) = \sum_{i=1}^{n} V_i(t).$$

One can view $V(t)$ as a sort of a variance function of Y. It is useful to observe that $C_i(x, t)$ depends on x and the first $i - 1$ coordinates of t. In particular, $C_1(x, t)$ depends only on x. Similarly, $V_i(t)$ depends only on the first $i - 1$ coordinates of t and so $V_1(t)$ is a constant. Finally, we set

$$C_Y = \max_{1 \le i \le n; x \in I; t \in \Omega} C_i(x, t) \qquad \text{and} \qquad V_Y = \max_{t \in \Omega} V(t).$$

Remark that the parameters just defined are invariant under shifting (for instance, $V_Y = V_{Y+s}$ for any number s). This allows us to assume later on, without loss of generality, that $\mathbb{E}(Y) = 0$.

The leading observation in this and the following subsection is that if C_Y and V_Y are relatively small, then Y is strongly concentrated. This is basically the content of our martingale lemma and also provides the reader a way to reason the definition of the bad set, which now follows.

For two arbitrary positive numbers $\mathbf{C}$ and $\mathbf{V}$, the set $\mathbb{B}$ of bad points (with respect to $\mathbf{C}$ and $\mathbf{V}$) is

$$\mathbb{B} = \{t | C(t) \geq \mathbf{C} \quad \text{or} \quad V(t) \geq \mathbf{V}\}.$$

For any $t \in \mathbb{B}$, let $i(t)$ be the smallest index i (between 1 and n) such that either $C_i(x, t) \geq \mathbf{C}$ for some x or $\sum_{j=1}^{i} V_j(t) \geq \mathbf{V}$. Let $\mathbb{B}_t = \{ z \in \Omega \mid z_i = t_i \text{ for all } i < i(t)\}$. It is clear that

- $\mathbb{B}_t \subset \mathbb{B}$.
- For any $t, t' \in \mathbb{B}$, $\mathbb{B}_t$ and $\mathbb{B}_{t'}$ are either identical or disjoint.

It follows that $\mathbb{B}$ is a disjoint union of subhypercubes. Now define the approximating function Y' as follows

- $Y'(z) = Y(z)$ if $z \notin \mathbb{B}$.
- $Y'(z) = \mathbb{E}_{\mathbb{B}_t}(Y)$ if $z \in \mathbb{B}_t \subset \mathbb{B}$.

The following properties are immediate:

$$C_{Y'} \leq \mathbf{C},$$

$$V_{Y'} \leq \mathbf{V},$$

$$Pr(Y \neq Y') \leq Pr(\mathbb{B}).$$

We next show that $\mathbb{E}(Y') = \mathbb{E}(Y)$. This is trivial by definition in the case the atom variables t_i's have discrete distributions. In the general case, it needs a little proof. For all $i = 1, 2, \ldots, n$, let T_i be the set of all bad points t, where $i(t) = i$. Each T_i is measurable and is a disjoint union of hypercubes of dimension $n - i + 1$. Moreover, the projection of T_i onto the hyperplane spanned by the first $i - 1$ coordinates is measurable and therefore Fubini's theorem implies

$$\int_{T_i} Y = \int_{T_i} Y',$$

concluding the proof.

Now we discuss the important question of how to bound $Pr(\mathbb{B})$. One of the simplest, but most useful, ways is the following. First observe that $V(t) \leq C_Y \sum_{i=1}^{n} \int C_i(x, t)\, d^i x$; moreover, $C_Y \geq \mathbf{C}$ if and only if there is a tuple i, x, t such that $C_i(x, t) \geq \mathbf{C}$. Therefore,

$$Pr(\mathbb{B}) \leq Pr\left(\sum_{i=1}^{n} \int C_i(x,\, t)\, d^i x \geq \mathbf{V}/\mathbf{C} \right) + \sum_{i=1}^{n} Pr(C_i(x,\, t) \geq \mathbf{C} \quad \text{for some } x). \tag{5}$$

Next, we find functions $W_i(t)$ and $W(t)$ such that $C_i(x,\, t) \leq W_i(t)$ for all x, i and t and $\sum_{i=1}^{n} \int C_i(x,\, t)\, d^i x \leq W(t)$ for all t. Trivially

$$Pr(\mathbb{B}) \leq Pr(W(t) \geq \mathbf{V}/\mathbf{C}) + \sum_{i=1}^{n} Pr(W_i(t) \geq \mathbf{C}). \tag{6}$$

To make ground for an inductive argument, we next find an index function Ind such that with a proper choice of W_i's and W, Ind(W_i)'s and Ind(W) are smaller than Ind(Y). If this is possible, we can bound the right-hand side in (6) using induction on Ind.

The most convenient situation is when Y is a polynomial. In this case, we can choose W_i to be (roughly) the first-order partial derivative of Y with respect to t_i and so the degree of Y serves as a natural index. Bounding $Pr(W(t) \geq \mathbf{V}/\mathbf{C})$ is usually a simple matter as W can often be defined as a (sort of) weighted sum of the W_i's. In general, when Y is defined based on a underlying hypergraph (see Sections 4.1 and 4.5), the maximum size of a hyperedge is a natural choice for the index.

There are, of course, lots of ways to improve or modify the above argument. For instance, if t_i is concentrated on a few points, we can define, for each x in the support of t_i, a function $W_{i,x}(t)$ such that $W_{i,x}(t) \geq C_i(x,\, t)$ for all t. This way, we obtain

$$Pr(W_i(t) \geq \mathbf{C}) \leq \sum_{x} Pr(W_{i,x}(t) \geq \mathbf{C}). \tag{7}$$

We expect that several modifications like this, which is tailored to a specific situation, will naturally arise with the emergence of new problems. To complete this subsection, let us mention one important and interesting remark. For the same function, there might exist several different indices and the quality of the concentration bound depends on the index we use. Typically, a smaller index leads to a better bound. For certain polynomials, there is, in fact, an index which serves better than the degree (see Section 6.6).

3.4. A Martingale Lemma

The following lemma shows that Y' is strongly concentrated (with respect to $\mathbf{C}$ and $\mathbf{V}$).

Lemma 3.1. *Let X be a real-value function on Ω. Assume that $\mathbf{V}$, $\mathbf{C}$, λ are positive number satisfying $C_X \leq \mathbf{C}$, $V_X \leq \mathbf{V}$, and $\lambda \leq \frac{4V}{C^2}$. Then*

$$Pr(|X - \mathbb{E}(X)| \geq \sqrt{\lambda \mathbf{V}}) \leq 2e^{-\lambda/4}.$$

Consequently, for any function Y and positive numbers $\mathbf{V}$, $\mathbf{C}$, λ such that $\lambda \leq \frac{4V}{C^2}$

$$Pr(|Y - \mathbb{E}(Y)| \geq \sqrt{\lambda \mathbf{V}}) \leq 2e^{-\lambda/4} + Pr(\mathbb{B}),$$

where the set $\mathbb{B}$ is defined with respect to $\mathbf{C}$ *and* $\mathbf{V}$ *as above.*

Proof of Lemma 3.1. We can assume, without loss of generality that $\mathbb{E}(X) = 0$. So it suffices to prove that

$$Pr(|X| \geq \sqrt{\lambda \mathbf{V}}) \leq 2e^{-\lambda/4}.$$

Lemma 3.2. *Let Z be a function form Ω to $\mathbb{R}$ with mean 0. If $u \leq 1/C_Z$, then*

$$\mathbb{E}(e^{uZ}) \leq e^{u^2 V_Z}.$$

To see that Lemma 3.2 implies (8), set $u = \sqrt{\frac{\lambda}{4\mathbf{V}}}$. Since $\lambda \leq \frac{4\mathbf{V}}{\mathbf{C}^2}$, $u \leq \frac{1}{\mathbf{C}} \leq \frac{1}{C_x}$. Lemma 3.2 yields

$$\mathbb{E}(e^{uX}) \leq e^{u^2 V_X} \leq e^{u^2 \mathbf{V}}.$$

By Markov's inequality

$$Pr(X \geq \sqrt{\lambda \mathbf{V}}) = Pr(e^{uX} \geq e^{u\sqrt{\lambda \mathbf{V}}}) = Pr(e^{uX} > e^{\lambda/2}) \leq e^{u^2 \mathbf{V} - \lambda/2} = e^{-\lambda/4}.$$

Inequality (8) follows by symmetry.

Proof of Lemma 3.2. First we need the following simple statement.

Proposition 3.3. *Let μ be an arbitrary measure on $I = [0, 1]$. Suppose that $f(x)$ is a measurable function on I which has absolute value at most 1 and mean 0 (with respect to μ). Then*

$$\int_I e^{f(x)}\, d\mu \leq e^{\int_I f^2(x) d\mu}.$$

Equality holds if and only if $f \equiv 0$.

Proof of Proposition 3.3. Assume that f is not identically zero. By the Taylor's series expansion

$$\int_I e^{f(x)}\, d\mu = 1 + \int_I f(x)\, d\mu + \frac{1}{2!} \int_I f^2(x)\, d\mu + \cdots$$

Since $\int_I f(x)\, d\mu = 0$ and $|f(x)| \leq 1$, the right-hand side is at most

$$1 + \sum_{i=2}^{\infty} \frac{1}{i!} \int_I f^2(x)\, d\mu < 1 + \int_I f^2(x)\, d\mu < e^{\int_I f^2(x) d\mu}.$$

■

Remark. As $c_0 = \sum_{i=2}^{\infty} \frac{1}{i!} < 1$, we can improve the proposition by replacing $\int_I f^2(x)\, d\mu$ in the exponent by $c_0 \int_I f^2(x)\, d\mu$. This may lead to a constant better than 1/4 in the exponent of the bound in Lemma 3.1. However, from the practical point of view, this does not make a big difference, and we do not try to optimize the constants in this paper.

The proof of Lemma 3.2 uses induction on n. If $n = 1$, then, for all $x \in I$, $C_1(x, t) = |Z(x)|$ and $V(x) = \int_I Z^2(x)\, dx$, where the measure is generalized by the unique variable $t_1 = x$. This yields that $|uZ(x)| \le uC_Z \le 1$ for all x and the statement of the lemma follows from Proposition 3.3 by substituting $f = uZ$.

Now consider a generic n. First notice $C_1(x, t) = |\mathbb{E}(Z|t_1 = x)|$ does not depend on t. Consequently, $V_1(t) = \int_I C_1^2(x, t)\, d^1x$ is a constant and we can set $V_1 = V_1(t)$.

Let Ω' be the $(n - 1)$ dimensional subcube spanned by $t_2, \ldots, t_n$. For each $x \in I$, consider the following function from Ω' to $\mathbb{R}$

$$Z_x(t') = Z(x, t_2, \ldots, t_n) - \mathbb{E}(Z|t_1 = x).$$

By definition, $Z_x(t')$ has mean 0, for any $x \in I$. Moreover, $V_{Z_x} \le V_Z - V_1$. By the induction hypothesis

$$\mathbb{E}_{\Omega'}(e^{uZ_x}) \le e^{u^2(V_{Z_x})} \le e^{u^2(V_Z - V_1)}.$$

On the other hand, by Fubini's theorem

$$\begin{aligned}
\mathbb{E}_{\Omega}(e^{uZ}) &= \mathbb{E}_{\Omega'}\left(\int_I e^{uZ_x} e^{u\mathbb{E}(Z \mid t_1 = x)}\, d^1x \right) \\
&= \int_I e^{u\mathbb{E}(Z \mid t_1 = x)} (\mathbb{E}_{\Omega'} e^{uZ_x})\, d^1x \\
&\le \int_I e^{u\mathbb{E}(Z \mid t_1 = x)} e^{u^2(V_Z - V_1)}\, d^1x \\
&= e^{u^2(V_Z - V_1)} \int_I e^{u\mathbb{E}(Z \mid t_1 = x)}\, d^1x.
\end{aligned}$$

Set $f(x) = u\mathbb{E}(Z|t_1 = x)$. By the assumption on u, $f(x)$ satisfies the conditions of Proposition 3.3, so

$$\int_I e^{u\mathbb{E}(Z|t_1 = x)}\, d^1x = \int_I e^{f(x)}\, d^1x \le e^{\int_I f^2(x)\, d^1x} = e^{u^2 V_1},$$

completing the proof. ∎

4. NEW CONCENTRATION RESULTS

4.1. The First Result

In [40] (see also [8], Chapter 7), J. H. Kim and the present author proved the first result on the concentration of polynomials. This result was originally formulated and proved as a lemma to the solution of a long standing question of Segre in finite geometry [41] (see Section 8.4 for more details). However, as it turned out to be of independent interest, we published it in a separate paper.

The setting of the result is as follows. Consider a hypergraph $H = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of vertices and E is the set of edges. We assume that each edge in E contains at most k vertices. Let t_i, $i \in V$, be mutually independent indicator random variables with expectation p_i. Consider the following polynomial:

$$Y = \sum_{e \in E} w_e \prod_{i \in e} t_i,$$

where w_e are positive coefficients. We allow $e = \emptyset$ in which case $\prod_{i \in e} t_i = 1$ by convention.

H is called the underlying hypergraph of Y, and we shall exploit its structure to extract information about the concentration of Y.

For each vertex set A with at most k elements, we define a polynomial Y_A as follows. For each monomial $\prod_{i \in e} t_i$ in Y with $A \subset e$ replace it by $\prod_{i \in e \setminus A} t_i$ and delete all other monomials. For instance, if $Y = 3t_1t_2 + 4t_1t_4 + 5t_3t_5$ and $A = \{1\}$, then $Y_A = 3t_2 + 4t_4$. If A is the empty set, then $Y_A = Y$.

Now comes the crucial definition. Set $\mathbb{E}_j(Y) = \max_{A, |A| \geq j} \mathbb{E}(Y_A)$, for all $j = 0, 1, \ldots, k$. Here and throughout the paper, $\mathbb{E}_j(Y)$ can be heuristically interpreted as the maximum *average* effect of a group of at least j atom variables on Y. In this setting, $\mathbb{E}_1(Y)$ is our candidate for the "average" Lipschitz coefficient.

Theorem 4.1. *For any k there are positive numbers a_k and b_k depending only on k such that with $\mathbb{E}_j(Y)$ defined as above*

$$Pr(|Y - \mathbb{E}(Y)| \geq a_k \lambda^k \sqrt{\mathbb{E}_0(Y)\mathbb{E}_1(Y)}) \leq b_k e^{-\lambda/4 + (k-1)\log n}. \tag{9}$$

Remark. In the above theorem, one can set $a_k = 8^k k!^{1/2}$ and $b_k = 2e^2$. By increasing a_k, one can replace $\lambda/4$ in the exponent by λ and remove b_k for a better-looking bound (this applies for all later theorems). However, as it is more convenient to prove Theorem 4.1 (and other theorems) in this form, we stay with it.

Assume that k is a constant. From the statement of the theorem and the definition of the $\mathbb{E}_j(Y)$'s, one can conclude that if $\mathbb{E}_0(Y) = \mathbb{E}(Y) \gg \mathbb{E}_1(Y) \log^{2k} n$, i.e., the expectation of Y exceeds the average effect of any group of at most k atom variables by a factor $\omega(\log^{2k} n)$, then Y is strongly concentrated. Indeed, in such a case one can choose $\lambda = \omega(\log n)$ so that the bound $b_k e^{-\lambda + (k-1)\log n}$ is super polynomially small while the tail $a_k \lambda^k \sqrt{\mathbb{E}_0(Y)\mathbb{E}_1(Y)}$ is still negligible compared to $\mathbb{E}(Y)$.

The proof of Theorem 4.1 uses an inductive argument on k, the maximal size of an

edge in the underlying hypergraph H. From this, the reader may gain some intuition about why we need to consider $\mathbb{E}(Y_A)$ for every plausible set A of size at most k.

To show the power of Theorem 4.1, let us apply it to obtain a bound on the probability considered in (2). To start, observe that the number of triangles in the random graph $G(N, p)$ can be written as a polynomial of degree 3 as follows:

$$Y = \sum_{1 \le i < j < l \le N} t_{ij} t_{ji} t_{il},$$

where t_{ij} is the binary variable representing the choice of the edge ij. To this end, we understand that t_{ij} and t_{ji} denote the same random variable.

For $p = \Theta(N^{-3/4})$, the expectation of Y is $\Theta(N^3 p^3) = \Theta(N^{3/4})$. Assume that $A = \{ij\}$; then

$$Y_A = \sum_{l \ne i,j} t_{ji} t_{il}.$$

Trivially,

$$\mathbb{E}(Y_A) = O(Np^2) = o(1).$$

If A has two elements, then Y_A is either 0 or t_{ij}, for some i and j. Finally, if A has three elements, then Y_A is either 0 or 1. It follows that

$$\mathbb{E}_1(Y) = \max_{|A| \ge 1} \mathbb{E}((Y_A)) = 1,$$

and

$$\mathbb{E}_0(Y) = \max_{|A| \ge 0} \mathbb{E}((Y_A)) = \mathbb{E}(Y).$$

Setting $\lambda = cN^{1/8}$, where c is a positive constant chosen such that $a_3 \lambda^3 \sqrt{\mathbb{E}(Y)} = \varepsilon \mathbb{E}(Y)$, Theorem 4.1 yields (recall that $n = \binom{N}{2}$)

$$Pr(|Y - \mathbb{E}(Y)| \ge \varepsilon \mathbb{E}(Y)) \le b_3 e^{-\lambda/4 + 2 \log n} = e^{-\Theta(N^{1/8})}. \tag{10}$$

An alert reader might have recognized that under the current circumstances, Y_A is exactly the partial derivative of Y with respect the t_i, $i \in A$. This is not at all accidental. The Lipschitz coefficient of Y can also be defined as the maximum value of a first partial derivative of Y. Thus, in order to find an "average" Lipschitz coefficient, it is natural to consider the expectation of partial derivatives. It has turned out, interestingly, that to define this average Lipschitz coefficient properly, one has to look at not only the partial derivatives of the first order, but of all orders (see Section 4.6 for a discussion of this point).

Theorem 4.1 has several deep and surprising applications [40, 41, 80]. On the other hand, a reader who is familiar with the theory of concentration might find the term λ^k a little bit ad hoc and might wonder whether it can be improved. A "natural" desire is,

perhaps, to replace λ^k by $\lambda^{1/2}$. In the case this could be done, one would obtain not only a more powerful inequality, but also a far more natural one. Our experience shows that in many cases, the quantity $\mathbb{E}_0(Y)\mathbb{E}_1(Y)$ matches the variance of Y, so an inequality with $\lambda^{1/2}$ (instead of λ^k) would imply that

$$Pr(|Y - \mathbb{E}(Y)| \geq \sqrt{\lambda\ Var(Y)}) \leq e^{-c\lambda},$$

which means that Y has a desirable sub-Gaussian tail distribution (see, for instance, Section 6.4).

It has turned out recently that this $\lambda^{1/2}$ replacement is, to some extent, possible. Moreover, we have also found out that the binary assumption, which plays certain role in the original proof of Theorem 4.1 in [40], is unnecessary. Together, these lead to a strengthened and generalized version of Theorem 4.1, which is the subject of the next subsection.

4.2. A More General Result

In this subsection, we consider independent random variables $t_1, \ldots, t_n$ with arbitrary distributions on the interval [0, 1]. A polynomial Y is *normal* if its coefficients are between 0 and 1. We define two parameters c_k and d_k recursively as follows: $c_1 = 1$, $d_1 = 2$, $c_k = 2k^{1/2}(c_{k-1} + 1)$, $d_k = 2(d_{k-1} + 1)$.

Assume that Y has degree k; for a multiset A of size at most k, $\partial_A Y$ denotes the partial derivative of Y with respect to A. For instance, if $Y = t_1^2 t_2^2 t_3 + t_4^5$ and $A_1 = \{1, 2\}$, $A_2 = \{1, 1, 3\}$, then $\partial_{A_1}(F) = 4t_1 t_2 t_3$ and $\partial_{A_2}(F) = 2t_2^2$, respectively. If the set A is empty, then $\partial_A Y = Y$. Finally, for all $0 \leq j \leq k$, let

$$\mathbb{E}_j(Y) = \max_{|A| \geq j} \mathbb{E}(\partial_A(Y)),$$

$$M_j(Y) = \max_{t \in \Omega, |A| \geq j} \partial_A Y(t).$$

The reader should notice that the above definition of $\mathbb{E}_j(Y)$ is a generalization of the definition given in the previous subsection.

Theorem 4.2. *Let Y be normal polynomial of degree k and assume that $\mathbb{E}_j(Y)$'s and $M_j(Y)$'s are defined as above. For any integer $1 \leq \tilde{k} \leq k$ such that $M_{\tilde{k}}(Y) \leq 1$ and any collection of positive numbers $\mathscr{E}_0 > \mathscr{E}_1 > \cdots > \mathscr{E}_{\tilde{k}} = 1$ and λ satisfying*

- $\mathscr{E}_j \geq \mathbb{E}_j(Y)$, $0 \leq j \leq \tilde{k} - 1$,
- $\mathscr{E}_j / \mathscr{E}_{j+1} \geq \lambda + 4j \log n$, $0 \leq j \leq \tilde{k} - 1$,

the following holds:

$$Pr(|Y - \mathbb{E}(Y)| \geq c_k \sqrt{\lambda \mathscr{E}_0 \mathscr{E}_1}) \leq d_k e^{-\lambda/4}. \tag{11}$$

Notice that an integer $\tilde{k}$ satisfying $M_{\tilde{k}}(Y) \leq 1$ always exists given that Y is normal (for instance, $\tilde{k} = k$ satisfies the requirement).

The proof of Theorem 4.2 again uses an inductive argument on k. However, due to the definition of $\tilde{k}$, when we consider the $\tilde{k}$th step, all functions in concerned are (roughly) the partial derivatives of order $\tilde{k}$ and hence bounded by a constant. Thus the induction hypothesis becomes trivial, and we could end the induction process at this point, reducing the number of steps from k to $\tilde{k}$. This usually results in a stronger bound if $\tilde{k} < k$ and the reconsideration of the triangle counting problem will provide such an example.

There are two natural ways to apply Theorem 4.2. If λ is given, one can find the smallest tail by optimizing $\mathscr{E}_0\mathscr{E}_1$. If the tail is given, one can find the best deviation bound (with respect to our theorem) by optimizing λ.

The reader might notice that there is no restriction on k in Theorems 4.1 and 4.2, so in general k may depend on n. However, due to the term λ^k and the fact that a_k, b_k, c_k, d_k are large (order $e^{k \log k}$), the tails in these theorems are small (compared to the mean) only when k is sufficiently small. Typically, our theorems give a reasonable bound for k up to $O(\log n/\log\log n)$. In all applications considered in this paper, k is a constant.

The power of Theorem 4.2 relies on the flexibility of choosing $\tilde{k}$ and the sequence $\mathscr{E}_j$. This will be made more clear in the next subsection, where a few corollaries are derived. To conclude this subsection, let us reconsider the triangle counting problem (see Sections 2.2 and 4.1) and compare the performance of Theorem 4.2 on this problem with that of Theorem 4.1.

Again let Y denote the number of triangles in $G(N, p)$. By the calculation in the previous subsection, a partial derivative of order at least 2 of Y is either t_{ij} for some i, j or 0 or 1. Therefore,

$$M_2(Y) = \max_{t\in\Omega,|A|\geq 2} \partial_A(Y(t)) = 1.$$

To apply Theorem 4.2, we choose $\tilde{k} = 2$ (thus $\tilde{k} < k = 3$) and set $\mathscr{E}_0 = \mathbb{E}(Y)$, $\mathscr{E}_1 = \sqrt{\mathbb{E}(Y)}$, $\mathscr{E}_2 = 1$, and $\lambda = a\sqrt{\mathbb{E}(Y)}$, where a is a positive constant chosen so that conditions of Theorem 4.2 are met and the tail $c_3\sqrt{\lambda\mathscr{E}_0\mathscr{E}_1}$ is at most $\varepsilon\mathbb{E}(Y)$. Theorem 4.2 then yields

$$Pr(|Y - \mathbb{E}(Y)| \geq \varepsilon\mathbb{E}(Y)) \leq d_3 e^{-\lambda} = e^{-\Theta(N^{3/8})}, \tag{12}$$

which is a significant improvement over (10).

4.3. Corollaries

In this subsection, we derive several corollaries of Theorem 4.2, which are somewhat more handy in applications.

Notice that c_k is $e^{\Theta(k \log k)}$, so if $\lambda \geq ck \log k$, for some sufficiently large constant c (which does not depend on k), then we can get rid of the parameter d_k by slightly increasing c_k. The assumption $\lambda \geq ck \log k$ is immediate in most applications, where k is a constant and $\lambda = \omega(1)$.

Corollary 4.3. *There is a constant c such that for each k there is a number c_k such that the following holds. For any positive numbers $\mathscr{E}_0 > \mathscr{E}_1 > \cdots > \mathscr{E}_{\tilde{k}} = 1$ and $\lambda \geq ck \log k$ satisfying*

- $\mathscr{E}_j \geq \mathbb{E}_j(Y)$, $0 \leq j \leq \tilde{k} - 1$,
- $\mathscr{E}_j/\mathscr{E}_{j+1} \geq \lambda + 4j \log n$, $0 \leq j \leq \tilde{k} - 1$,

one has

$$Pr(|Y - \mathbb{E}(Y)| \geq c_k \sqrt{\lambda \mathscr{E}_0 \mathscr{E}_1}) \leq e^{-\lambda/4}, \tag{13}$$

where $\mathbb{E}_j(Y)$'s are defined as in Theorem 4.2.

In certain applications, one can set the $\mathscr{E}_j$'s so that the product $\mathscr{E}_0\mathscr{E}_1$ is the same (up to a constant factor) as the variance of Y. In such a case, (13) yields that Y has sub-Gaussian tail distribution, namely,

$$Pr(|Y - \mathbb{E}(Y)| \geq \sqrt{\lambda\ Var(Y)}) \leq e^{-\Omega(\lambda)}, \tag{14}$$

in some range of λ.

In a general situation, we may not assume that Y is normal, and so the index $\tilde{k}$ with the property that $M_j(Y) \leq 1$ for all $j \geq \tilde{k}$ may not exist. One can easily overcome this obstacle by scaling down Y in a natural way.

Corollary 4.4. *For any positive numbers $\mathscr{F}_0 > \mathscr{F}_1 > \cdots > \mathscr{F}_k$ and λ satisfying*

- *for all* $0 \leq j \leq k$, $\mathscr{F}_j \geq \mathbb{E}_j(Y)$,
- *for all* $0 \leq j \leq k - 1$, $\mathscr{F}_j/\mathscr{F}_{j+1} \geq \lambda + 4j \log n$,

the following holds:

$$Pr(|Y - \mathbb{E}(Y)| \geq c_k \sqrt{\lambda \mathscr{F}_0 \mathscr{F}_1}) \leq d_k e^{-\lambda/4},$$

where $\mathbb{E}_j(Y)$'s are defined as in Theorem 4.2.

Proof. We have, by definition, that

$$M_k(Y) = \mathbb{E}_k(Y) \leq \mathscr{F}_k.$$

Set $\tilde{k} = k$, $\mathscr{E}_j = \mathscr{F}_j/\mathscr{F}_k$ for all j and define $Z = Y/\mathscr{F}_k$. Apply Theorem 4.2 for Z and the sequence $\mathscr{E}_0, \ldots, \mathscr{E}_k$. ∎

To conclude, let us derive Theorem 4.1 from Corollary 4.4, using a particular setting of the parameters $\mathscr{F}_j$'s. Recall that $\mathbb{E}_1(Y) = \max_{j \geq 1} \mathbb{E}_j(Y)$. Set $\mathscr{F}_k = \mathbb{E}_1(Y)$, $\mathscr{F}_i = \mathscr{F}_{i+1}(\lambda + 4i \log n)$ for all $i = 1, \ldots, k - 1$ and $\mathscr{F}_0 = \mathbb{E}_0(Y) \prod_{i=0}^{k-1} (\lambda + 4i \log n)$, we have

$$\mathscr{F}_0\mathscr{F}_1\lambda = \lambda \mathbb{E}_0(Y)\mathbb{E}_1(Y) \prod_{i=0}^{k-1} (\lambda + 4i \log n) \prod_{i=1}^{k-1} (\lambda + 4i \log n).$$

As $(\lambda + 4i \log n)(\lambda + 4(k - 1 - i)\log n) \leq (\lambda + 2(k-1)\log n)^2$, it follows that

$$\sqrt{\mathcal{F}_0\mathcal{F}_1\lambda} \leq \sqrt{[\lambda + 2(k-1)\log n]^{2k}\mathbb{E}_0(Y)\mathbb{E}_1(Y)}.$$

Corollary 4.4 implies

$$Pr(|Y - \mathbb{E}(Y)| \geq c_k[\lambda + 2(k-1)\log n]^k \sqrt{\mathbb{E}_0(Y)\mathbb{E}_1(Y)}) \leq d_k e^{-\lambda/4}, \tag{15}$$

which is equivalent to (9). It should be clear that the previous choice of $\mathcal{F}_j$'s is far from optimal. In most cases (e.g., the triangle counting problem), Theorem 4.2 or Corollaries 4.3 and 4.4 give stronger bounds than Theorem 4.1.

4.4. Concentration of Polynomials with Small Expectations

In this subsection, we consider polynomials with rather small expectations (of order $polylog(n)$, say). In this case, Theorem 4.2 and its corollaries are not very effective. Let us assume, for instance, that $\mathbb{E}(Y) = O(\log n)$ and the parameter $\tilde{k}$ in Theorem 4.2 is at least 2. To satisfy the lower bound of the ratios $\mathcal{E}_j/\mathcal{E}_{j+1}$, $\mathcal{E}_0$ should be $\Omega(\log^2 n)$ and $\mathcal{E}_1$ should be $\Omega(\log n)$. Thus, the tail in Theorem 4.2 is $\Omega(\log^{3/2} n)$. Although Theorem 4.2 still gives something, it is not particularly useful because we usually require the tail to be smaller than the mean of Y. It should be obvious that the situation is even worse with Theorem 4.1, due to the term λ^k.

In a recent paper [78], the present author proved several results in order to handle this situation, and in the rest of this subsection we shall describe these results. To this end, we shall consider only binary random variables. Every monomial of binary random variables can be reduced to a product of different variables (for instance, $t_1^2 t_2^3 = t_1 t_2^2 = t_1 t_2$). We say that a polynomial is *simplified* if its monomials are reduced. It is obvious that every polynomial has a unique simplification.

Define a function f as follows: $f(K) = \max\{1, \lfloor (K/k!)^{1/k} \rfloor - 1\}$. Furthermore, let

$$r(k, K, n, \delta) = \left(\log_2 \frac{1}{\delta}\right) \frac{n^k \delta^{f(K/2)/2}}{f(K/2)!} + (\delta/K^8)^{\lfloor 1/8k \log 1/\delta \rfloor}.$$

It is useful to notice that if δ is a negative power of n (with constant degree), then for any fixed k and β, there is a number $K(k, \beta)$ such that for all $K \geq K(k, \beta)$, $r(k, K, n, \delta) \leq n^{-\beta}$.

Given a normal polynomial Y, define $h(k, K, n, \delta)$ recursively in the following way:

$$h(1, K, n, \delta) = 0;\ h(k, K, n, \delta) = h(k-1, K, n + \lceil \mathbb{E}(Y) \rceil, \delta) + nr(k, K, n, \delta).$$

Given a set A, we denote by $\partial_A^* Y$ the polynomial obtained from the partial derivative $\partial_A Y$ by subtracting its constant coefficient. Define $\mathbb{E}_j^*(Y) = \max_{|A| \geq j} \mathbb{E}(\partial_A^* Y)$. From the definitions, it is clear that if Y has degree k, then $\mathbb{E}_k^*(Y) = 0$.

Theorem 4.5. *Let Y be a simplified normal polynomial of degree k. Suppose that there are positive numbers δ and K satisfying $K \geq 2k$, $\mathbb{E}_1^*(Y) \leq \delta \leq 1$, and $4kK\lambda \leq \mathbb{E}(Y)$.*

Then

$$Pr(|Y - \mathbb{E}(Y)| \geq 2\sqrt{\lambda k K \mathbb{E}(Y)}) \leq 2ke^{-\lambda/4} + h(k, K, n, \delta).$$

The following two corollaries of Theorem 4.5 are easier to apply when $\mathbb{E}(Y)$ falls into certain range. Notice that if Y is homogeneous (that is, every monomial has the same degree), then, for any set A with cardinality at most $k - 1$, $\partial_A^* Y = \partial_A Y$.

Corollary 4.6. *For any positive constants k, α, β, ε there is a constant $Q = Q(k, \varepsilon, \alpha, \beta)$ such that the following holds. If Y is a normal positive homogeneous polynomial of degree k, $n/Q \geq \mathbb{E}(Y) \geq Q \log n$ and $\mathbb{E}(\partial_A(Y)) \leq n^{-\alpha}$ for every nonempty set A of cardinality at most $k - 1$, then*

$$Pr(|Y - \mathbb{E}(Y)| \geq \varepsilon \mathbb{E}(Y)) \leq n^{-\beta}.$$

Corollary 4.7. *Assume that Y is a normal positive homogeneous polynomial of degree k and the expectation of Y is $g \log n$, where $0 < g \leq 1$ can be a function depending on n. Assume, furthermore, that, for all A, $1 \leq |A| \leq k - 1$, $\mathbb{E}(\partial_A(Y)) \leq n^{-\alpha}$ for some positive constant α. Then there are positive constants $c = c(\alpha, k)$ and $d = d(\alpha, k)$ such that for any $0 \leq \varepsilon \leq 1$,*

$$Pr(|Y - \mathbb{E}(Y)| \geq \varepsilon \mathbb{E}(Y)) \leq de^{-c\varepsilon^2 \mathbb{E}(Y)}.$$

In Corollaries 4.6 and 4.7, Y does not need to be simplified and this proves convenient in certain applications (see Section 7). The reader can try to deduce both corollaries from Theorem 4.5 as an exercise or check [78] for a proof.

The proof of Theorem 4.5 is actually more complicated than that of Theorems 4.1 and 4.2. This proof combined our general method (see Section 3) with the following result.

Theorem 4.8. *Let δ be a positive number at most 1. Suppose that Y is normal and $\mathbb{E}_0^*(Y) \leq \delta$. Then for any $K > 0$*

$$Pr(Y \geq K) \leq 2\frac{b(k, n)\delta^{f(K/2)/2}}{f(K/2)!} + (\delta^{1/8}/K)^{\lfloor 1/8k \log 1/\delta \rfloor},$$

where $b(k, n) = \sum_{i=1}^{k} \binom{n}{i}$.

Here is the reason why we need this theorem. When Y has small expectation, we usually have to set the quantity $\mathbf{C}$ (see Section 3.3) be a constant. So, following (6), we need to bound $Pr(W_i \geq \mathbf{C})$, where W_i is a polynomial and $\mathbf{C}$ is a constant. Unfortunately, for such a small tail, our general inductive argument does not work anymore and we need to prove Theorem 4.8 using different arguments. These arguments are purely combinatorial and strongly require the binary assumption on the atom variables. It remains an interesting question whether Theorems 4.5 and 4.8 can be extended to atom variables with arbitrary distributions in the unit interval.

Theorem 4.8 yields the following corollary.

Corollary 4.9. *For any positive constants α and β and a positive integer k, there is a*

positive constant $K = K(k, \alpha, \beta)$ *such that if* Y *is normal of degree* k *and* $\mathbb{E}_0^*(Y) \leq n^{-\alpha}$, *then* $Pr(Y \geq K) \leq n^{-\beta}$.

To conclude this subsection, let us mention a result of a little bit different flavor. This result was needed in the proof of Theorem 4.8 and proved useful in several other situations (see [8] or Section 7, for instance). The proof is simple and can be found in Chapter 8 of [8].

Consider a polynomial Y which is a sum of different monomials with coefficient 1. Two monomials of Y are disjoint if their supports are disjoint. At a point t, we say that a monomial is *alive* if it is not zero. We denote by $Disj(Y(t))$ the maximum number of pairwise disjoint alive monomials of Y at t.

Proposition 4.10. *For* Y *as above and any positive integer* K

$$Pr(Disj(Y) \geq K) \leq \mathbb{E}(Y)^K/K!.$$

4.5. Concentration of More General Functions

While our results focus on polynomials, it is worth mentioning that our method is not restricted to these functions. In fact, the proofs for Theorems 4.1 and 4.2 use the maximum degree of the monomials as an index function, but never use the fact that these monomials are products of the atom variables in an essential way. It is thus very easy to modify these proofs to obtain results for a larger class of functions and below we provide two examples. These examples are, by no mean, exclusive, and we believe that the emergence of new applications will naturally lead to many other extensions.

The writing of this subsection was inspired by a recent paper of S. Janson and A. Ruciński [34], who proved several bounds for the upper tail probability $Pr(Y \geq \mathbb{E}(Y) + T)$ under various general assumptions, using the so-called deletion lemma combined with an inductive argument similar to ours. Their bounds are more or less equivalent to ours in all applications found so far, perhaps due to the similarity between the inductions.

In the first example, we consider a more general setting of Theorem 4.1. Again we have an underlying hypergraph H whose edges are of size at most k. On each edge $e = \{i_1, \ldots, i_l\}$, consider a nonnegative, monotone increasing (in every variables) function Z_e which depends only on the random variables $t_{i_1}, \ldots, t_{i_l}$. (If Z_e is a multiple of the product of these random variables, then we obtain a polynomial.) Our objective function Y is the sum of the Z_e's.

Next, we slightly modify the definition of $\mathbb{E}_j(Y)$'s. It will be useful to keep in mind that we want $\mathbb{E}_j(Y)$ to be the maximum average effect of a group of at least j variables. Similarly to Section 4.1, for each vertex set A of at most k elements, we define a function Y_A as follows. In all Z_e, where e contains A, we substitute 1 for every random variable t_i where $i \in e$ and let Y_A be the sum of these new functions. (Also notice this is a generalization of the definition of Y_A in Section 4.1.) Because of monotonicity, this assignment makes $\mathbb{E}(Y_A)$ the largest among all possible assignments. Now we can define $\mathbb{E}_j(Y)$ the same way as before: $\mathbb{E}_j(Y) = \max_{A,|A|\geq j}\mathbb{E}(Y_A)$.

We can also take into account the quantity $\tilde{k}$ defined in Section 4.2. Let $M_j = \max_{t,|A|\geq j} Y_A(t)$ and assume that there is an index $1 \leq \tilde{k} \leq k$ such that $M_{\tilde{k}}(Y) \leq 1$.

The following theorem can be proved using essentially the proof of Theorem 4.2.

Theorem 4.11. *With $\mathbb{E}_j(Y)$'s defined as above and the numbers c_k and d_k as in Theorem 4.2 the following holds. For any collection of positive numbers $\mathscr{E}_0 > \mathscr{E}_1 > \cdots > \mathscr{E}_{\tilde{k}} = 1$ and λ satisfying*

- $\mathscr{E}_j \geq \mathbb{E}_j(Y)$, $0 \leq j \leq \tilde{k} - 1$,
- $\mathscr{E}_j/\mathscr{E}_{j+1} \geq \lambda + 4j \log n$, $0 \leq j \leq \tilde{k} - 1$,

we have

$$Pr(|Y - \mathbb{E}(Y)| \geq c_k\sqrt{\lambda \mathscr{E}_0 \mathscr{E}_1}) \leq d_k e^{-\lambda/4}. \tag{16}$$

Remark. Theorem 4.11 still holds (with a modification on the definition of $\mathbb{E}_j$) if we only assume that the Z_e's are monotone, but not necessarily increasing in each coordinate. (See Section 5.3 for more about this remark.)

In the second example, we show that one can even give up monotonicity, provided that each atom variable t_i has discrete support consisting of not too many points. This later assumption is satisfied in most combinatorial applications, where the t_i's are typically binary.

Assume that the support of t_i contains at most L points, for all i. We again consider a hypergraph H as before but now Z_e can be an arbitrary nonnegative function depending on the variables with indices in e. To define the quantities $\mathbb{E}_j(Y)$'s, we now need to be a little bit more technical. Since the functions Z_e's are no longer monotone, we cannot uniformly set the value of the random variables in A to one as before, but rather have to consider all possible assignments. Technically speaking, for a set $A = \{i_1, \ldots, i_l\}$ and a vector $x = (x_1, \ldots, x_l)$, where x_j is a vertex in the support of t_{i_j}, we first need to define $Y_{A,x}$ by setting the value of t_{i_j} to x_j and summing over all Z_e where $A \subset e$. Next, we define $\mathbb{E}_j(Y) = \max_{A,x;|A|\geq j}\mathbb{E}(Y_{A,x})$.

Let $M_j = \max_{t,x,|A|\geq j} Y_{A,x}(t)$ and assume that there is an index $1 \leq \tilde{k} \leq k$ such that $M_{\tilde{k}}(Y) \leq 1$.

Theorem 4.12. *With $\mathbb{E}_j(Y)$'s and $M_j(Y)$'s defined as above and the numbers c_k and d_k as in Theorem 4.2 the following holds. For any collection of positive numbers $\mathscr{E}_0 > \mathscr{E}_1 > \cdots > \mathscr{E}_{\tilde{k}} = 1$ and λ satisfying*

- $\mathscr{E}_j \geq \mathbb{E}_j(Y) 0 \leq j \leq \tilde{k} - 1$,
- $\mathscr{E}_j/\mathscr{E}_{j+1} \geq \lambda + 4j \log(nL)$, $0 \leq j \leq \tilde{k} - 1$,

we have

$$Pr(|Y - \mathbb{E}(Y)| \geq c_k\sqrt{\lambda \mathscr{E}_0 \mathscr{E}_1}) \leq d_k e^{-\lambda/4}. \tag{17}$$

The only formal difference between this theorem and Theorems 4.2 and 4.11 is that here we have $\log(nL)$ in the second condition instead of $\log n$. This does not change the bound significantly when $\lambda = \omega(\log n)$ and L is upper bounded by a fixed power of n.

It looks plausible that if we consider only the upper tail or lower tail probability ($Pr(Y \geq \mathbb{E}(Y) + T)$ or $Pr(Y \leq \mathbb{E}(Y) - T)$, respectively, then we may omit the

assumption that the t_i's have discrete supports (and correspondingly L from the theorem). Details will appear elsewhere.

4.6. Partial Derivatives of the First Order Are Not Sufficient

As we have pointed out in Section 4.1, the definition of $\mathbb{E}_j$'s involves partial derivatives of all order. A natural question is whether this is necessary. The following example shows that this is indeed the case.

Assume that n is divisible by 4 and set $l = n/4$. Let $t_1, \ldots, t_n$ by i.i.d. binary random variables with mean $n^{-1/2}$. Consider the following polynomial

$$Y = \left(\sum_{i=1}^{l} t_{2i-1} t_{2i} \right) \left(\sum_{j=2l+1}^{n} t_j \right).$$

A straightforward calculation shows that the expectation of Y is $n^{1/2}/8$ and the expectation of any first-order partial derivative of Y is at most 1/2. On the other hand, Y is not at all concentrated! It is easy to see, using Chernoff's bound, that with probability $1 - o(1)$, $\sum_{j=2l+1}^{n} t_j$ is at least $n^{1/2}/4$. Moreover, $\sum_{i=1}^{l} t_{2i-1} t_{2i}$ is either 0 or at least 1. Therefore, Y is, almost surely, 0 or at least twice its expectation!

Variants of the above example also rule out the possibility of defining average smoothness using partial derivatives up to any fixed order.

4.7. The Polynomial Method

In several applications, the function we are interested in might not be a positive polynomial. In such a case, we can use the following method, which will be later referred to as the polynomial method.

Given a function X from a product space Ω generated by independent variables $x_1, \ldots, x_m$ to $\mathbb{R}$. In order to show that X is strongly concentrated, we first find two polynomials Y_1 and Y_2 of small degree such that

$$Y_1(x) \leq X(x) \leq Y_2(x), \qquad \text{for all } x \in \Omega,$$

$$\mathbb{E}(Y_1) \approx \mathbb{E}(X) \approx \mathbb{E}(Y_2).$$

Once Y_1 and Y_2 are found, all we need is to show that both polynomials are strongly concentrated. Since X is sandwiched between the two, X itself should also be strongly concentrated.

There are several variants of this method; for instance, if we only need a upper tail bound on X, then it is sufficient to find Y_2 such that $X \leq Y_2$ and $\mathbb{E}(X) \approx \mathbb{E}(Y_2)$.

A very convenient way to find polynomial approximation is to truncate the Taylor series of X (if it has one). This approach works remarkably well in several applications (see Section 8.2, for instance).

4.8. Questions

In this subsections we present few questions which we think are important and may stimulate further research in the area.

Let us start with a specific question concerning the general version of the triangle counting problem considered in Section 2.2. For a fixed graph H, let Y_H be the number of copies of H in $G(N, p)$. Let μ denote the expectation of Y_H and let ε be a fixed small positive constant. We want to find a sharp estimate for

$$Pr(|Y_H - \mu| \geq \varepsilon\mu).$$

This problem is interesting by several reasons. First, in itself it is an old and fairly well-known problem in the theory of random graphs. Second, it seems to be a good toy problem for developing new ideas which can be applied to a more general setting (such as polynomials or the settings in Section 4.5). For instance, the study of this problem reveals the possibility of using different indices for the same function (see Section 6.6 and the last paragraph of Section 3.3).

The most recent lower bound and upper bound on the above probability can be found in [76] (see also Section 6.6). Particularly interesting is the lower bound, which involves the fractional independent number of H and thus gives the problem more combinatorial flavor. This bound turns out to be sharp for certain graphs and seems to be a good candidate even for most graphs.

A more general problem is to study the sharpness of our theorems presented in this section. Perhaps a solution to the above specific question will also shed some light on this matter. We know very little about the lower bounds for the probabilities considered in our theorems. The most useful ones are those obtained by using the argument of the lower bound in the subgraph counting problem [76].

The parameters a_k, b_k, c_k, d_k in our results are fairly generous, and we have not made any attempt to optimize them, as these parameters can always be ignored in our applications so far. However, it is still desirable to know how much could one reduce these parameters.

We also find it very interesting to have a more analytical proof for Theorem 4.2, which exploits more the properties of partial derivatives. Such a proof may easily lead to a new method and further developments.

5. PROOFS OF THEOREMS 4.2, 4.11, AND 4.12

In the first two subsections, we prove Theorem 4.2. The proofs of the other two theorems follow fairly easily in the last subsection.

5.1. Choice of W_i and W

We start by defining functions $W_i(t)$'s and $W(t)$ such that $W_i(t) \geq C_i(x, t)$ for all i, x and t and $\sum_{i=1}^{n} \int_I C_i(x, t)\, d^i x \leq W(t)$ for all t, following the plan described in subsection 3.3 [see the paragraph prior to (6)].

Notice that Y is nonnegative and monotone, so

$$C_i(x, t) = |\mathbb{E}(Y|t_1, \ldots, t_{i-1}, t_i = x) - \mathbb{E}(Y|t_1, \ldots, t_{i-1})| \leq \mathbb{E}(Y_i|t_1, \ldots, t_{i-1}, t_i = 1), \tag{18}$$

where Y_i is the sum of all monomials containing t_i (for the other two theorems $Y_i = \sum_{i \in e} Z_e$). Thus we can set $W_i(t) = \mathbb{E}(Y_i | t_1, \ldots, t_{i-1}, t_i = 1)$. Next, notice that

$$\begin{aligned} C_i(x, t) &= |\mathbb{E}(Y | t_1, \ldots, t_{i-1}, t_i = x) - \mathbb{E}(Y | t_1, \ldots, t_{i-1})| \\ &\leq \mathbb{E}(Y_i | t_1, \ldots, t_{i-1}, t_i = x) + \mathbb{E}(Y_i | t_1, \ldots, t_{i-1}), \end{aligned}$$

which yields

$$\begin{aligned} \sum_{i=1}^{n} \int_I C_i(x, t)\, d^i x &\leq \sum_{i=1}^{n} \int_I [\mathbb{E}(Y_i | t_1, \ldots, t_{i-1}, t_i = x) + \mathbb{E}(Y_i | t_1, \ldots, t_{i-1})]\, d^i x \\ &= \sum_{i=1}^{n} (1 + p_i) \mathbb{E}(Y_i | t_1, \ldots, t_{i-1}), \end{aligned}$$

For convenience, we replace $(1 + p_i)$ by 2 and set $W(t) = 2 \sum_{i=1}^{n} \mathbb{E}(Y_i | t_1, \ldots, t_{i-1})$. It is clear, by definition, that the W_i's and W are polynomials with degrees at most $k - 1$. Moreover, it is fairly easy to show

Lemma 5.1. *We have*

(A) $\mathbb{E}_j(W_i) \leq 2\mathbb{E}_{j+1}(Y)$ *for* $j = 0, \ldots, k - 1$ *and any* i,
(B) $M_j(W_i) \leq 2$ *for all* $j \geq \tilde{k} - 1$ *and any* i,
(C) $\mathbb{E}_j(W) \leq 2k\mathbb{E}_j(Y)$ *for* $j = 0, \ldots, k - 1$ *and any* i,
(D) $M_j(W) \leq 2k$ *for all* $j \geq \tilde{k}$ *and any* i.

To give the reader a feeling about this lemma, let us consider the special, but important, case when Y is multilinear. In this case, one can easily verify that $W_i(t)$ can be obtained by substituting all variables t_j $(j > i)$ in the first partial derivative of Y with respect to t_i with their expectations (this is the only place we use the multilinear assumption). It follows immediately that the expectation of any partial derivative of order j of W_i is at most the maximum expectation of a partial derivative of order $j + 1$ of Y. This implies $\mathbb{E}_j(W_i) \leq \mathbb{E}_{j+1}(Y)$, proving (A) (even without the factor 2 on the right-hand side). The proof for (B) is similar. To verify the statement in (C) (for $j = 0$, say), simply notice that each monomial of Y has at most k nonvanishing first-order partial derivatives and hence can appear at most k times in W.

The factor of 2 in (A) and (B) is due to the presence of atom variables with higher degrees when Y is not multilinear. (This constant factor never plays an essential role anyway.) The formal proof of the lemma is given in the appendix at the end of the paper.

5.2. The Completion of the Second Step

We use induction on k to show

$$Pr(|Y - \mathbb{E}(Y)| \geq c_k \sqrt{\lambda \mathscr{E}_0 \mathscr{E}_1}) \leq d_k e^{-\lambda/4}.$$

To start, consider $k = 1$. Since we assume that $\tilde{k} > 0$, $\tilde{k}$ should be 1. Set $\mathbf{C} = \mathscr{E}_1$ and $\mathbf{V} = \mathscr{E}_0\mathscr{E}_1$. By definition, $C_Y \le \mathbf{C}$ and $V_Y \le \mathbf{V}$, and the bad set $\mathbb{B}$ is empty. Moreover, $4\mathbf{V}/\mathbf{C}^2 \ge \lambda$; thus, Lemma 3.1 applies and yields that

$$Pr(|Y - \mathbb{E}(Y)| \ge \sqrt{\lambda\mathscr{E}_0\mathscr{E}_1}) \le 2e^{-\lambda/4},$$

completing the proof.

We now consider a generic $k \ge 1$. Set $\mathbf{C} = 2(c_{k-1} + 1)\mathscr{E}_1$ and $\mathbf{V} = 4k(c_{k-1} + 1)^2\mathscr{E}_0\mathscr{E}_1$. A simple calculation yields that $\mathbf{V}/\mathbf{C} = 2k(c_{k-1} + 1)\mathscr{E}_0$ and $4\mathbf{V}/\mathbf{C}^2 \ge \mathscr{E}_0/\mathscr{E}_1 \ge \lambda$. Define Y' and $\mathbb{B}$ with respect to these parameters. Due to Lemma 3.1 and the recursive definition of d_k, it suffices to show that

$$Pr(\mathbb{B}) \le 2d_{k-1}e^{-\lambda/4}. \tag{19}$$

Taking into account the definitions of $\mathbf{C}$, $\mathbf{V}$, c_k, and d_k, inequality (19) is a straightforward consequence of the following two claims and inequality (6) in Section 3.3. The proofs of both claims require only a formal verification of the conditions of the induction hypothesis.

Claim 5.2. *For all* $i = 1, \ldots, n$

$$Pr(W_i(t) \ge 2\mathscr{E}_1(c_{k-1} + 1)) \le d_{k-1}e^{-\lambda/4 - \log n}.$$

Claim 5.3. *We have*

$$Pr(W(t) \ge 2k(c_{k-1} + 1)\mathscr{E}_0) \le d_{k-1}e^{-\lambda/4}.$$

Proof of Claim 5.2. We distinguish two cases: $\tilde{k} = 1$ and $\tilde{k} > 1$.

If $\tilde{k} = 1$, then by (B) of Lemma 5.1, $W_i(t) \le 2$ for all t. On the other hand, $2k(c_{k-1} + 1)\mathscr{E}_0 > 2$. Thus, the claim is trivial.

Assume that $\tilde{k} > 1$. Let $X_i(t) = W_i(t)/2$, $\lambda' = \lambda + 4 \log n$, $\mathscr{E}'_j = \mathscr{E}_{j+1}$. By (A) and (B) of Lemma 5.1, X_i satisfies the following:

- $\mathbb{E}_j(X_i) \le \mathscr{E}_{j+1} = \mathscr{E}'_j$,
- $M_j(X_i) \le 1$, for all $j \ge \tilde{k} - 1$,
- $\mathscr{E}'_j/\mathscr{E}'_{j+1} \ge \lambda' + 4j \log n$.

Since $X_i(t)$ is a polynomial of degree $k - 1$, the induction hypothesis applies and yields

$$Pr(|X_i - \mathbb{E}(X_i)| > c_{k-1}\sqrt{\lambda'\mathscr{E}'_0\mathscr{E}'_1}) \le d_{k-1}e^{-\lambda'/4} = d_{k-1}e^{-\lambda/4 - \log n}.$$

On the other hand, since $\tilde{k} > 1$, $\mathscr{E}'_0/\mathscr{E}'_1 = \mathscr{E}_1/\mathscr{E}_2 \ge \lambda'$. Thus, the tail is at most $c_{k-1}\mathscr{E}_1$. Because $\mathbb{E}(X_i) \le \mathscr{E}_1$, the claim follows. ∎

Proof of Claim 5.3. Set $X(t) = W(t)/2k$. Since $\tilde{k} > 0$, we have $\mathscr{E}_0/\mathscr{E}_1 \ge \lambda$. By the last two statements of Lemma 5.1, we have

- $\mathbb{E}_j(X) \le \mathscr{E}_j$,
- $M_j(X) \le 1$, for all $j \ge \tilde{k}$.

The induction hypothesis, applied for $X(t)$, yields

$$Pr(|X - \mathbb{E}(X)| \ge c_{k-1}\sqrt{\lambda \mathscr{E}_0 \mathscr{E}_1}) \le d_{k-1}e^{-\lambda/4}.$$

As $\lambda \mathscr{E}_0 \mathscr{E}_1 \le \mathscr{E}_0^2$ we obtain

$$Pr(X \ge \mathbb{E}(X) + c_{k-1}\mathscr{E}_0) \le d_{k-1}e^{-\lambda/4}.$$

Since $\mathbb{E}(X) \le \mathscr{E}_0$ and $X(t) = W(t)/2k$, the claims follows. ∎

5.3. Proofs of Theorems 4.11 and 4.12

Notice that in the previous proof, the only place where properties of polynomials are used is the proof of Lemma 5.1 (in fact, we do not use much about polynomials in this proof, either). On the other hand, this lemma is straightforward in the setting of Theorem 4.11, due to the new definition of $\mathbb{E}_j(Y)$'s. Thus Theorem 4.11 follows without any modification.

To prove Theorem 4.12, we apply inequality (7) from Section 3.3, taking into account the fact that the support of each t_i has at most L elements. The most trivial choice for $W_{i,x}(t)$ is, perhaps, $W_{i,x}(t) = \mathbb{E}(Y_i|t_1, \ldots, t_{i-1}, t_i = x) + \mathbb{E}(Y_i|t_1, \ldots, t_{i-1})$. With this choice of $W_{i,x}(t)$, we can apply the induction hypothesis to bound $Pr(W_{i,x}(t) \ge C)$ for every pair i, x. Compared to the proof of Theorem 4.2, the only technical difference is that now we have (at most) nL functions instead of n functions. Therefore, in Claim 5.2, we need to change $\log n$ to $\log(nL)$. This results in the appearance of the term $\log(nL)$ in the statement of Theorem 4.12. For $W(t)$ we can still use the same definition as in the proof of Theorem 4.2 and Claim 5.3 also remains the same.

Finally, we discuss the remark following Theorem 4.11. If a function Z_e is monotone, but not necessarily increasing in each coordinate, then for each i, we need to consider two functions $W_{i,0} = \mathbb{E}(Y_i|t_1, \ldots, t_{i-1}, t_i = 0)$ and $W_{i,1} = \mathbb{E}(Y_i|t_1, \ldots, t_{i-1}, t_i = 1)$ and use the trick in the proof of Theorem 4.12. Instead of n functions W_i's, we now need to consider $2n$ functions and this only results in a slight change of c_k.

6. APPLICATIONS TO RANDOM GRAPHS

In this section, we describe some applications of our results to the classical subgraph counting problem. Through these applications, we rediscover important notions such as balanced and strictly balanced graphs. Several results are new, and we are not aware of any other method which could produce the same results.

6.1. Introduction

Fix a small graph G with m vertices and k edges, X_G counts the number of subgraphs of $G(N, p)$ isomorphic to G. The ratio k/m is the density of G, and we call G *balanced* if its density is not smaller than that of the subgraphs. If the density of G is larger than the

density of every subgraph, then we call G *strictly balanced.* The study of X_G is a classical topic in the theory of random graphs, and we refer the reader to [11] and [34] for the background.

In this section, we use our concentration bounds to prove several properties of the distribution of X_G. As a warm-up, we give a new proof for a classical theorem of Erdős and Rényi on the threshold probability of X_G for the case G is strictly balanced. Next, in Section 6.3 we prove a new exponential concentration bound on X_G, improving a bound given in [40]. This bound also generalizes the bound shown for the triangle problem in our principal example. In Section 6.4, we show that, for sufficiently large p, X_G has sub-Gaussian tail distribution in a large interval around its mean. It has been known for a long time that the normalized version of X_G tends to the Gaussian distribution in probability, but little has been known about the speed of convergence. In relation to this problem, our last result is quite interesting, since it, in a sense, implies that the normalized version of X_G assumes a sub-Gaussian behavior very fast. Section 6.5 describes various extensions and other problems. Section 6.6 discusses the possibility of choosing different indices which lead to better bounds.

Although the distribution of X_G has been studied intensively for decades, we are not aware of any other method which would provide the results of the last two applications. Another interesting feature is that, in the proofs, the notions of balanced and strictly balanced graphs emerge in a natural way, through the context of polynomials.

To start, notice that X_G can be expressed as a polynomial of degree k as follows

$$X_G = \frac{1}{|Aut(G)|} \sum_{x_1, \ldots, x_m \in [N]} \prod_{i \sim_G j} t_{x_i x_j},$$

where $i \sim_G j$ means i and j are adjacent in G. It is more convenient to work with $Y_G = \Sigma_{x_1, \ldots, x_m \in [N]} \Pi_{i \sim_G j} t_{x_i x_j}$, ignoring the constant factor $|Aut(G)|^{-1}$. It is clear that

$$\mathbb{E}(Y_G) = \Theta(N^m p^k).$$

For any set A of at most k edges, let $v(A)$ denote the number of vertices in A. Furthermore, set $v(j) = \min_{A, |A| \geq j} v(A)$. A simple consideration shows that

$$\mathbb{E}(\partial_A(Y_G)) = \Theta(N^{m - v(A)} p^{k - |A|}), \tag{20}$$

which implies

$$\mathbb{E}_j(Y_G) = \max_{h \geq j} \Theta(N^{m - v(h)} p^{k-h}). \tag{21}$$

(21) yields the following vital fact, which underlines most proofs in this section

$$\mathbb{E}(Y_G)/\mathbb{E}_j(Y_G) = \Theta(\min_{h \geq j}(N^{v(h)} p^h)). \tag{22}$$

Observe that the right-hand side of (22) is (up to a constant factor) exactly the minimum expectation of the number of copies of a subgraph of H with at least j vertices.

6.2. Threshold Probability

In their seminal paper [19], which founded the theory of random graph, Erdős and Rényi proved the following result on the threshold probability of Y_G.

Theorem 6.1. *Assume that G is strictly balanced. Then*

$$Pr(Y_G = 0) = 1 - o(1) \qquad \text{if } p \ll N^{-m/k},$$
$$Pr(Y_G > 0) = 1 - o(1) \qquad \text{if } p \gg N^{-m/k}.$$

The first statement is trivial by a first moment argument and we now give a short proof for the second statement. For $p \gg N^{-m/k}$, $\mathbb{E}(Y_G) = \omega(1)$. Since the probability $Pr(Y_G > 0)$ is increasing in p, we can assume, without loss of generality, that $\mathbb{E}(Y_G) = g(n)\log N$, where $0 < g(n) \leq 1$. We next apply Corollary 4.7. To apply this corollary, we need the following claim.

Claim 6.2. *If G is strictly balanced and* $\mathbb{E}(Y_G) = N^{o(1)}$, *then* $\mathbb{E}(\partial_A(Y_G)) \leq N^{-\alpha}$ *for some constant* α *and all* A, $k - 1 \geq |A| \geq 1$.

Proof of the Claim. By (20)

$$\mathbb{E}(\partial_A(Y_G)) = \Theta(N^{m-v(A)}p^{k-|A|}) = \Theta(\mathbb{E}(Y_G))N^{-v(A)}p^{-|A|}.$$

We need to show that $\mathbb{E}(Y_G))\, N^{-v(A)}p^{-|A|}$ is bounded from above by a negative power of N. Here the strictly balance assumption emerges naturally. For any set A of edges which forms a proper subgraph of G, the density $|A|/v(A)$ is less than k/m, the density of G; therefore, $v(A)/|A| > m/k$. Moreover, since the expectation of Y_G is $N^{o(1)}$, $N^{m/k}p = N^{o(1)}$. Therefore, there is a positive constant α so that $\mathbb{E}(\partial_A(Y_G)) \leq N^{-\alpha}$. ∎

Corollary 4.7 yields

$$Pr(Y_G = 0) \leq Pr(|Y_G - \mathbb{E}(Y_G)| \geq \mathbb{E}(Y_G)) = O(e^{-c\mathbb{E}(Y_G)^2}) = o(1),$$

proving the theorem. ∎

6.3. Exponential Concentration Bounds

Let us reconsider the question posed in our principal example:

Bound

$$Pr(|X_G - \mathbb{E}(X_G)| \geq \varepsilon\mathbb{E}(X_G)).$$

The case when $\mathbb{E}(X_G)$ is small (of order $polylog(N)$) was treated in earlier papers (see [28] and its references). So here we address the case of larger $\mathbb{E}(X_G)$ only. For the sake of simplicity, we think of ε as a small positive constant, although our arguments apply for arbitrary ε.

We would like to apply Theorem 4.2 with $\tilde{k} = k$. Set $\lambda = aN^{v/k}p$ and $\mathscr{E}_j = (b\lambda)^{k-j}$, where a, b are positive constants to be chosen. It is clear that the condition $\mathscr{E}_j/\mathscr{E}_{j+1} \geq$

$\lambda + 4j \log n$ is satisfied with any $b > 1$, since $\lambda \gg \log n$. Moreover, we can set a sufficiently small so that the tail $c_k\sqrt{\lambda \mathscr{E}_0 \mathscr{E}_1}$ is at most $\varepsilon \mathbb{E}(Y_G)$. The only condition we need to check is whether $\mathscr{E}_j \geq \mathbb{E}_j(Y_G)$. Assume that b is sufficiently large, to guarantee that $\mathscr{E}_j \geq \mathbb{E}_j(Y_G)$ it is enough to have

$$(N^{m/k}p)^{k-j} \geq N^{m-v(h)}p^{k-h}, \tag{23}$$

for any $h \geq j$. By finding a necessary and sufficient condition for (23), we will see that the notion of balanced graphs arise in a natural way. In (23), assume that $h = j$, then the inequality becomes

$$(N^{m/k}p)^{k-j} \geq N^{m-v(j)}p^{k-j}, \tag{24}$$

which is satisfied if and only if $m/k \geq (m - v(j))/(k - j)$. The latter is equivalent to $k/m \geq j/v(j)$. This simply means that the density of G should be at least as large as that of any subgraph, that is, G should be balanced.

The condition G is balanced is not only necessary, but also sufficient for (23). Indeed, if G is balanced then for any h, $k/m \geq h/v(h)$ which implies $m/k \geq (m - v(h))/(k - h)$. On the other hand, since $N^{m/k}p \gg 1$ and $k - j \geq k - h$, it follows that

$$(N^{m/k}p)^{k-j} \geq (N^{m/k}p)^{k-h} \geq (N^{(m-v(h))/(k-h)}p)^{k-h} = N^{m-v(h)}p^{k-h}.$$

Thus, Theorem 4.2 implies the following corollary.

Corollary 6.3. *If G is balanced and $\mathbb{E}(Y_G) \gg \log^k N$, then for any positive constant ε, there is a positive constant c such that*

$$Pr(|Y_G - \mathbb{E}(Y_G|) \geq \varepsilon \mathbb{E}(Y_G)) \leq e^{-c\mathbb{E}(Y_G)^{1/k}}.$$

Remarks. For certain G and p, Corollary 6.3 can be furthered improved by choosing a smaller $\tilde{k}$ (see the treatment of triangles in subsection 3.3, where $\tilde{k} = 2$ and $k = 3$). Corollary 6.3 strengthens an earlier result in [40]. Prior to our method, we were not aware of any other general method which could provide bounds of comparable strength for an arbitrary graph G, although there are several ad hoc arguments which work for certain graphs (see [34], Chapter 2).

6.4. Sub-Gaussian Distribution

A significant part of the research on the distribution of Y_G focus on the limit distribution of the normalized version of Y_G. For instance, a typical result is the following, proven by Ruciński [57]:

If $1/2 \geq p \gg n^{-1/m(G)}$, where $m(G)$ is the maximal density of a subgraph of H, then $\overline{Y_G} = \frac{Y_G - \mathbb{E}(Y_G)}{Var^{1/2}(Y_G)}$ tends to the normal distribution $N(0, 1)$ in distribution.

This is more or less equivalent to saying that for any positive constant λ

$$Pr(|Y_G - \mathbb{E}(Y_G)| \geq \sqrt{\lambda\ Var(Y_G)}) \leq \exp(-c\lambda),$$

where one can set $c = 1/2 - o(1)$. A question of great importance is the speed of the convergence and here very little has been known. In particular, it is not clear that whether the bound above holds for λ tending to infinity together with N. Using Theorem 4.2, we could prove that with a different constant c this is the case, provided that p is sufficiently large.

Corollary 6.4. *For any fixed graph G and any $p = N^{-\alpha}$, where $\alpha < 1/m'(G)$, where $m'(G) = \max_{2\leq j\leq k}(j-1)/(v(j)-2)$, there are positive constants c, δ, Δ such that for any $N^{\delta} > \lambda > \Delta$*

$$Pr(|Y_G - \mathbb{E}(Y_G)| \geq \sqrt{\lambda\ Var(Y_G)}) \leq e^{-c\lambda}.$$

Proof. In this range of p, the expectation of every subgraph of G is at least a positive constant power of N. By Claim 6.2, there is a positive constant β such that for any $j \geq 1$

$$\mathbb{E}(Y_G)/\mathbb{E}_j(Y_G) \geq N^{\beta}.$$

For any $j \geq 2$, consider

$$R_j = N^{v(j)}p^j/N^2p = N^{v(j)-2}p^{j-1} = N^{((v(j)-2)/(j-1)-\alpha)(j-1)}.$$

By the condition on α, there is a positive constant γ such that for any $j \neq 1$, R_j is at least N^{γ}. It follows that $\max_{j\geq 1}\mathbb{E}_j(Y_G) = \mathbb{E}_1(Y) = \Theta(\mathbb{E}(Y)/N^2p)$. Now set $\mathscr{C}_0 = a\mathbb{E}(Y_G)$, $\mathscr{C}_j = b(\lambda + 4j\log n)^{1-j}\mathbb{E}(Y_G)/N^2p$ for all $1 \leq j \leq k-1$ and $\mathscr{C}_k = 1$. There is a positive constant δ such that for all $k \log k \ll \lambda \leq N^{\delta}$ one can find positive constants a and b so that the conditions of Corollary 4.3 hold. This corollary implies that for some positive constant c

$$Pr(|Y_G - \mathbb{E}(Y_G)| \geq \sqrt{\lambda\mathscr{C}_0\mathscr{C}_1}) \leq e^{-c\lambda}.$$

It is well known (and easy to verify) that

$$Var(Y_G)) = \Theta(\mathbb{E}(Y_G)\max_{j\geq 1}\mathbb{E}_j(Y_G))$$

(see [11, 34], for instance). In our case, $\max_{j\geq 1}\mathbb{E}_j(Y_G)$ is attained at $j = 1$. Therefore, $Var(Y_G) = \Theta(\mathbb{E}_0\mathbb{E}_1) = \Theta(\mathscr{C}_0\mathscr{C}_1)$, and this completes the proof. ∎

6.5. Variations

Counting Extensions and the Zero–One Law. Let L be a graph with vertices labeled by $r_1, \ldots, r_l, v_1, \ldots, v_m$, where $R = \{r_1, \ldots, r_l\}$ is a special subset, called the roots.

The v_j are *free* points, and an edge with at least one free endpoint is called a *free edge*. The pair (R, L) will be dubbed as a rooted graph. Let H be a graph on $[N]$ and identify R with a set of l points in H (to simplify the notation, we also call these points $r_1, \ldots, r_l$). In a rooted graph we pay no attention to the edges between the roots.

Consider a subgraph L' of H on $\{r_1, \ldots, r_l\} \cup W$, where $|W| = m$. We say that this subgraph is an extension if one can label the vertices of W as $w_1, \ldots, w_m$ so that

- $w_i \sim w_j$ if and only if $v_i \sim v_j$,
- $w_i \sim r_j$ if and only if $v_i \sim r_j$.

In other words, (R, L') is a copy of (R, L). Consider $G(N, p)$; we denote by $Y_{(R,L)}$ the number of extensions corresponding to a given pair (R, L) and a fixed set of vertices $r_1, \ldots, r_l$. If $l = 0$ (i.e., there is no root), then $Y(\emptyset, L) = Y_L$ is the number of copies of L. Therefore, the problem of counting extensions can be seen as a generalization of the problem of counting subgraphs considered in the previous subsection. We denote by $\mathbb{E}(Y_{(R,L)})$ the expectation of $Y_{(R,L)}$ in $G(N, p)$; it is clear that $\mathbb{E}(Y_{(R,L)}) = \Theta(N^m p^k)$, where m is the number of free vertices and k is the number of free edges.

The investigation of the number of rooted subgraphs is motivated by a theorem of Shelah and Spencer on zero–one laws. In [61] Shelah and Spencer proved the following important result

Theorem 6.5. *If $p = N^{-\gamma}$ for γ irrational, then p satisfies zero–one law.*

We omit the (rather involved) definition of zero-one laws and refer to [61]. Consider a rooted graph (R, L) with m free vertices and k edges. The ratio k/m is the *density* of (R, L). Similar to the case of graphs, we say that (R, L) is *balanced* if its density is not smaller than the density of any proper subgraph; if the density of (R, L) is definitely larger than that of any proper subgraph, then we say that (R, L) is *strictly balanced.* We say that p is *safe* if the expectation of $Y_{(R,L)}$ in $G(N, p)$ is lower bounded by a positive constant power of N.

A key tool in the proof of Shelah and Spencer's theorem is a concentration result on the number of rooted subgraphs. This concentration result was later strengthened by Spencer in another paper [63] to the following:

Theorem 6.6. *If p is safe and (R, L) is strictly balanced, then for any positive constant ε*

$$Pr(|Y_{(R,L)} - \mathbb{E}(Y_{(R,L)})| > \varepsilon \mathbb{E}(Y_{(R,L)})) = o(n^{-r}).$$

We use Theorem 4.2 to improve Theorem 6.6 by giving a sharper bound and weakening the assumption that (R, L) is strictly balanced. Notice that if p is safe than $\mathbb{E}(Y_{(R,L)})^{1/k} \gg \log N$. The proof of the following corollary is similar to that of Corollary 6.3 and is omitted.

Corollary 6.7. *If p is safe and (R, L) is balanced, then for any positive constant ε there is a positive constant $c(\varepsilon)$ such that*

$$Pr(|Y_{(R,L)} - \mathbb{E}(Y_{(R,L)})| > \varepsilon \mathbb{E}(Y_{(R,L)})) < \exp(-c(\varepsilon)\mathbb{E}(Y_{(R,L)})^{1/k}).$$

One can also prove an analog of Corollary 6.4, and we leave it to the reader as an exercise.

Counting Induced Subgraphs. Our results can also be used to derive strong concentration results on the number of induced subgraphs. For instance, the number of induced paths of length 2 can be expressed as follows:

$$Y = \sum_{x,y,z} t_{xy}t_{yz}(1 - t_{xz}).$$

The term $1 - t_{xz}$ is needed to guarantee that there is no edge between x and z; in other words, the path is induced. Although Y is not positive, we can write it as the difference of 2 positive polynomials Y_1 and Y_2, where Y_1 has degree 2 and Y_2 has degree 3. Then apply our concentration results to Y_1 and Y_2. An alternative way is to use Theorem 4.11 or 4.12 and think of $t_{xy}t_{yz}(1 - t_{xz})$ as a function Z_e.

Counting Number of Pairs of Fixed Distance. One can use the same idea to prove the strong concentration of the numbers of more complicated objects. For instance, consider the number of pairs of vertices of distance 2 in a random graph. This number can be written as

$$Y = \sum_{x,y} (1 - t_{xy})\left(1 - \prod_{z} (1 - t_{xz}t_{yz})\right).$$

The term $1 - t_{xy}$ equals 1 if and only if x and y are not adjacent. The more complicated term $(1 - \Pi_z (1 - t_{xx}t_{yz}))$ equals 1 if there is a vertex z adjacent to both x and y and 0 otherwise. Now we exploit the polynomial method. In certain range of p, Y can be very well approximated by the polynomials consisting of low degree terms of Y. A strong concentration result on these polynomials yields a strong concentration result on Y. Again one can also use Theorem 4.11 or 4.12 as suggested for the previous problem.

Generalization. Our method is not restricted to $G(N, p)$. There are several directions to generalize the results derived in this section.

1. Our concentration results do not require the random variables t_i to be identically distributed. Thus, we can consider more general models where the edges can have different distribution.
2. We can also consider more general random structures. For instance, instead of $G(N, p)$, we can study the random hypergraph $H(r, N, p)$, where the edges are random subsets of size r chosen independently with probability p. All results derived for $G(N, p)$ still hold without a slightest modification.

6.6. Better Bound with a Different Induction

The proof of Theorem 4.2 used induction on the degree of the polynomial. In the case of subgraph counting, it means that we use induction on k, the number of edges of the graph

in question. In [76], the present author observed that one can also use induction on m, the number of vertices. Since m is typically smaller than k, this reduces the number of induction steps and results in a stronger bound. For instance, we have shown [76]

Theorem 6.8. *If H is balanced and $\varepsilon^2\mathbb{E}(Y_H)^{1/(m-1)} = \omega(\log n)$, then there is a positive constant $c = c(K, H)$ such that*

$$Pr(Y \geq (1+\varepsilon)\mathbb{E}(Y_H)) \leq e^{-c\varepsilon^2\mathbb{E}(Y_H)^{1/(m-1)}}.$$

This theorem provides a better bound than Corollary 6.3 for most graphs. More interestingly, the bound turns out to be sharp (up to a logarithmic term in the exponent) for certain graphs. It is a tantalizing question to obtain a sharp bound for $Pr(Y \geq (1+\varepsilon)\mathbb{E}(Y_H))$ for an arbitrary graph H. Some progress has been made, very recently, on this problem by J. H. Kim and the present author, using a variant of Lemma 3.1 combined with a refinement of the first step in the proving scheme presented in Section 3.

Among others, the paper [76] also provides a nontrivial lower bound for the probability in Theorem 6.8 and a variant of Theorem 6.8 for counting extensions.

7. APPLICATIONS IN ADDITIVE NUMBER THEORY

7.1. Introduction

In additive number theory, one often asks whether or not there exists a sequence with certain properties. One of the essential ways to obtain an affirmative answer is to use the probabilistic method, established by Erdős. To show that a sequence with some property $\mathcal{P}$ exists, it suffices to show that a properly defined random sequence satisfies $\mathcal{P}$ with positive probability. The power of the probabilistic method has been justified by the fact that in most problems solved by this method, it seems almost impossible to come up with a constructive proof.

In this section, $\mathbb{N}$ denotes the set of positive integers. One usually defines a random sequence by choosing each x from $\mathbb{N}$ with some probability p_x, independently. We will use the binary random variable t_x to represent this choice. The probability space we are talking about is the (infinite dimension) product space spanned by the atom variables t_x's.

Quite frequently, the property $\mathcal{P}$ requires that for all sufficiently large $n \in \mathbb{N}$, some relation $\mathcal{P}(n)$ holds. The general strategy to handle this situation is the following. For each n, one first shows that $\mathcal{P}(n)$ fails with a small probability, say $s(n)$. If $s(n)$ is sufficiently small so that $\sum_{n=1}^{\infty} s(n)$ converges, then by Borel and Cantelli's lemma, $\mathcal{P}(n)$ holds for all sufficiently large n with probability 1 (see, for instance, [31], Chapter 3).

The main issue in the above argument is to show that for each n, $\mathcal{P}(n)$ holds with high probability. It turns out that in many problems, this is equivalent to showing that a properly defined polynomial Y_n (with variables t_x, $x \leq n$) is strongly concentrated and this is where our concentration results appear very useful.

In the next subsection, we shall present several results on additive bases which can be proved using our bounds in a rather canonical way. These include an answer to a 20 years old question of Nathanson on the existence of thin Waring bases.

7.2. Thin Bases

A subset X of $\mathbb{N}$ is a basis of order k if every sufficiently large number $n \in \mathbb{N}$ can be represented as a sum of k elements of X. The notion of bases is fundamental in additive number theory. It is clear that $\mathbb{N}$ itself is a basis of any order. On the other hand, the most interesting bases are, perhaps, the set of all rth powers, for arbitrary $r \geq 2$. The famous Waring's conjecture (proved by Hilbert, Hardy-Littlewood, Vinogradov, and many other by the beginning of the last century) asserts that for any fixed r and sufficiently large k, $\mathbb{N}^r$, the set of all rth powers is a k basis.

An important question is how to measure the quality of a basis. The most obvious way is to look at the density. We might want to say that a basis is better if it has smaller density. A finer measurement is to look at the number of representations of each number $n \in \mathbb{N}$. If X is a basis of order k, we denote this number by $R_X^k(n)$. We say that X is *thin* if, for each n, $R_X^k(n)$ is positive, but small. The study of thin bases was started by Rohrbach and Sidon back in 1930s and has since then attracted considerable attention from both combinatorialists and number theorists (see [15, 31, 25] and their references).

Now let us take a look at the classical bases $\mathbb{N}^r$ (including the case $r = 1$). These bases are very far from being thin, due to the following deep theorem of Vinogradov [70].

Theorem 7.1. *For any fixed $r \geq 1$, there is a constant $k(r)$ such that if $k \geq k(r)$, then, $R_{\mathbb{N}^k}^k(n)$, the number of representations of n as a sum of k rth powers satisfies*

$$R_{\mathbb{N}^r}^k(n) = \Theta(n^{k/r-1}),$$

for every positive integer n.

It is natural to ask whether $\mathbb{N}^r$ contains a thin subbasis. The very first case is $r = 1$ and $k = 2$. Sidon, in the 1930s, conjectured that $\mathbb{N}$ has a 2-basis such that $R_2(n) = n^{o(1)}$ for all large n. Twenty years later, Erdős confirmed this conjecture by showing [15] the following:

Theorem 7.2. *There is a subset $X \subset \mathbb{N}$ such that $R_X^2(n) = \Theta(\log n)$, for all sufficiently large n.*

In 1990, Erdős and Tetali [23] generalized Theorem 7.2 to arbitrary k (still $r = 1$).

Theorem 7.3. *There is a subset $X \subset \mathbb{N}$ such that $R_X^k(n) = \Theta(\log n)$, for all sufficiently large n.*

The case $r \geq 2$ appears much more difficult. For many years, researchers focused on a simpler question that whether $\mathbb{N}^r$ contains a subbasis of small density. This question has been investigated intensively for $r = 2$ [18, 13, 82, 83, 81, 63]. Choi, Erdős, and Nathanson proved in [13] that $\mathbb{N}^2$ contains a subbasis X of order 4, with $X(m) \leq m^{1/3+\epsilon}$, where $X(m)$ denotes the number of elements of X not exceeding m. Improving this result, Zöllner [82, 83] shows that for any $k \geq 4$ there is a subbasis $X \subset \mathbb{N}^2$ of order k satisfying $X(m) \leq m^{1/k+\epsilon}$ for arbitrary positive constant ϵ. Wirsing [81], sharpening Zöllner's theorem, proved that for any $k \geq 4$ there is a subbasis $X \subset \mathbb{N}^2$ of order k satisfying $X(m) = O(m^{1/k}\log^{1/k} m)$. It is easy to see, via the pigeon-hole principle, that Wirsing's

result is best possible, up to the log term. A short proof of Wirsing's result for the case $k = 4$ was given by Spencer in [63]. For $r \geq 3$, much less was known. In 1980, Nathanson [53] proved that $\mathbb{N}^r$ contains a subbasis with density $o(m^{1/r})$. In the same paper, he raised the following question.

Question. *Let $r \geq 2$ and k be fixed, positive integers, where k is sufficiently large compared to r. What is the smallest density of a subbasis of order k of $\mathbb{N}^r$? Can it be $m^{1/k+o(1)}$?*

It is clear that the conjectured density $m^{1/k+o(1)}$ is best possible up to the $o(1)$ term. Very recently, we succeeded to prove the extension of Theorem 7.3 to $\mathbb{N}^r$ for arbitrary r [77].

Theorem 7.4. *For any fixed $r \geq 2$, there is a constant $k(r)$ such that if $k \geq k(r)$, then $\mathbb{N}_0^r$ contains a subset X such that*

$$R_X^k(n) = \Theta(\log n),$$

for every positive integer $n \geq 2$ and $R_X^k(1) = 1$.

This theorem implies, via the pigeon hole principle that X has density $O(m^{1/k}\log^{1/k} m)$, settling Nathanson's question.

In the rest of this section, we first use our concentration results to give a simple and short proof for a slightly more general version of Theorem 7.3. Next, based on the frame work presented in this proof, we sketch the ideas behind the proof of Theorem 7.4.

7.3. Thin Linear Bases

In this section, we extend Theorem 7.3 by allowing a representation to be a linear combination with fixed coefficients.

Fix k positive integers $a_1, \ldots, a_k$, where $gcd(a_1, \ldots, a_k) = 1$. Let $Q_X^k(n)$ be the number of representations of n of the form $n = a_1x_1 + \cdots + a_kx_k$, where $x_i \in X$. We shall prove:

Theorem 7.5. *There is a subset $X \subset \mathbb{N}$ such that $Q_X^k(n) = \Theta(\log n)$, for all sufficiently large n.*

The proof of Theorem 7.5 is based essentially on the scheme developed for the more difficult Theorem 7.4 in [77]. However, since the proof of Theorem 7.5 is much simpler, we present it first in order to give the reader a better understanding of our method.

The assumption $gcd(a_1, \ldots, a_k) = 1$ is necessary, due to a simple number theoretic reason. Theorem 7.3 follows from Theorem 7.5 by setting all $a_i = 1$.

To start the proof, we use Erdős's idea and define a random set X as follows. For each $x \in \mathbb{N}$, choose x with probability $p_x = cx^{1/k-1}\log^{1/k} x$, where c is a positive constant to be determined. Let t_x be the indicator random variable of this choice; thus, t_x is a $\{0, 1\}$ random variable with mean p_x.

Fix a number n, and let $\mathbf{Q}_n$ be the set of all k-tuples $(x_1, \ldots, x_k)$, where x_i are positive

integers and $\sum_i a_i x_i = n$. The number of representations of n using elements from the random sequence X can be expressed as a random variable in the following way

$$Y_n = \sum_{(x_1, \ldots, x_k) \in \mathbf{Q}_n} t_{x_1} \cdots t_{x_k}. \tag{25}$$

It is obvious that Y is a polynomial of degree k in $t_1, \ldots, t_n$. We now show that with probability close to 1, Y_n is $\Theta(\log n)$ for any sufficiently large n. It is easy to show that $\mathbb{E}(Y_n)$ is of the right order, namely $\log n$. Next, we want to make use of Corollary 4.6. The main obstruction here is that Y_n, as a polynomial, does have partial derivatives with large expectations which violate the condition of Corollary 4.6. For instance, consider the representation $a_1 K + a_2 x_2 + \cdots + a_k x_k$, where K is a constant. The partial derivative with respect to $t_{x_2}, \ldots, t_{x_k}$ has expectation $p_K = \Theta(1)$. However, we could easily overcome this obstruction by splitting Y_n into two parts, as follows. Set $a = 0.4$ (0.4 can be any small constant) and let $\mathbf{Q}_n^{[1]}$ be the subset of $\mathbf{Q}_n$ consisting of all tuples whose smallest element is at least n^a and $\mathbf{Q}_n^{[2]} = \mathbf{Q}_n \backslash \mathbf{Q}_n^{[1]}$. We break Y_n into the sum of two terms corresponding to $\mathbf{Q}_n^{[1]}$ and $\mathbf{Q}_n^{[2]}$, respectively:

$$Y_n = Y_n^{[1]} + Y_n^{[2]},$$

where

$$Y_n^{[j]} = \sum_{(x_1, \ldots, x_k) \in \mathbf{Q}_n^{[j]}} t_{x_1} \cdots t_{x_k}.$$

Intuitively, $Y_n^{[1]}$ should be the main part of Y_n, since in most solutions of $\sum_{i=1}^k a_i x_i = n$, all $x_i = \Theta(n)$. The theorem follows immediately from the following two statements and Borel-Cantelli's lemma.

(A) $\mathbb{E}(Y_n^{[1]}) = \Theta(\log n)$ and $Pr(|Y_n^{[1]} - \mathbb{E}(Y_n^{[1]})| \geq \mathbb{E}(Y_n^{[1]})/2) \leq n^{-2}$.
(B) For almost every sequence X, there is a finite number $M(X)$ such that $Y_n^{[2]} \leq M(X)$ for all sufficiently large n.

(A) and (B) confirm our intuition. The main part of Y_n indeed comes from $Y_n^{[1]}$ as $Y_n^{[2]}$'s contribution is bounded by a constant.

In order to apply Corollary 4.6 to verify (A), we first need the following lemma, which asserts that $\mathbb{E}(\partial_A Y_n^{[1]})$'s are sufficiently small.

Lemma 7.6. *For all nonempty multisets A of size at most $k - 1$,*

$$\mathbb{E}(\partial_A Y_n^{[1]}) = O(n^{-a/2k}).$$

Proof. Consider a (multi-) set A of $k - l$ elements $y_1, \ldots, y_{k-l}$. For a permutation $\pi \in S_k$ (where S_k denotes the symmetric group on $\{1, 2, \ldots, k\}$), let $\mathbf{Q}_{n,l,\pi}^{[1]}$ be the set of l-tuples $(x_1, \ldots, x_l)$ of positive integers satisfying $x_i \geq n^a$ for all i and

$$\sum_{i=1}^{l} a_{\pi(i)} x_i = n - \sum_{j=1}^{k-l} a_{\pi(l+j)} y_j.$$

A simple consideration shows that

$$\partial_A(Y_n^{[1]}) \le b(k) \sum_{\pi \in S_k} \sum_{(x_1, \ldots, x_l) \in \mathbf{Q}_{n,l,\pi}^{[1]}} t_{x_1} \cdots t_{x_l},$$

where $b(k)$ is a constant depending on k. By symmetry, it now suffices to verify the following:

$$\mathbb{E}\left(\sum_{(x_1, \ldots, x_l) \in \mathbf{Q}_{n,l,\pi_0}^{[1]}} t_{x_1} \cdots t_{x_l} \right) = O(n^{-a/2k}),$$

where π_0 is the identity permutation. Without loss of generality, we can assume that $x_l = \max(x_1, \ldots, x_l)$. Set $m = n - \sum_{j=1}^{k-l} a_{l+j} y_j$; since $\sum_{i=1}^{l} a_i x_i = m$ and the a_i's are fixed numbers, it follows that $x_l = \Omega(m/l)$. Using the fact that $\sum_{x=1}^{m} x^{1/k-1} \approx \int_1^m z^{1/k-1}\, \partial z \approx m^{1/k}$, we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{(x_1, \ldots, x_l) \in \mathbf{Q}_{n,l,\pi_0}^{[1]}} t_{x_1} \cdots t_{x_l} \right) &= O\left(\sum_{(x_1, \ldots, x_l) \in \mathbf{Q}_{1,l,\pi_0}} p_{x_1} \cdots p_{x_l} \right) \\
&= O(\log n) \sum_{\substack{n^a \le \min(x_1, \ldots, x_l) \\ a_1 x_1 + \cdots + a_l x_l = m}} x_1^{1/k-1} \cdots x_l^{1/k-1} \\
&= O\left(\left(\sum_{x=1}^{m} x^{1/k-1} \right)^{l-1} (m/l)^{1/k-1} \log n \right) \\
&= O(m^{(l-1)/k} (m/l)^{1/k-1} \log n) \\
&= O(m^{(l-k)/k} \log n) \\
&= O(n^{-a/2k}),
\end{aligned}$$

since $k - l \ge 1$ and $m \ge n^a$ by the definition of $\mathbf{Q}_{1,l,\pi}$. This ends the proof of the lemma. ∎

The last step in the previous calculation explains the restriction $\min(x_i)_{i=1}^k \ge n^a$. This assumption guarantees that every partial derivatives of $Y_n^{[1]}$ has small expectation.

From the above calculation, it follows immediately (by setting $l = k$ and $m = n$) that $\mathbb{E}(Y_n^{[1]}) = O(\log n)$. Moreover, a straightforward argument shows that if $c \to \infty$, then $\mathbb{E}(Y_n^{[1]})/\log n \to \infty$. Indeed, there are at least $\Theta(n^{k-1})$ k-tuples $(x_1, x_2 \ldots, x_k)$ in $\mathbf{Q}_n^{[1]}$, where $x_i = \Theta(n)$ for all $i \le k$, where the constants in the Θ's depend only on k and the a_i's. On the other hand, each such tuple contributes at least $c^k n^{1-k} \log n$ to $\mathbb{E}(Y_n^{[1]})$. Therefore, by increasing c, we can assume that $\mathbb{E}(Y_n^{[1]})$ satisfies the condition of Corollary 3.6. Corollary 3.6 then applies and implies (A).

Before continuing with the proof of (B), let us pause for a moment and show why we

could not apply Azuma's inequality to prove (A). The reason is that the Lipschitz coefficient of $Y_n^{[1]}$ is way too large. It is clear that there is a number x which appears in $\Omega(n^{k-2})$ tuples in $\mathbf{Q}_n^{[1]}$ (in fact, most number x do so). For such an x, changing t_x, in the worst case, might change $Y_n^{[1]}$ by $\Omega(n^{k-2})$. Thus, the Lipschitz coefficient of $Y_n^{[1]}$ is $\Omega(n^{k-2})$. This coefficient is clearly too large for Azuma's inequality to deliver a non-trivial bound.

Now we turn to the proof of (B), which is purely combinatorial. We say that a l-tuple $(x_1, \ldots, x_l)$ $(l \leq k)$ is an l-representation of n if there is a permutation $\pi \in S_k$ such that $\sum_{i=1}^{l} a_{\pi(i)} x_i = n$. For all $l < k$, let $Q_X^l(n)$ be the number of l-representations of n. With essentially the same computation as in the previous lemma, one can show that $\mathbb{E}(Q_X^l(n)) = O(n^{-1/k} \log n) = O(n^{-1/2k})$. Proposition 4.10 then implies that for a sufficiently large constant M_1, with probability $1 - O(n^{-2})$, the maximum number of disjoint representations of n in $Q_X^l(n)$ is at most M_1. By Borel and Cantelli's lemma, we conclude that for almost every random sequence X there is a finite number $M_1(X)$ such that for any $l < k$ and all n, the number of disjoint l-representations of n from X is at most $M_1(X)$.

Using a computation similar to the one in the proof of Lemma 7.6, one can deduce that $\mathbb{E}(Y_n^{[2]}) = O(n^{(a-1)/k} \log n) = O(n^{-1/2k})$. Indeed, since $x_1 \leq n^a$, instead of $(\sum_{x=1}^{n} x^{1/k-1})^{k-1}$, one can write $\sum_{x=1}^{n^a} x^{1/k-1} (\sum_{x=1}^{n} x^{1/k-1})^{k-2}$ and the bound follows. So, again by Proposition 4.10 and Borel and Cantelli's lemma, there is a constant M_2 such that, almost surely, the maximum number of disjoint k-representations of n in $Y_n^{[2]}$ is at most M_2 for all large n. From now on, it would be useful to think of $Y_n^{[2]}$ as a family of sets of size k, each corresponds to a representation of n.

We say that a sequence X is *good* if it satisfies the properties described in the last two paragraphs.

To finish the proof, it suffices to show that if X is good, then $Y_n^{[2]}$ is bounded by a constant. This follows directly from a well-known combinatorial result of Erdős and Rado's [21], stated below. A collection of sets $A_1, \ldots, A_r$ forms a sun flower if the sets have pair-wise the same intersection. Erdős and Rado have shown:

Lemma 7.7. *If H is a collection of sets of size at most k and $|H| > (r-1)^k k!$, then there are r sets forming a sun flower.*

Set $M(X) = (\max(M_1(X)k!, M_2))^k k!$. Assume that n is sufficiently large. It is clear that if $Y_n^{[2]} > M(X)$, then by Erdős and Rado's sunflower lemma, $Y_n^{[2]}$ contain a $M_3 = \max(M_1(X)k!, M_2) + 1$ sunflower. If the intersection of this sunflower is empty, then the petals form a family of M_3 disjoint k-representations of n. Otherwise, assume that the intersection consists of $y_1, \ldots, y_j$, where $1 \leq j \leq k-1$. By the pigeonhole principle, there is a permutation $\pi \in S_k$ such that one can find $M_1(X) + 1$ $(k-j)$-representations of $m = n - \sum_{i=1}^{j} a_{\pi(i)} y_i$ among the sets obtained by the petals minus their common intersection. These $M_1(X) + 1$ sets are disjoint due to the definition of the sun flower. Therefore, in both cases we obtain a contradiction. ∎

Remark. One can prove a statement similar to Theorem 7.5 when $Q_X^k(n)$ is required to be $\Theta(g(n))$, for any "reasonable" function $g(n) \gg \log n$ [reasonable here means that one should be able to set p_x so that the expectation of $Q_X^k(n)$ is $g(n)$]. In fact, in such a case, one can easily show that $Q_X^k(n)$ is asymptotically $g(n)$, since now in the proof of (A) one can have a deviation tail $o(g(n))$.

7.4. Thin Waring Bases

Here we sketch the proof of Theorem 7.4, using the framework provided in the previous subsection. To start, we define a random subset of $\mathbb{N}^r$ as follows. Choose, for each $x \in \mathbb{N}$, x^r with probability $p_x = cx^{-1+r/k} \log^{1/k} x$, where c is a sufficiently large positive constant. Again let t_x denote the characteristic random variable representing the choice of x^r: $t_x = 1$ if x^r is chosen and 0 otherwise. Similar to (25), the number of representations of n (not counting permutations), restricted to X, can be expressed as follows

$$R_X^k(n) = \sum_{x_1^r + \cdots + x_k^r = n} \prod_{j=1}^{k} t_{x_j} = Y(t_1, \ldots, t_{\lfloor n^{1/r} \rfloor}). \tag{26}$$

Given the framework of the previous proof, the remaining difficulty is to estimate the expectations of Y and its partial derivatives. In the following, we shall focus on the expectation of Y. Notice that

$$\mathbb{E}(Y) = \sum_{x_1^r + \cdots + x_k^r = n} c^k \prod_{j=1}^{k} x_j^{-1+r/k} \log^{1/k} x_j.$$

To see that the right hand side has order $\Theta(\log n)$, one may argue as follows. A typical solution $(x_1, \ldots, x_k)$ of $\sum_{j=1}^k x_j^r = n$ should satisfy $x_j = \Theta(n^{1/r})$, for all j. Thus, a typical term in the sum has order $\Theta(n^{-k/r+1} \log n)$. On the other hand, by Vinogradov's theorem, the number of terms is $\Theta(n^{k/r-1})$, and we would be done by taking the product. The trouble is that there could be many nontypical solutions with larger contribution. For instance, assume that $\mathbf{x} = (x_1, \ldots, x_k)$ is a solution where $1 \le x_j \le P_j$ and some of the P_j's are considerably smaller than $n^{1/r}$ (for example, $P_1 = n^{\varepsilon}$ with $\varepsilon \ll 1/r$). The contribution of the term corresponding to $\mathbf{x}$ is at least $\Omega(\prod_{j=1}^s P_j^{-1+r/k})$, which is significantly larger than the contribution of a typical term.

To overcome this trouble, we need an upper bound on the number of solutions of the equation $\sum_{j=1}^k x_j^r = n$, restricted to a box of type which is the product of s intervals $[1, P_j]$ $(j = 1, \ldots, s)$, for arbitrary positive integers $P_1, \ldots, P_k$. Denote this number by $Root(P_1, \ldots, P_k)$. We proved the following lemma, which generalizes the lower bound in Vinogradov's theorem.

Lemma 7.8. *For a fixed positive integer $r \ge 2$, there exists a constant k_r such that the following holds. For any constant $k \ge k_r$, there is a positive constant $\delta = \delta(r, k)$ such that for every sequence $P_1, \ldots, P_k$ of positive integers*

$$Root(P_1, \ldots, P_k) = O\left(n^{-1} \prod_{j=1}^{k} P_j + \prod_{j=1}^{k} P_j^{1-r/k-\delta} \right),$$

for all n.

The proof of this lemma requires a sophisticated application of the Hardy and

Littlewood's circle method and is beyond the scope of this paper. The reader may consult [77] for the full proof.

8. THE SEMIRANDOM METHOD

To prove that a certain object exists, it is enough to show that one can obtain one with positive probability as an output of a random process. This original observation of Erdős has become one of the most powerful methods in combinatorics, the so-called "probabilistic method."

A typical application of the probabilistic method is usually an "one-round" argument, in which one first generates a proper random space, and next shows that the set of desired objects has positive measure in this space. Quite frequently, this measure is close to 1, namely, the set of desired objects is abundant. Some examples of this type have already been presented in the last section.

The semirandom method is a sophisticated version of Erdős's argument which enables us to prove the existence of a very rare object by showing that such object can be obtained with positive probability as the output of a randomized algorithm running in many rounds. This powerful method has been the backbone of several breakthrough developments in combinatorics in the last two decades (see next subsection).

In this part of the paper, we first provide a description of the semirandom method, following [41]. Next, via a proof of a refinement of a theorem of Pippenger, we show how our polynomial method can be used to analyze a semirandom process in a convenient and robust way. The most appealing about this analysis is that polynomials not only naturally arise, but they also capture the heart of the problem, and give a quantitative explanation about the condition in the theorem. We conclude by describing two other results which can be proven using the combination of the semirandom method and the polynomial method. The first is a general result on list-coloring locally sparse graphs. The second is the solution to an old question of Segre in finite geometry. This latter result is, perhaps, the deepest application so far of our method. It is also the problem which motivated our development on concentration of polynomials; our first result on polynomials, Theorem 4.1, was originally proved as a lemma in order to complete this result.

8.1. Description of the Semirandom Method

Assume that we want to construct an object with certain structural constrains (such as a matching in a hypergraph or a proper coloring of a graph), random greedy construction is considered a natural way to generate it: Randomly order all possible elements of the desired object and *select* each of them one by one in the order if and only if it together with already selected ones causes no conflict, i.e., no violation to the given constrains. Here we mean by "select" that we choose and permanently add it to the desired object being constructed. We may discard at each step all elements that cause any conflict with already selected ones and then randomly select a nondiscarded one. This is an equivalent construction and will be called RGC which stands for random greedy construction. For example, the RGC of a matching in a hypergraph is the following. Initially, the matching being constructed is empty. At each step, consider the set M of selected edges and select one new edge among those which do not intersect any edge in M uniformly at random.

The semirandom method, or dynamic random construction using nibble (DRC), is an

approximated version of RGC. DRC has been initiated by a seminal paper of Ajtai, Komlós, and Szemerédi [2] to construct a large independent set in a triangle-free graph and become well known to combinatorialists by Rödl [56], who used the construction to settle the Erdős-Hannani conjecture regarding Steiner systems in design theory (see subsection 8.2). It has been developed and become more sophisticated and powerful to solve intriguing combinatorial problems regarding packings and edge-colorings of hypergraphs or multigraphs ([55, 36, 37]), chromatic numbers of sparse graphs ([38, 35]), Ramsey numbers ([39]), and some general graph coloring problems ([52]).

Rather than select one element at each step, DRC randomly and independently choose elements, not select yet though, with certain probability so that a bunch of elements are chosen together. This is called a nibble. The size of a nibble is the number of chosen elements or sometimes its expectation. Since the set of chosen elements may violate the constrains, we take a subset of it satisfying the constrains. Elements of this subset are called selected in the above meaning. Though the way constructing this subset varies depending on problems and/or for the sake of simplicity, chosen elements contribute no conflicts with other chosen ones are usually selected. We discard each unchosen element that may cause any new conflict if it were added to chosen elements regardless what the selected elements are. Since not all chosen elements are selected, some elements are unnecessarily discarded but the set of remaining, i.e., nondiscarded unchosen, elements is defined with respect to randomly and independently chosen elements so that the structure of the set might be well understood. After a step, we continue if the remaining structure still looks like random. The proof consists of showing that with a proper choice of the size of the nibbles, one has a positive chance to continue the algorithm until the desired object is found.

To carry out the above plan, the most crucial issue is to show that after each step, with positive probability, the remaining structure looks like a random substructure of the original one. Usually, it suffices to show that few important parameters behave essentially as their expectations predict. This, in turn, requires strong concentration results and here is the point where our results become ultimately useful.

8.2. Nearly Perfect Matchings in Hypergraphs

Let us start with the celebrated conjecture of Erdős and Hanani [16]. A partial Steiner system $S(t, r, m)$ is an r-uniform hypergraph on m vertices so that every set of t vertices is contained in at most one edge. Steiner systems are fundamental objects in design theory, an independent branch of discrete combinatorics (see [12] and its references).

From the definition, it is obvious that any partial Steiner system has at most $\binom{m}{t}/\binom{r}{t}$ edges. In the 1960s, Erdős and Hanani [16] made the following famous conjecture:

Erdős and Hanani's Conjecture. *Assume that t and r are fixed and $m \to \infty$; then there is a $S(t, r, m)$ partial Steiner system with $(1 - o(1))\binom{m}{t}/\binom{r}{t}$ edges.*

This conjecture was confirmed by Rödl in a seminal paper [56], which formalized the semi-random method. Later, Pippenger [54] (see also [27]) realized that Rödl's theorem is a special case of a more general statement. Given a hypergraph H, we denote by $codeg(H)$ the maximum number of edges sharing two points in common, and by $\mathcal{U}(H)$ the minimum number left uncovered by a matching. Pippenger showed:

Theorem 8.1. *Assume that H is an r-uniform and D-regular hypergraph on n vertices and $codeg(H) = C = o(D)$, then there is a matching which covers all but $o(n)$ vertices, i.e., $\mathcal{U}(H) = o(n)$.*

To see that Pippenger's theorem generalizes Rödl's result, define a special hypergraph H in the following way. H has $n = \binom{m}{t} = \Theta(m^t)$ vertices, where each vertex represents a t-tuple of a ground set of m elements. An edge of H consists of those t-tuples which are subsets of the same r set. Thus H is $\binom{r}{t}$-uniform and $\binom{m-t}{r-t}$-regular. Moreover, the codegree of H is $\binom{m-t-1}{r-t-1} = C = o(D)$. To conclude, notice that a large partial Steiner system corresponds to a large matching in H.

Theorem 8.1, however, does not supply an explicit estimate for the error term $o(n)$. For instance, it is not clear how C and D contribute in $\mathcal{U}(H)$. Sharpening this error term, motivated by applications from diverse areas, is a challenging problem which attracted the attention of several researchers (see [29, 30, 5, 43, 79] and their references).

In [79], we successfully combined the semirandom method with our concentration results to prove the following theorem, which gives the strongest explicit bound known for $\mathcal{U}(H)$, under the assumption of Theorem 8.1. This bound improves upon an earlier bound obtained by Kostochka and Rödl [43], and generalizes a bound obtained by Alon, Kim, and Spencer [5].

Theorem 8.2. *Let k be a fixed positive integer which is at least 3. If H is a $(k+1)$-uniform, D-regular hypergraph on n vertices with codegree C, then $\mathcal{U}(H) = On(\frac{C}{D})^{1/k}\log^{c/k}D$, where c is a positive constant not depending on k.*

In the rest of this subsection, we sketch the proof of this theorem in order to illustrate our ideas. While Pippenger result gives a very elegant generalization to Rödl's result, it is not so clear why the codegree is the right parameter to look at. Our analysis gives a nice explanation for this question and also opens a new direction for further improvements.

To start, let us describe a randomized algorithm to generate a large matching. Pick each edge in the hypergraph independently with probability $p = \theta/D$, where θ is a small number to be determined. An edge is *lonely* if it is chosen and does not intersect any other chosen edge. The set of lonely edges form a matching by definition. Remove from the hypergraph all vertices covered by the chosen edges, and repeat the operation on the remaining (induced) subhypergraph.

Here is a rough analysis of this process. Consider the first step. A vertex is *dead* if it is covered by a chosen edge (and then removed by the end of the step); otherwise, it *survives*. Clearly, a vertex survives if none of the edges adjacent to it was chosen. This happens with probability

$$p_{sur} = (1-p)^D \approx e^{-pD} \approx 1 - \theta.$$

So after the first step, we expect that the number n' of surviving vertices be about $n(1-\theta)$. Next, let us consider the degrees in the remaining hypergraph. Fix a vertex v, we say that an edge e adjacent to v remains *v-active* if all vertices in $e\backslash v$ survive.

Assume, for a moment, that the survival events are independent. Under this assumption, we expect that for a fixed pair v and e, e remains v-active is

$$p_{act} = p_{sur}^{|e\setminus v|} = p_{sur}^{k} \approx (1 - \theta)^k.$$

Thus, in the new hypergraph, one may expect that the new degrees are roughly $Dp_{act} = D(1 - \theta)^k$.

Now let us check how big a matching has been extracted at this step. Since each edge is chosen with probability p and there are $nD/(k + 1)$ edges, we have chosen about $pnD/(k + 1) = \theta n/(k + 1)$ edges. For any edge e, the probability that no edges adjacent to e are chosen is $(1 - p)^{(k+1)D} \approx (1 - p(k + 1)D) = 1 - (k + 1)\theta$. Therefore, with this probability, a chosen edge would be lonely. Together, we would have

$$M = \frac{\theta n}{k + 1}(1 - (k + 1)\theta),$$

lonely edges.

In general, after the ith step, we expect to obtain a hypergraph on $n_i \approx n(1 - \theta)^i$ vertices, in which every vertex has degree roughly $D_i \approx D(1 - \theta)^{ki}$, and a set of $M_i = \frac{\theta n_{i-1}}{k+1}(1-(k+1)\theta) = \frac{\theta n_{i-1}}{k+1}(1-O(\theta))$ lonely edges. (Here we understand that $n_0 = n$.)

Let T be an integer so that $D_T \approx C \log^c n$, for some appropriately chosen constant c. (Perhaps we can set $c = 1$, but we do not want to optimize it at this point.) Thus, $(1 - \theta)^{kT} \approx (C/D)\log^c n$. After step T, the hypergraph has $n_T \approx n(1 - \theta)^T \approx (C/D)^{1/k} n \log^{c/k} n$ vertices. Moreover, up to this point, we obtained a matching of size approximately

$$M = \sum_{i=1}^{T} M_i = \sum_{i=1}^{T} \frac{\theta n_{i-1}}{k + 1}(1 - O(\theta)).$$

Recall that $n^i \approx n(1 - \theta)^i$, arithmetic shows that the last sum is approximately

$$\frac{n}{k + 1}(1 - O(\theta)).$$

This matching covers $n(1 - O(\theta))$ vertices. So the number of vertices uncovered is

$$(C/D)^{1/k} n \log^{c/k} n + O(\theta n).$$

Set θ small ($(C/D)^{1/k}\log^{c/k} n$), say), one may conclude that there are only $O((C/D)^{1/k} n \log^{c/k} n)$ vertices left uncovered.

To make the above analysis rigorous, it is essential to show that with positive probability, the parameters n_i, D_i, and M_i behave almost as expected. Before starting to prove this, let us notice that the survival events are not independent as we assumed, but rather positively correlated. If v is adjacent to u, then the fact that v survives seems to increase the probability that u also survives.

In the following, let us show how one can handle the situation in the first and typical step. To each edge e, define a binary atom variable t_e as follows: $t_e = 1$ if e is chosen and 0 otherwise. Obviously, the t_e's are independent and have mean p.

The number of survival vertices n' is relatively easy to handle, based on the following observation: n', as a function in the t_e, has bounded Lipschitz coefficient. Indeed,

changing any t_e from 0 to 1 or vice versa could change n' by at most $k+1$, the size of one edge. Thus, one can use a Azuma or Talagrand type inequality to show that n' is strongly concentrated around its mean, which is approximately $n(1-\theta)$. The size of the lonely set can be dealt with almost the same way (see [79] for details).

The harder part of the proof is to deal with D'. There are two main obstacles. First, the previous estimate $D(1-\theta)^k$ on the expectation is not rigorous, since the assumption that the survival events are independent is false. Second and more crucially, D', as a function in the t_e's, does have a large Lipschitz coefficient which kills the possibility of applying a Azuma or Talagrand type inequality. To see this, assume that u and v share C edges in common, then the choice of any edge containing u may effect the degree D'_v of v by as much as C. Quite amazingly, it has turned out that our polynomial method could be used to handle both obstacles at the same time in a simple and robust way, which, in addition, is easy to adapt to other problems. The purpose of the next few paragraphs is to show how this can be achieved.

For a fixed point v, we first write the number of v-active edges as a function in the atom variables. It is clear that a point u survives if and only if no edges adjacent to u were chosen. So the indicator function I(u survives) of this event can be expressed as

$$\mathrm{I}(u \text{ survives}) = \prod_{u \in f} (1 - t_f).$$

On the other hand, an edge e adjacent to v is v-active if all k vertices in the set $e_v = e \backslash v$ survive. Therefore,

$$\mathrm{I}(e\ v\text{-active}) = \prod_{u \in e_v} \mathrm{I}(u \text{ survives}) = \prod_{u \in e_v} \prod_{u \in f} (1 - t_f).$$

Now the number of v-active edges is

$$D' = \sum_{v \in e} \mathrm{I}(e\ v\text{-active}) = \sum_{v \in e} \prod_{u \in e_v} \prod_{u \in f} (1 - t_f).$$

Needless to say, D' is a polynomial in the atom variables. However, D' is not positive and its degree is high. The key tool here is the polynomial method, described in Section 4.7. Using this method, we first approximate D' by lower degree polynomials. As already pointed out, the most natural to do this is to consider the polynomials formed by the low degree terms of D'.

It is very useful to notice that for binary variables t_i, the following inequalities hold

$$1 - \sum_{i=1}^{M} t_i \le \prod_{i=1}^{M} (1 - t_i) \le 1 - \sum_{i=1}^{M} t_i + \sum_{1 \le i < j \le M} t_i t_j.$$

Therefore, we can sandwich D' between two polynomials Y_1 and Y_2 ($Y_1 \le D' \le Y_2$ everywhere), where

$$Y_1 = \sum_{e \ni v} \left(1 - \sum_{u \in e_v} \sum_{f \ni u} t_f\right),$$

$$Y_2 = Y_1 + \sum_{e \ni v} \sum_{u,u' \in e_v} \sum_{f \ni u, f' \ni u'} t_f t_{f'} = Y_1 + X.$$

In a term $t_f t_{f'}$, f and f' are not necessarily different. But it turns out that the number of terms $t_f t_{f'}$, where $f = f'$ is negligible, so in the following we will assume, for the sake of convenience, that $f \neq f'$. Consequently, $\mathbb{E}(t_f t_{f'}) = p^2$.

Next, we verify that the expectations of Y_i are close to each other. Indeed,

$$\mathbb{E}(Y_1) = D(1 - kDp) = D(1 - k\theta),$$

$$\mathbb{E}(Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(X) \leq D(1 - k\theta) + Dk^2D^2p^2 = D(1 - k\theta + k^2\theta^2).$$

With a proper choice of θ, we can afford an error term $O(\theta^2)$. So it suffices to show that with high probability, both Y_1 and Y_2 are in the range $D(1 - k\theta \pm O(\theta^2))$.

What does high probability mean? Recall that we want D' to behave nicely at every vertex v. On the other hand, the number of vertices in the hypergraph does not depend on the degree D, so it is obvious that we cannot use a straightforward union bound to achieve our goal, but rather need to invoke the Lovász Local Lemma to show that there is a positive probability that all degrees behave as expected. Here is the description of this lemma.

Consider a set of events $A_1, \ldots, A_m$. The *dependency* graph of $A_1, \ldots, A_m$ is a graph on $\{1, \ldots, m\}$ such that A_i is mutually independent of all events A_j where i is not adjacent to j. Let d_i be the degree of i. One of the variants of the famous Lovász Local Lemma [17, 8] is the following

Lemma 8.3. *If the probability that A_i holds is at least $1 - 1/4d_i$ for all i, then there is a positive probability that all of the A_i hold.*

In our setting, at each vertex v, we consider the event that D'_v is sandwiched between $D(1 - k\theta - O(\theta^2))$ and $D(1 - k\theta + O(\theta^2))$. It is clear that if the hypergraph distance between u and v is at least 5, then any choice of an edge that affects D'_v cannot affect D'_u. Therefore, in the setting of the Local Lemma, each event has degree $O(D^4)$. To apply the Local Lemma, it suffices to show that each event holds with probability at least $1 - o(1/D^4)$.

Now it remains to show that for some constant K, the probability that Y_1 and Y_2 deviate from their means by more than $K\theta^2D$ is $o(1/D^4)$. The proof for Y_1 is simple, since Y_1 is a sum of independent variables. All we need now is to show that X is strongly concentrated.

Claim 8.4. *There is a constant K such that*

$$Pr(X \geq K\theta^2D) \leq e^{-5 \log D}.$$

Recall that

$$X = \sum_{e \ni v} \sum_{u,u' \in e_v} \sum_{f \ni u, f' \ni u'} t_f t_{f'},$$

and $\mathbb{E}(X) \leq k^2\theta^2 D$. To apply our concentration results, let us compute $\mathbb{E}_j(X)$, for $j = 1$ and 2. Denote by S_f the set of f' such that $t_f t_{f'}$ appear in X

$$\partial_{t_f} X = \sum_{f' \in S_f} t_{f'}.$$

So $\mathbb{E}(\partial_{t_f} X) = p|S_f|$. To estimate $|S_f|$, notice that $f' \in S_f$ if and only if there are two nodes $u \in f$ and $u' \in f'$ such that u, u' are contained in some edge e adjacent to v. Since f is fixed, there are $(k + 1)$ ways to choose u. Choose and edge e containing both u and v. Once e is fixed, there are k ways to choose u', and with each u' there are D' ways to choose f'.

But how many e can we choose? This number is exactly the codegree of u and v and can be at most C. Therefore,

$$|S_f| \leq (k + 1)CkD = k(k + 1)CD.$$

It follows that

$$\mathbb{E}(\partial_{t_f} X) = p|S_f| \leq k(k + 1)\theta C.$$

Next, we consider a second-order partial derivative of X with respect to t_f and $t_{f'}$. Assume that $f \neq f'$, then the partial derivative is the coefficient of $t_f t_{f'}$. This coefficient is the number of ways we can choose $u \in f$ and $u' \in f'$ and an edge e through u, u' and v. Since the number of edges through u, u' and v is at most C, we obtain

$$\partial_{(t_f, t_{f'})} X \leq (k + 1)^2 C.$$

Together, we have that both $\mathbb{E}_1(X)$ and $\mathbb{E}_2(X)$ are at most $(k + 1)^2 C$. We can now apply Corollary 4.4 to settle the claim. Let

$$\lambda = \sqrt{\frac{\theta^2 D}{(k + 1)^2 C}},$$

and $\mathcal{F}_j = KD\theta^2/(2\lambda)^j$ for $j = 0, 1, 2$, where K is a properly chosen constant. Recall that $\theta = (C/D)^{1/k} \log^{c/k} n)$ and $k \geq 3$, so provided that $D/C \geq \log^c D$ for an appropriate constant c, we have $\lambda \gg \log D$. For K sufficiently large, the conditions of Corollary 4.4 are met and it implies

$$Pr(X \geq K\theta^2 D) \leq Pr(|X - \mathbb{E}(X)| \geq \tfrac{1}{2} K\theta^2 D) \leq e^{-\lambda} \leq e^{-5 \log D},$$

which is our desired bound. This completes the proof of the claim. ∎

At this point, we hope that the reader has obtained an idea of how polynomials arise,

and of how our results can be applied. The really fascinating fact here is how well the concentration result captures the essence of the problem. First, it tells how we could use the bound C on the codegrees. More importantly, it also tells that why this parameter is the one to look at.

There are several technical tricks one needs to use in order to make the above sketch a complete proof. However, the role of polynomials remains crucial. The interested reader is referred to [79] for the complete proof.

The idea presented above can be further refined to obtain a more general result. As we have seen, we need to terminate the process once the degrees of the hypergraph more or less match its maximum codegree. The reason is that at this point, the expectation of a partial derivative of X becomes too large compared to the expectation of the polynomial. Moreover, in principle, there are hypergraphs with large codegree which do not have any large matching.

The source of further improvement is the following simple observation: If we repeat the analysis of $D(v)$ to $codeg(u, v)$, we can see that $codeg(u, v)$ also decreases until it matches the maximum triple codegree (the maximum number of edges containing the same three vertices). In general, denote by $C_j(H)$ the maximum number of edges in H sharing a set of j vertices, it is possible to show that for any j, $C_j(H)$ decreases, given that $C_{j+1}(H)$ satisfies a proper bound. Thus, if we have bounds on the codegrees of not only 2, but also of 3, 4, ..., s vertices, then we may be able to use them to obtain better estimate of $\mathcal{U}(H)$. The following theorem, proved in [79], quantified this intuition.

Theorem 8.5. *Let H be a $(k+1)$-uniform, D-regular hypergraph on n vertices. Assume that for some $s \le k+1$ there are $D_1 = D \ge D_2 \ge \cdots \ge D_s > 0$ and $x > 0$ such that*

(1) For every $j < s$, $C_j(H) \le D_j$,
(2) For every $j \ge s$, $C_j(H) \le D_s$,
(3) $x^3 \le D_j/D_{j+1}$ for all $j \le s-1$,
(4) $x^{k-s+2} \le D_{s-1}/D_s$.

Then $\mathcal{U}(H) = \tilde{O}(x^{-1}n)$.

It is a routine to verify that Theorem 8.5 implies Theorem 8.2. In several applications (such as Erdős and Hanani's problem), where the higher codegrees are easy to compute, Theorem 8.5 does provide a significant improvement upon Theorem 8.2. Due to space limitation, we do not discuss this matter here and refer the interested reader to [79] for more details.

In a recent paper by N. Alon, B. Bollobás, J. H. Kim, and the present author [4], we extended Theorem 8.2 in a different way by allowing k to tend to infinity together with n. This extension also has a more exact log term in the bound compared to Theorem 8.2. More interestingly, it has few quite surprising and nontrivial geometrical applications. For instance, we used it to obtain an asymptotic answer to the following question: How many lines do one need to separate n random points dropped in the unit square?

It also seems that Theorem 8.5 holds as k tends to infinity with n. Such an extension is motivated by another geometrical problem (work in progress with J. H. Kim and B. Sudakov) and details will appear in a future paper.

8.3. List Coloring of Locally Sparse Graphs

For a graph G, $d(G)$ denotes the maximum degree in G. In this subsection, we always assume that $d(G)$ is sufficiently large and the asymptotic notations is used under the assumption that $d(G) \to \infty$.

Given a graph G, the choice number of G is defined as follows. Assign to each vertex v in the graph a list L_v of k colors (different vertices may have different lists), a list coloring is a coloring in which every vertex is colored by a color from its own list. The choice number (or list chromatic number) $\chi_l(G)$ of G is the least number k such that there exists a proper list coloring for every assignment of lists of size k to the vertices. If we require that all the lists are the same, then we obtain the classical definition of the chromatic number.

The choice number was introduced by Erdős, Rubin, and Taylor [22] and independently by Vizing [71], as a natural extension of the chromatic number. The problem of bounding the choice number, using structural properties of the graph, becomes an exciting research topic in the last ten years, leading to many fascinating results and questions (see [3, 69] for surveys).

Despite the similarity in their definitions, the choice number and the chromatic number differ at a very crucial point. Consider a graph G on the vertex set V and partition V arbitrarily into V' and V''. Let G' and G'' be the induced subgraphs spanned by V' and V'', then $\chi(G') + \chi(G'') \geq \chi(G)$. In other words, the chromatic number is *subadditive*. This property, unfortunately, does not hold for the choice number. For instance, the complete bipartite graph $K_{n,n}$ has list chromatic number $\Theta(\log n)$ (see [22, 3]). It is thus clear that to bound the choice number by a number K it is not sufficient to find K independent sets to cover the vertex set. Therefore, one needs to find another way to upper bound the choice number.

A natural approach is to use the greedy algorithm. With the standard greedy algorithm, it is simple to show that $\chi_l(G) \leq d(G) + 1$. Consider a step of the greedy algorithm when it arrives at a vertex v. The algorithm looks at the neighbors of v, and color v by a color not used in this neighborhood. In the worst case, the neighbors of v could have used up to $d(G)$ colors (for instance, if the neighbors of v form a clique and all of them are exposed prior to v). This worst case forces the list of v to have at least $d(G) + 1$ colors. On the other hand, if the neighbors of v do not span too many edges, then it seems plausible that a color could be used several times and probably this fact could be exploited to reduce the size of the list of v. Therefore, one would hope that if the graph is locally sparse, then a bound significantly better than $d(G) + 1$ might hold.

Recently, this idea has been worked out at different levels in several papers [38, 51, 35, 74, 79], improving the trivial bound $d(G) + 1$. Most proofs made use of the semi-random method. In [51], Molloy and Reed showed that for every positive constant ε, there is a positive constant δ so that if the neighborhood of any points has at most $d(G)^2(1 - \varepsilon)/2$ edges then $\chi_l(G) \leq d(G)(1 - \delta)$. A stronger bound was obtained when the graph is triangle free, that is, every neighborhood is empty. In a beautiful paper [35], Johansson showed:

Theorem 8.6. *If G is triangle free then $\chi_l(G) = O(d(G)/\log d(G))$.*

Johansson's theorem solves the first case of the famous Brooks' conjecture on coloring K_r-free graphs and strengthens an earlier theorem of Kim [38], which proved the same

bound under a stronger assumption. In a recent paper [80], we managed to extend Johansson's result to the following:

Theorem 8.7. *Given a graph G and a number $f > 1$ such that in any neighborhood subgraph of G, the number of edges is at most $d(G)^2/f$. Then $\chi_l(G) = O(d(G)/\log f)$. This bound is sharp up to a constant factor.*

Theorem 8.7 was inspired by a result of Alon, Krivelevich, and Sudakov [6], who proved the same bound for the chromatic number. However, their proof used the subadditivity and cannot be extended for the list-chromatic number.

Theorem 8.7 implies that if for any edge, the two end points have at most $d(G)^{1-\varepsilon}$ common neighbors, then $\chi_{\ldots}(G) = O(d/\log d)$. This result was proved earlier in [74], motivated a question from Theoretical Computer Science. As an application, let us consider the random graph $G(N, p)$ with $p = O(N^{-\varepsilon})$; this random graph satisfies the assumption almost surely. Indeed, in expectation, the number of common neighbors of any two points is $O(N^{1-2\varepsilon})$. Thus, a.s., this number is of order $O(N^{1-2\varepsilon} + \log N)$. Moreover, $d(G(N, p)) = O(Np)$ a.s. Therefore, one could conclude that $\chi_l(G(N, p)) = O(Np/\log(np)) = O(\chi(G(N, p))$, for any $p = N^{-\varepsilon}$, where ε is a positive constant less than 1 (the last equality is a well known fact in the theory of random graph [11, 34]. So we can conclude that for these p, the choice number and the chromatic number of a random graph are of the same order of magnitude, a rather interesting fact given that for several simple graphs (such as $K_{n,n}$) these two quantities are rather far apart. A result of the same nature but with a smaller range of p was obtained in [5] and together they motivate further investigation on the choice number of random graphs and hypergraphs. Several new and exciting results obtained recently in this direction can be found in [7, 44, 74, 75, 80, 45, 46].

Another application of Theorem 8.7 involves the strong chromatic index. Given a graph G, construct a graph $L_1(G)$ as follows: The vertex set of $L_1(G)$ is the edge set of G, and two vertices are adjacent if the corresponding edges have graph distance at most 1. The strong chromatic index of G is the chromatic number of $L_1(G)$ and can be interpreted as the minimum number of induced matchings one needs to cover G. Erdős and Nešetřil conjectured that

$$\chi(L_1(G)) \leq 5d(G)^2/4$$

(see [26] for the history of this conjecture).

It seems plausible that the bound Erdős–Nešetřil conjecture also holds for χ_l. Theorem 8.7 confirmed that this is the case when G is locally sparse. It is easy to show that if any two vertices in G have at most $d(G)/g$ common neighbors (for any $1 < g \leq \Delta$), than any neighborhood in $L_1(G)$ has at most $O(d(G)^3 + d(G)^4/g)$ edges. Thus, it follows from Theorem 8.7 that:

Corollary 8.8. *If every two vertices in G has at most $d(G)/g$ common neighbors, then*

$$\chi(L_1(G)) \leq \chi_l(L_1(G)) = O(d(G)^2/\log g).$$

Corollary 8.8 strengthens an earlier bound of Faudree, Gyárfás, Schelp, and Tuza [26] and generalizes a result Madhian [49] on $\chi_l(L_1(G))$.

The proof of Theorem 8.7 is more complicated than that of Theorem 8.2, but again it is based on our combination of the semi-random method and the polynomial method. The basic idea is to color the graph in a random manner in several steps. In each step, we generate randomly a small set of colors from the union of the lists. Each color has a tag attached to it indicating the list it comes from. A particular color, say Red, may appear with different tags. If a color c with tag v is chosen, and no other c color was chosen with a tag adjacent to v, then we can color v with c. Having done this, a small subset of vertices get colored. The colors chosen, but not used are discarded. In order to avoid conflicts in future steps, we look at each uncolored vertex v, and remove a color c from its lists, if c is already used to color a neighbor of v. Thus, after each step, the graph shrinks and the lists also shrink. The key observation is that if the small set of colors is properly generated, then with positive probability, after the first few steps, the graph will shrinks faster than the lists and then at some point, each list will contain more colors than the maximum degree of the remaining graph. At this point, we end the random process and finish the coloring using the trivial greedy algorithm.

Similar to the proof of Theorem 8.2, the essential part of this proof is to show that the lists and the degrees of the graph shrink at the desired speeds. The task of controlling the degrees was done with the polynomial method, in a way similar to the proof presented in the previous subsection. The degrees of the vertices can be approximated by polynomials of low degrees, and our concentration results finish the job. We refer the interested reader to [80] for details.

8.4. Segre's Conjecture in Finite Geometry

Let us end this paper with our favorite, and perhaps, deepest application. This application essentially settles a long-standing problem in finite geometry, posed by Segre in the late 1950s.

A finite projective planes of order q consists of a set of $q^2 + q + 1$ points and a set of $q^2 + q + 1$ lines, each line contains exactly $q + 1$ points and every two points lie in exactly one line. From this definition, it is easy to derive that every two lines intersect in exactly one point, and each point is contained in $q + 1$ lines.

The most important plane is the so-called Galois plane, $PG(2, q)$, constructed as follows. Consider a 3-dimensional vector space V over the finite field $GF(q)$. The points and the lines of $PG(2, q)$ are the 1- and 2-dimensional subspaces of V, respectively. A line l contains a point p if p is a subspace of l. It is known that there are infinite number of projective planes not isomorphic to $PG(2, q)$, as $q \to \infty$. For more information about these planes, we refer to [14, 32].

In the 1950's, B. Segre (see [59, 60]), one of the founders of the italian school of finite geometry, introduced the notion of arcs. An *arc* in a finite projective plane is a set of points with no three on a line. Maximal arcs under the set inclusion are called complete arcs. A line containing two points of an arc is called a *secant.* By definition, an arc is complete if and only if its secants cover the whole plane. Since Segre's introduction, the study of complete arcs has become a main research topic in finite geometry (see, for instance, the survey of Szőnyi [65] or [32] and their references).

One of the key problems concerning complete arcs is to determine the minimum size

of a complete arc in a given plane. Given a plane P of order q, we denote by $n(P)$ the minimum size of a complete arc in P.

Shortly after Segre's introduction, two other Italian geometers, Lunelli and Sce [48], gave a lower bound $(2q)^{1/2}$ for $n(P)$. Their argument is very simple: Since the union of all secants of a complete arc covers all $q^2 + q + 1$ points of a plane and each secant (which is a line) covers $q + 1$ points, a complete arc must have at least $(q^2 + q + 1)/(q + 1) \geq q$ secants. To have q secants, at least $(2q)^{1/2}$ points are required. The only improvement on this trivial lower bound obtained in the last fourty years was due to Blokhuis [10] and Ball [9], who used a nontrivial algebraic consideration to improve Lunelli and Sce's bound to $(3q)^{1/2}$ for $PG(2, q)$, given q is a prime or the square of a prime, respectively. It is thus widely believed that $q^{1/2}$ is close to the truth. However, this conjecture resisted a number of attacks by several geometers for more than forty years.

Not totally accidentally—constructing a small complete arc is indeed a very difficult task. To illustrate this, let us point out that it is already quite involved to construct a complete arc having ϵq points, for a small constant ϵ. The first important result in this direction was a construction of a complete arc of size roughly $q/3$ in the Galois plane given by Abatangelo [1], whose proof made use of a deep theorem of Weil in algebraic geometry. Korchmáros [42] improved the bound to $q/4$ by similar arguments. For bounds better than $\Omega(q)$, one needs more sophisticated algebraic techniques which have been developed in a sequence of papers [47, 72, 73, 58, 65, 66]. The best construction for the Galois plane $PG(2, q)$ due to Szőnyi [66] yields $n(PG(2, q)) \leq cq^{3/4}$. This bound is still far from $q^{1/2}$. For a general plane, nothing better than a trivial upper bound $O(q)$ was known prior to 1998.

In 1998, J.H. Kim and the present author achieved a significant improvement on the upper bound of $n(P)$ [41]. We were able to prove, using the combination of the semirandom method and the polynomial method, that Lunelli and Sce's lower bound is at most a polylogarithmic term from the truth. Furthermore, our result holds for any plane of order q, with no restriction to the Galois plane.

Theorem 8.9. *There is a positive constant c such that the following holds. Every projective plane of order q contains a complete arc of size at most $q^{1/2}\log^c q$.*

The proof of Theorem 8.9 is fairly long (50 pages) and truly complicated. The semirandom method was used to produce an arc of size $O(q^{1/2})$. This arc is not yet complete, but its secants cover all but $O(q^{1/2}\log^c q)$ points. The existence of such an arc immediately implies the theorem. This critical near-complete arc was built in several steps, basically as follows. Star with the empty set. At each step we add few (random) points to the current arc, and remove from the plane all points covered by the secants of the new arc. The essential part of the proof is, again, to show that after each step, several parameters (such as the number of surviving points on a line) of the remaining structure behave almost as their expectations predict. In other words, we need to show that these parameters, as outputs of our random process, are strongly concentrated around their means. Due to the complexity of the algorithm, at the time being, no classical tools were powerful enough to carry out the task. This motivated us to develop our own tool and this was the way our first result on polynomials, Theorem 4.1, was born.

APPENDIX

Proof of Lemma 5.1. Consider a polynomial Y of degree k. For all $1 \leq i \leq n$, we can write Y as follows

$$Y(t) = \sum_{j=0}^{k} t_i^j Y_{i,j}(t),$$

where the polynomials $Y_{i,j}(t)$, $j = 0, 1, \ldots, k$, do not depend on t_i. Set $\bar{Y}_{i,j}(t) = \mathbb{E}(Y_{i,j}|t_1, \ldots, t_{i-1})$. It is easy to verify that with $p_{i,j} = \int_I t_i^j dt_i$

$$W_i(t) = \sum_{j=1}^{k} \bar{Y}_{i,j}(t)$$

$$W(t) = 2 \sum_{i=1}^{n} \sum_{j=1}^{k} p_{i,j} \bar{Y}_{i,j}(t).$$

(A) We write $X(t) \prec X'(t)$ if $X'(t) - X(t)$ is nonnegative for all t. Since Y is a positive polynomial,

$$\sum_{m=1}^{k} Y_{i,m} \prec \sum_{m=1}^{k} \frac{1}{m!} \frac{\partial^m Y}{\partial^m t_i}.$$

Consequently, for any set A not containing i,

$$\sum_{m=1}^{k} \partial_A Y_{i,m} \prec \sum_{m=1}^{k} \frac{1}{m!} \partial_{i^{[m]} \cup A} Y,$$

where $i^{[m]}$ means that the index i has multiplicity m. It follows that

$$\mathbb{E}_j(W_i) \leq \mathbb{E}_j\left(\sum_{m=1}^{k} Y_{i,m} \right) \leq \mathbb{E}_{j+1}(Y) \sum_{m=1}^{k} \frac{1}{m!} \leq 2\mathbb{E}_{j+1}(Y).$$

(B) The proof is similar to the previous one, using the relation

$$M_j(W_i) \leq M_j\left(\sum_{m=1}^{k} Y_{i,m} \right).$$

(C) First notice that

$$\mathbb{E}(\partial_A(p_{i,m} \bar{Y}_{i,m})) \leq \mathbb{E}(\partial_A(t_i^m Y_{i,m})).$$

Indeed, if A contains i, then the left-hand side is 0. Otherwise the inequality holds since $\mathbb{E}(t_i^m) = p_{i,m}$ by definition. Summing up over $i = 1, \ldots, n$ and $m = 1, \ldots, k$, we obtain

$$\mathbb{E}(\partial_A W) = 2 \sum_{i=1}^{n} \sum_{m=1}^{k} \mathbb{E}(\partial_A(p_{i,m}\bar{Y}_{i,m})) \leq 2 \sum_{i=1}^{n} \sum_{m=1}^{k} \mathbb{E}(\partial_A(t_i^m Y_{i,m})).$$

On the other hand, since Y is positive and has degree k

$$\sum_{i=1}^{n} \sum_{m=1}^{k} t_i^m Y_{i,m} \prec kY,$$

which implies that

$$\mathbb{E}(\partial_A(W)) \leq 2k\mathbb{E}(\partial_A Y).$$

This completes the proof.

(D) The proof of (D) is similar to that of (C) and omitted. ∎

ACKNOWLEDGMENTS

We would like to thank the referee for his very careful reading. We would also like to thank A. Frieze and S. Janson for useful comments and S. Janson and A. Ruciński for sending us their paper [33].

REFERENCES

[1] V. Abatangelo, A class of complete $((q + 8)/3)$-arcs of $PG(2, q)$, with $q = 2^h$ and $h > 6$ even, Ars Combinat 16 (1983), 103–111.

[2] M. Ajtai, J. Komlós, and E. Szemerédi, A dense infinite Sidon sequence, Eur J Combinat 2 (1981), 1–11.

[3] N. Alon, Restricted colorings of graphs, Survey in combinatorics, Proc 14th British Combinatorial Conf, London Mathematical Society Lecture Notes Series 187, K. Walker (Editor), Cambridge University Press, Cambridge, 1993, pp. 1–33.

[4] N. Alon, B. Bollobás, J. H. Kim, and V. H. Vu, Economical covers and geometric applications, submitted.

[5] N. Alon, J. H. Kim, and J. Spencer, Nearly perfect matching in regular simple hypergraphs, Isr J Math 100 (1997), 171–187.

[6] N. Alon, M. Krivelevich, and Sudakov, List coloring of random and pseudo-random graphs, submitted.

[7] N. Alon, M. Krivelevich, and Sudakov, Coloring graphs with sparse neighborhoods, JCT Ser B 77(1) (1999), 73–82.

[8] N. Alon and J. Spencer, The probabilistic methods, 2nd edition, Wiley, New York, 2000.

[9] S. Ball, On small complete arcs in a finite plane, Discrete Math 174(1–3) (1997), 29–34.

[10] A. Blokhuis, Polynomials in finite geometry and combinatorics, Survey in combinatorics, London Mathematical Society Lecture Notes Series 187, K. Walker (Editor), Cambridge University Press, Cambridge, 1993, pp. 35–52.

[11] B. Bollobás, Random graphs, Academic Press, London, 1985.

[12] A. Brower, Block designs, Handbook of combinatorics, R. Graham, M. Grötschel, and L. Lovász (Editors), North-Holland, Amsterdam, 1995, Chap. 14.

[13] S. L. G. Choi, P. Erdős, and M. Nathanson, Lagrange's theorem with $N^{1/3}$ squares, Proc Am Math Soc 79 (1980), 203–205.

[14] P. Dembowski, Finite geometries, Springer-Verlag, New York, 1968.

[15] P. Erdős, Problems and results in additive number theory, Coll Théory de Nombre (CBRM), Bruxelles, 1956, pp. 127–137.

[16] P. Erdős and H. Hanani, On a limit theorem in combinatorial analysis, Publ Math Debrecen 10 (1963), 10–13.

[17] P. Erdős and L. Lovász, Problems and results on 3-chromatic hypergraphs and some related questions, Infinite and finite sets, A. Hajnal et al. (Editors), North-Holland, Amsterdam, 1975, pp. 609–628.

[18] P. Erdős and M. Nathanson, Largange's theorem and thin subsequences of squares, Contribution to probability, J. Gani and V. K. Rohatgi (Editors), Academic Press, New York, 1981, pp. 3–9.

[19] P. Erdős and A. Rényi, Additive properties of random sequences of positive integers, Acta Arith 6 (1960), 83–110.

[20] P. Erdős and A. Rényi, One the evolution of random graphs, Publ Math Inst Hung Acad Sci 5 (1960), 17–61.

[21] P. Erdős and R. Rado, Intersection theorems for systems of sets, J London Math Soc 35 (1960), 85–90.

[22] P. Erdős, A. L. Rubin, and H. Taylor, Choosability in graphs, Proc West Coast Conf Combinatorics, Graph Theory and Computing, Congressus Numerantium XXVI, 1979, pp. 125–157.

[23] P. Erdős and P. Tetali, Representations of integers as sum of k terms, Random Struct Alg 1 (1990), 245–261.

[24] P. Erdős and P. Turán, On a problem of Sidon in additive number theory and some related problems, J London Math Soc 16 (1941), 212–215.

[25] P. Erdős, A. Sárközy, and T. Sós, Problems and results on additive properties of general sequences III, Stud Sci Math Hung 22 (1987), 53–63.

[26] R. Faudree, A. Gyárfás, R. Schelp, and Zs. Tuza, The strong chromatic index of graphs, Ars Combinat 29-B (1990), 205–211.

[27] P. Frankl and V. Rödl, Near perfect covering in graphs and hypergraphs, Eur J Combinat 6 (1985), 317–326.

[28] A. Frieze, On small subgraphs of random graphs, Proc Random Graphs, Poznan, 1989, pp. 67–90.

[29] D. Grable, More than nearly perfect packings and partial designs, Combinatorica 19 (1999), 221–239.

[30] D. Grable, Nearly perfect hypergraph packing in NC, Inform Process Lett 60 (1996), 295–299.

[31] H. Halberstam and K. F. Roth, Sequences, Springer-Verlag, New York, 1983.

[32] J. W. P. Hirschfeld, Projective geometries over finite fields, Claredon Press, Oxford, 1971.

[33] S. Janson and A. Ruciński, The deletion method for upper tail estimates, preprint.

[34] S. Janson, T. Łuczak, and A. Ruciński, Random graphs, Wiley, New York, 2000.

[35] A. Johansson, An improved upper bound on the choice number for triangle free graphs, Technical report, Dimacs.

[36] J. Kahn, One some hypergraph problems of P. Erdős and the asymptotics of matchings, covers and colorings, The mathematics of P. Erdős, R. Graham and J. Nesetril (Editors), Springer-Verlag, New York, 1997.

[37] J. Kahn, Asymptotically good list-colorings, J Combinat Theory Ser A 73 (1996), 1–59.

[38] J. H. Kim, On Brooks' theorem for sparse graphs, Combinat Probab Comput 4 (1995), 97–132.

[39] J. H. Kim, The Ramsey number $R(3, t)$ has order of Magnitude $t^2/\log t$, Random Struct Alg 7 (1995), 173–207.

[40] J. H. Kim and V. H. Vu, Concentration of multivariate polynomials and its applications, Combinatorica 20 (2000), 417–434.

[41] J. H. Kim and V. H. Vu, Small complete arcs in projective planes, submitted, http://www.math.ucsd.edu/vanvu/.

[42] G. Korchmáros, New example of complete k-arcs in $PG(2, q)$, Eur J Combinat 4 (1983), 329–334.

[43] A. Kostochka and V. Rödl, Partial Steiner systems and matchings in hypergraphs, Random Struct Alg 13 (1997), 335–347.

[44] M. Krivelevich, The choice number of dense random graphs, Combinat Probab Comput 9 (2000), 19–26.

[45] M. Krivelevich and V. H. Vu, The weak choice number of random hypergraphs, J Combinat Theory Ser B, to appear.

[46] M. Krivelevich, B. Sudakov, V. H. Vu, and N. Wormald, Random regular graphs of high degree, Random Struct Alg 18 (2001), 346–363.

[47] L. Lombardo-Radice, Sui problema dei k-archi completi di $S_{2,q}$, Boll Un Mat It 11 (1956), 178–181.

[48] L. Lunelli and M. Sce, Considerazioni arithmetiche e risultati sperimentali sui $\{K; n\}_q$ archi, Ist Lombardo Accad Sci Rend A 98 (1964), 3–52.

[49] M. Madhian, The strong chromatic index of graphs, M.Sc. Thesis, Department of Computer Science, Toronto University, 2000.

[50] C. McDiarmid, Concentration, Probabilistic method for algorithmic discrete mathematics, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed (Editors), Springer-Verlag, New York, 1997, pp. 195–248.

[51] M. Molloy and B. Reed, A bound on the strong chromatic index of a graph, J Combinat Theorey, Series B, 69, 1997, no 2, 103–109.

[52] M. Molloy and B. Reed, A bound on the total chromatic number, Combinatorica 18 (1998), 241–280.

[53] M. Nathanson, Waring's problem for sets of density zero, Analitic number theory, M. Knopp (Editor), Lecture Notes in Mathematics 899, Springer-Verlag, New York, 1980, pp. 302–310.

[54] N. Pippenger, unpublished.

[55] N. Pippenger and J. H. Spencer, Asymptotic behaviour of the chromatic index for hypergraphs, J Combinat Theory Ser A 51 (1989), 24–42.

[56] V. Rödl, On a packing and covering problem, Eur J Combinat 5 (1985), 69–78.

[57] A. Ruciński, When are small subgraphs of a random graph normally distributed?, Probab Related Fields 78 (1988), 1–10.

[58] R. Schoof, Non-singular plane cubic curve s over finite fields, J Combinat Theory Ser A 46 (1987), 183–211.

[59] B. Segre, Le geometrie di Galois, Ann Mat Pura Appl 48 (1959), 1–97.

[60] B. Segre, Introduction to Galos geometries (ed. J. W. P. Hirschfeld), Mem Accad Naz Lincei 8 (1967), 133–263.

[61] S. Shelah and J. Spencer, Zero–one law for sparse random graphs, J Am Math Soc 1 (1988), 97–115.

[62] J. Spencer, Counting extensions, J Combinat Theory Ser A 55 (1990), 247–255.

[63] J. Spencer, Four squares with few squares, Number theory, New York Seminar 1991–1995, D. V. Chudnovsky et al. (Editors), Springer-Verlag, New York, 1996, pp. 295–297.

[64] M. Steel, Probability theory and combinatorial optimization, Conference Board of the Mathematical Sciences 69, Philadelphia, 1997.

[65] T. Szőnyi, Complete arcs in $PG(2, q)$, Quad Sem Geom Comb Univ Roma ("La Sapienza"), 94 (1989), 15–29.

[66] T. Szőnyi, Arcs, caps, codes and 3-independent subsets, Giornate di geometrie combinatoire, G. Faina and G. Tallini (Editors), Università di Perrugia, 1993, pp. 57–80.

[67] M. Talagrand, A new look at independence, Ann Probab 24 (1996), 1–34.

[68] M. Talagrand, Concentration of measures and isoperimetric inequalities in product spaces, Publ Math IHES 81 (1996), 73–205.

[69] Z. Tuza, Graph coloring with local constrains—a survey, Discuss Math Graph Theory 17(2) (1997), 161–228.

[70] R. C. Vaughan, The Hardy–Littlewood method, Cambridge University Press, Cambridge, 1981.

[71] V. G. Vizing, Coloring the vertices of a graph in prescribed colors (in Russian), Diskret Anal No. 29, Met Diskret Anal Teor Kodov Shem 101 (1976), 3–10.

[72] J. F. Voloch, On the completeness of certain plane arcs II, Eur J Combinat 11 (1990), 491–496.

[73] J. F. Voloch, Complete arcs in Galois planes of non-square order, Advance in finite geometries and designs, J. W. P. Hirschfeld, D. R. Huges, and J. A. Thas (Editors), Oxford University Press, Oxford, 1991, pp. 401–406.

[74] V. H. Vu, On some degree conditions which guarantee the upper bound of chromatic (choice) number of random graphs, J Graph Theory 31 (1999), 201–226.

[75] V. H. Vu, On the choice number of random hypergraphs, Combinat Probab Comput 9 (2000), 79–95.

[76] V. H. Vu, A large deviation result on the number of small subgraphs of a random graph, Combinat Probab Comput, to appear, http://www.math.ucsd.edu/vanvu/.

[77] V. H. Vu, On a refinement of Waring's problem, Duke Math J 105 (2000), 107–134.

[78] V. H. Vu, On the concentration of multivariate polynomials with small expectation, Random Struct Alg 16 (2000), 344–363.

[79] V. H. Vu, New bounds on nearly perfect matchings in hypergraphs: Higher codegrees do help, Random Struct Alg 17 (2000), 29–63.

[80] V. H. Vu, On the list chromatic number of locally sparse graphs, Combinat Probab Comput, to appear, http://www.math.ucsd.edu/vanvu/.

[81] E. Wirsing, Thin subbases, Analysis 6 (1986), 285–308.

[82] J. Zöllner, Der Vier-Quadrate-Satz und ein Problem von Erdős and Nathanson, Ph.D thesis, Johannes Gutenberg-Universität, Mainz, 1984.

[83] J. Zöllner, Über eine Vermutung von Choi, Erdős and Nathanson, Acta Arith 45 (1985), 211–213.