Report on the thesis
INEQUALITIES FOR SUMS OF RANDOM VARIABLES: A COMBINATORIAL PERSPECTIVE
by Matas Šileikis

Prof. dr hab. Paweł Hitczenko
Department of Mathematics, Drexel University

1. Topics

The topics covered in the proposed thesis lie on the border of probability theory and combinatorics, and more specifically, graph theory. The author presents a number of results concerning estimates for the probabilities that a sum of random variables falls in a given interval in $\mathbb{R}$ (either bounded or unbounded). The common feature is that the summands are either independent or posses a high degree of independence. The most important and most common situation concerns the intervals of the form (x, ∞) (upper tail) or $(-\infty, x)$ (lower tail); here x is a real number.

Results of this type have preoccupied probabilists for a long time and are central to many further developments in probability theory, mathematics, and other disciplines of science that use probabilistic techniques. One such area is a relatively new sub–field of graph theory, namely random graphs. Various aspects of graph theory have found applications in theoretical computer science and have become an integral part of this rapidly expanding area of research.

Major part of the thesis concerns estimates for the tail probabilities (upper and lower) for a sum of a specific indicator random variables. The sum under consideration counts the number of copies of a fixed deterministic graph in a large random graph. 'Random graph' refers here to the most heavily studied in the literature model the so–called Erdös – Rényi $\mathbb{G}(n, p)$ model introduced over fifty years ago. In this part of the thesis the author partially substantiates a very recent conjecture of DeMarco and Kahn concerning the bounds on the probability that the number of copies of a given graph in a large random graph exceeds a fixed multiple of the expected number of such graphs.

Two other parts of the thesis are devoted mainly to the concentration inequalities for weighted sums of independent Bernoulli random variables and they fall firmly into classical probability theory (although some have an interesting combinatorial flavor).

2. Results

Original results are presented in Chapters 2, 3 and 4. However, as the author himself states, the content of Chapter 2 is primarily expository. It concerns the concentration inequalities for the Lipschitz functions on a discrete cube $\{0, 1\}^{\mathbb{N}}$ equipped with the product measure. Results that are formally new in this chapter are minor improvements over existing results. It should be noted that this is a very classical and well developed area that has been studied for a long time. Therefore, it is unreasonable to expect major new results.

Results presented in Chapter 3 concern applications of combinatorial reasoning (related to combinatorics of ordered sets) to provide elegant and sharp bounds on the probabilities of linear combinations (with suitably balanced coefficients) of independent, symmetric Bernoulli random variables falling into a fixed interval or an interval of fixed and finite length. Symmetric Bernoulli random variables are random variables taking on values ± 1, each with probability $1/2$). The bounds are expressed in terms of probabilities involving a simple symmetric random walk (that is just a sum of independent symmetric Bernoulli variables - because of that such probabilities are easy to find in terms of binomial coefficients).

The results are interesting and were published in the *SIAM Journal on Discrete Mathematics* a major journal of very high standard and international reach. It should be emphasized while the paper is co-authored with two other researchers, as far as I could determine they are the candidate's peers as far as the stage of their professional careers is concerned (one is a PhD candidate himself and the other seems to be either a recent PhD or a candidate, too). Thus, it seems reasonable to assume that all three of the the co–authors contributed more or less equally to the paper, both on technical as well as on the conceptual level. Having said that, let me add that I would have liked to see a clarifying statement to that effect.

The most valuable in my view are results presented in Chapter 4. Some (but by no means all) of the results presented in that part of a thesis were published in a paper (with the candidate being the sole author) in *The Electronic Journal of Combinatorics*, a major venue for dissemination of knowledge in combinatorics. The journal has an international reach and global influence.

The problem concerns the count of the number of appearances of a copy of a specific graph in a large random graph. To make things more precise, one considers the Erdös–Rényi model of a random graph $\mathbb{G}(n,p)$ on n vertices, with each of the possible $\binom{n}{2}$ edges put in place with probability p, independently of all other edges. One then chooses a specific graph G and lets X_G be the number of (isomorphic to G) copies that $\mathbb{G}(n,p)$ contains. Of interest is then obtaining the bounds on the probability that X_G deviates significantly from its expected value, $\mathbb{E}X_G$. More precisely, one is interested in estimating (from above and below) the probabilities of the form $\mathbb{P}(X_G > t\mathbb{E}X_G)$ (upper tail) and $\mathbb{P}(X_G < t\mathbb{E}X_G)$ (lower tail). This is considered to be a problem of central interest in the theory of random graphs and has been a focus of substantial attention by prominent researchers in this area. Despite all the efforts the problem does not have a complete solution. The candidate discusses nicely the history in his thesis, so let me just briefly describe the situation: one is actually facing four separate problems: lower and upper bounds for the lower tail and lower and upper bounds for the upper tail. The problem is further complicated by the fact that one usually assumes that $p = p_n$ is a function of n and gets (or does not get) the results depending on the range of p as a function of n.

The bound on the lower tail was settled by Janson in 1990, but there are only partial results for the upper tail. Based on those DeMarco and Kahn in their recent work conjectured the asymptotic behavior as $n \to \infty$ of the upper tail (or, to be precise on the order of the magnitude of $\log \mathbb{P}(X_G > t\mathbb{E}X_G)$).

The candidate was able to make some progress on that conjecture. His contribution, presented in the last chapter of this thesis is as follows:

- the author confirms that the conjectured quantity is, indeed, the lower bound on the upper tail (in the range of p's not covered by earlier work). He proved it by himself for the so–called strictly balanced graphs and jointly with Janson in full generality.
- for a specific types of graphs G the author provides the upper bound that is consistent with the conjecture of DeMarco–Kahn.

This part of the proposed thesis makes a non–trivial contribution to an important open problem in random graph theory. It is clear from the up to date history that it is a difficult problem that has eluded the experts for more than twenty years now. The progress over that time was incremental and while the results presented in the thesis thesis do not provide a complete solution of the problem they widen a class of graphs for which the conjecture is true, thus bringing us closer to a complete solution.

Researchers who have contributed in the past to the (still only partial) solution include, among others, Chatterjee, Janson, Kahn, Kim, Łuczak, Ruciński, and Vu (this list by itself is a good testimony to the difficulty of the problem). I am convinced that in future discussions of a history of this problem the name of Mr. Šileikis and the results he proved in his thesis will be added to this list.

3. Comments

The thesis is very well and carefully written. I have only a few points, most of them minor. Specific comments are given below.

3.1. General comments.

- My main criticism of the thesis concerns the discussion of the history of the concentration results for Lipschitz functions of independent random variables. The author limits the discussion almost entirely to the martingale methods. While I have some sympathy for that, in my view any serious discussion of a history and background of concentration results for sums of independent random variables, *must* include in a prominent role the body of results obtained (mostly in late nineties) by Talagrand. His work on concentration transformed that area an any discussion omitting the role of his contribution is simply inadequate. (To some extent this is also true about an approach based on analytic inequalities and largely mastered by Ledoux and his collaborators.) While it is true that the author mentions Talagrand, it is done only marginally and does not fairly reflect the contributions to the body of knowledge in this area made by various approaches.
- Some proof techniques used in the chapter about random graphs connect nicely with two preceding chapters on inequalities for sums of random variables. In particular, in the proofs of Theorems 4.8 and 4.9 the author obtains upper bounds on the tail probabilities for the sum of n iid random variables $\binom{B_i}{r}$ (or just its upper bound – the sum of n iid B_i^r random variables), where B_i has a binomial distribution with parameters n and p. While I have not seen inequalities of this type for such random variables in the literature, the topic is *very* classical and there has been a lot of of work on inequalities for sums of iid random variables. As a matter of fact the author follows the most common approach, that is, he obtains a bound on the moment generating function of the sum and then optimizes over an

extra parameter that this procedure allows one to introduce (denoted by h in this case). Therefore, I wish the author had put his proof in a bit wider context and perhaps discussed its relationship to what is normally achievable by this method. In particular, if there are any subtleties as compared to what is normally being done, they would be worth emphasizing.

3.2. **Minor omissions, typographical errors.**

- p. 19, formula (2.30): I believe '$\mathbb{P}_p(A_t) \geq$' should read '$\mathbb{P}_p(A_{t-1}) \geq$'.
- p. 21 line 8 (not counting the header and the title): 'the bound is depends' should read 'the bound depends'.
- p. 25, line 16: 'let numbers' should read 'let the numbers'.
- p. 34, line 23: 'behaves like a sum' should read 'behaves like the tail of a sum'.
- p. 43: there is a double 'that' in a line immediately following the end of a proof of Proposition 4.12.
- p. 51, line 1: the sentence 'Let us show that for every $H \subseteq F$ that $\Psi_{H,R} = n^{\Omega(1)}$.' sounds a bit awkward. Perhaps using 'one has' or 'we have' would be better.
- p. 52_7: 'Spencer' is unnecessarily repeated.
- p. 57, line 3: actually the fact that B_v^r are also identically distributed is used here as well. It would be better to emphasize this.
- p. 60_8: although '$1 - e^{-x} \asymp x$ as $x \to 0$' is formally correct, perhaps $1 - e^{-x} \sim x$ would be more consistent with the notation used throughout the thesis.

4. Conclusion

The results presented in the thesis constitute a valuable contribution to the theory of probability and the theory of random graphs. Some of the results were already published or accepted for publication in the prestigious journals with worldwide readership like *SIAM Journal of Discrete Mathematics* (joint with two other authors) or *Electronic Journal of Combinatorics*.

I have no doubts that the proposed thesis complies with the requirements set forth by the Polish law. I therefore recommend that the thesis is accepted and that the PhD defense of Mr. Matas Šileikis proceeds to its next phase.

Paweł Hitczenko

Warszawa, June 16, 2012