On measure concentration in graph products

Matas ŠILEIKIS
Institute of Mathematics and Informatics
Akademijos 4, LT-08663 Vilnius, Lithuania
e-mail: matas.sileikis@gmail.com

Abstract. Bollobás and Leader [1] showed that among the \( n \)-fold products of connected graphs of order \( k \) the one with minimal \( t \)-boundary is the grid graph. Given any product graph \( G \) and a set \( A \) of its vertices that contains at least half of \( V(G) \), the number of vertices at a distance at least \( t \) from \( A \) decays (as \( t \) grows) at a rate dominated by \( \Pr(X_1 + \ldots + X_n \geq t) \) where \( X_j \) are some simple i.i.d. random variables. Bollobás and Leader used the moment generating function to get an exponential bound for this probability. We insert a missing factor in the estimate by using a somewhat more subtle technique (cf. [3]).

Keywords: graph product, discrete isoperimetric inequalities, concentration of measure, sums of independent random variables, tail probabilities, large deviations.

1. Introduction and theorem

Consider a finite set \([k]\) consisting of \( k \) elements: \(\{0, 1, \ldots, k-1\}\). We may define various metrics (distances) \( d \) on \([k]\). One of the ways to do that is to consider a graph \( G = (V, E) \) with a vertex set \( V = [k] \) and define the distance \( d(a, b) \), as the length of the shortest path between \( a \) and \( b \). In order to have a finite metric, we will, of course, put a restriction that the graph \( G \) is connected.

If, for example, we choose \( G \) to be a path \( P_k \), i.e., graph with the edge set \( E = \{\{0, 1\}, \{1, 2\}, \ldots, \{k-2, k-1\}\} \) then the resulting metric is the one inherited from the real line with the Euclidean distance. On the other hand, if \( G \) is a complete graph \( K_k \) on \( k \) vertices, consisting of all possible pairs of vertices, then \( d(a, b) = 1 \) if and only if \( a \neq b \). Let us consider a product \([k]^n\) of metric spaces \(([k], d_1), \ldots, ([k], d_n)\) each with the same number of elements but probably distinct metrics \(d_i\). Let us denote elements of \([k]^n\) as \( a = (a_1, \ldots, a_n) \).

It is easy to see that the \( l_1 \)-type metric on \([k]^n\) defined as

\[
d(a, b) = d_1(a_1, b_1) + \cdots + d_n(a_n, b_n)
\]

is indeed a metric. We choose this way of defining a metric on the product space because we can reconstruct a graph on \([k]^n\) by considering a pair \( (a, b) \) an edge if and only if \( d(a, b) = 1 \).

If metrics \( d_i \) are induced by graphs \( G_i \), we shall refer to the graph reconstructed from the metric \( d \) as the cartesian product of graphs \( G_i, i = 1, \ldots, n \), denoting it \( G = G_1 \times \cdots \times G_n \). We can equivalently define \( G \) by saying that a pair \( (a, b) \) of vertices is an edge whenever there is \( i \) such that \( a_i, b_i \) is an edge in \( G_i \) and \( a_j = b_j \) for all \( j \neq i \).
Consider the example where \( G_i = P_k \). Multiplying a path by itself we obtain so-called \( n \)-dimensional grid graphs. Given a subset of vertices \( A \subset V \) of a graph \( G \) which is not too small (say, has at least \( |V|/2 \) elements), how big is its neighbourhood, i.e., vertices having a neighbour in the set \( A \)? More generally, how many vertices are there at a distance from \( A \) at most \( t \)?

Let us denote \( t \)-neighbourhood of \( A \) as \( A_t := \{ a \in V : d(a, b) \leq t \text{ for some } b \in A \} \).

Given a graph, we are interested in finding a set that has the smallest \( t \)-boundary, it is, determining the quantity

\[
\min_{|A| \geq |V|/2} |A_t|.
\]

(1)

It turns out that in the case of product graphs of high dimension a striking phenomenon (known as concentration of measure) is observed: \( A_t \) is almost all of \( V \) whenever \( t \) is a small proportion of the diameter of \( G \).

We may pose a question from another point of view: given a class of graphs, which one has the slowest growth of \( A_t \), or, seeking a slightly weaker answer, what is a good lower bound for (1)? This was fully answered by Bollobás and Leader [1] in case when the class consists of all \( n \)-fold products of graphs on \( k \) vertices.

Consider, for \( r \geq 0 \), balls around zero \( B_k^{(n)}(r) = \{ a \in [k]^n : \sum_i a_i \leq r \} \).

THEOREM 1 [Bollobás and Leader, [1]]. Let \( G_1, \ldots, G_n \) be connected graphs of order \( k \). Let \( G = \prod_{i=1}^n G_i \) be their product. Suppose \( r \in \{0, 1, 2, \ldots\} \), and \( A \subset V(G) \) is such that \( |A| \geq |B_k^{(n)}(r)| \). Then, for \( t = 0, 1, 2, \ldots \)

\[
|A_t| \geq |B_k^{(n)}(r + t)|.
\]

The lower bound given by Theorem 1 can be interpreted using probability. Let \( X_1, \ldots, X_n \) be independent copies of a random variable \( X \) distributed uniformly over \([k] \):

\[
P(X = j) = 1/k \quad \text{for all } j \in [k].
\]

(2)

Now we can estimate \( |B_k^{(n)}(r + t)| \) by the means of the following representation:

\[
|B_k^{(n)}(r + t)|/k^n = P(X_1 + \cdots + X_n \leq r + t).
\]

(3)

Let

\[
Y_i = X_i - \mathbb{E}X_i, \quad i = 1, 2, \ldots; \quad S_n = Y_1 + \cdots + Y_n.
\]

(4)

Bollobás and Leader [1] estimated the moment generating function \( \exp\{hS_n\} \) by calculating moments of \( S_n \) and then used Chebyshev’s inequality

\[
P(S_n \geq t) \leq \inf_{h > 0} \exp\{h(S_n - t)\}
\]

(5)

to obtain the following statement.
THEOREM 2 [Bollobás and Leader [1]]. Let $G_1, \ldots, G_n$ be connected graphs of order $k$ and let $G = \prod_{i=1}^n G_i$. Suppose $A \subset V(G)$ is such that $|A| \geq |V(G)|/2$. Then, for $t = 0, 1, 2, \ldots$, we have

$$1 - |A_t|/k^n \leq \mathbb{P}\{S_n \geq t\} \leq \exp\left\{-\frac{6t^2}{(k^2 - 1)n}\right\} = \exp\left\{-\frac{t^2}{2n\sigma^2}\right\},$$

(6)

where $S_n$ is the random variable defined in (4) and $n\sigma^2 = \text{Var}S_n$.

Using the Central Limit Theorem we can see that the constant $6/(k^2 - 1)$ in (6) cannot be improved. However, one could expect a bound similar to the right tail of a normal random variable with variance $n\sigma^2$. We show that this is indeed the case.

THEOREM 3. For the random variable $S_n$ defined in (4) and $t \in \mathbb{R}$ we have

$$\mathbb{P}\{S_n \geq t\} \leq c I(t\sigma/\sqrt{n}) \leq \sqrt{\frac{2}{\pi}} c \sqrt{\frac{\sigma}{n}} t \exp\left\{-\frac{t^2}{2n\sigma^2}\right\},$$

(7)

where $I(x) = 1 - \Phi(x)$ is the survival function of a standard normal random variable, $c = 5!e^{3/3} = 5.699\ldots$, and $\sigma^2 = (k^2 - 1)/12 = \text{Var}S_n/n$.

The author conjectures that the constant $c = 5.699\ldots$ can be replaced by a constant $c = 3!e^{3/3} = 4.463\ldots$.

Theorem 3 gives an improvement upon the bound (6) whenever $t$ is of order larger than $\sigma/\sqrt{n}$ which is the case when we set $t$ to be a 'small fixed proportion' of the diameter of the grid graph, namely $t = \epsilon \text{diam}(P^n_k) = \epsilon n(k - 1)$, with an arbitrarily small $\epsilon > 0$.

Proof of Theorem 3. Consider, for any $h < t$, a function $x \mapsto (x - h)_+^5$. As $\mathbb{I}\{x \geq t\} \leq (x - h)_+^5/(t - h)_+^5$, we get

$$\mathbb{P}\{S_n \geq t\} = \mathbb{E}\mathbb{I}\{S_n \geq t\} \leq \inf_{h < t} \frac{\mathbb{E}(S_n - h)_+^5}{(t - h)_+^5}. \quad (7)$$

Applying Lemma 3 and Lemma 1.1 of [2] we conclude the proof.

2. Lemmas and their proofs

Consider a random variable $\tau = \tau(b, \sigma^2)$ which assumes values $\{-b, 0, b\}$, with probabilities

$$\mathbb{P}(\tau = -b) = \mathbb{P}(\tau = b) = \frac{\sigma^2}{2b^2} \quad \text{and} \quad \mathbb{P}(\tau = 0) = 1 - \frac{\sigma^2}{b^2}.$$

LEMMA 1. For any $h \in \mathbb{R}$ we have

$$\mathbb{E}(Y - h)_+^5 \leq \mathbb{E}(\tau - h)_+^5, \quad (8)$$

where $Y$ is a centered discrete uniform random variable on $[k]$ as defined in (4) and $\tau = \tau(\max Y, \text{Var}Y)$. 

\textbf{Proof.} Note that \( Y \) is symmetric and so satisfies the conditions of Lemma 3 of [5] with \( b = \max Y \) and \( \sigma^2 = \text{Var} Y \). Therefore we get that for \( h \in \mathbb{R} \)

\[
\mathbb{E}(Y - h)^5_+ \leq \mathbb{E}(\tau - h)^5_+ .
\] (9)

To prove (8) it suffices to show that \( f(h) = \mathbb{E}(\tau - h)^5_+ - \mathbb{E}(Y - h)^5_+ \geq 0 \). The function \( h \mapsto (x - h)^5_+ \) has the second continuous derivative. Therefore we can differentiate \( f \) under the integral to obtain

\[
f'(h) = -5\mathbb{E}(\tau - h)^4_+ + 5\mathbb{E}(Y - h)^4_+, \quad f''(h) = 20\mathbb{E}(\tau - h)^3_+ - 20\mathbb{E}(Y - h)^3_+ .
\]

It is obvious that \( f(b_1) = f'(b_1) = 0 \). Moreover, \( f \) is convex because from (9) we have \( f'' \geq 0 \). Therefore \( f \geq 0 \).

The following result is probably the essence of the paper.

\textbf{Lemma 2.} Let \( \tau = \tau(b, \sigma^2) \) with \( b \) and \( \sigma \) satisfying \( \sigma^2 / b^2 \geq 1/3 \). Then for all \( h \in \mathbb{R} \) we have

\[
\mathbb{E}(\tau - h)^5_+ \leq \mathbb{E}(\eta - h)^5_+ ,
\]

where \( \eta \) is a normal random variable with mean zero and variance \( \sigma^2 \).

\textbf{Proof.} For simplicity and without loss of generality we may assume that \( b = 1 \), because the general case follows by rescaling. Under this assumption we have that \( \sigma^2 \geq 1/3 \). To prove the lemma it suffices to show that \( \mathbb{E}(\eta - h)^5_+ - \mathbb{E}(\tau - h)^5_+ =: f(h) \geq 0 \).

\textbf{Case 1.} If \( h \geq 1 \), then \( (\tau - h)_+ \equiv 0 \) so \( f \geq 0 \) holds trivially.

\textbf{Case 2.} If \( h \leq -1 \), then

\[
f(h) = \mathbb{E}(\eta - h)^5_+ - \mathbb{E}(\tau - h)^5_+ \geq \mathbb{E}(\eta - h)^5 - \mathbb{E}(\tau - h)^5
\]

\[
= ( - 5h(\eta^4 - 10h^3 \eta^2 - h^5) - (- 5h(\tau^4 - 10h^3 \tau^2 - h^5))
\]

\[
= 5h(- 3\sigma^4 + 2 \cdot 1^4 \cdot \sigma^2 / 2) \geq 0
\]

since \( \sigma^2 \geq 1/3 \) and odd moments of symmetric random variables vanish.

\textbf{Case 3.} \( h \in [-1, 0] \). We may reduce this case to the Case 4 as soon as we show that \( f(-t) \geq f(t) \) for all \( t \in [0, 1] \). Since \( (\eta - t)_+ = (- \eta + t)_- \) is equal in distribution to \( (\eta + t)_- \) (here \( (x)_- = \max\{-x, 0\} \)), and \( \sigma^2 \geq 1/3 \), we have

\[
f(-t) - f(t) = \mathbb{E}(\eta + t)^5_+ - \mathbb{E}(\eta - t)^5_+ - \mathbb{E}(\tau + t)^5_+ + \mathbb{E}(\tau - t)^5
\]

\[
= \mathbb{E}(\eta + t)^5 - (1 + t)^5 \sigma^2 / 2 - t^5 (1 - \sigma^2) + (1 - t)^5 \sigma^2 / 2
\]

\[
= 5t\eta^4 + 10t^3 \eta^2 + t^5 - 5\sigma^2 - 10t^3 \sigma^2 - t^5
\]

\[
= 5t \cdot 3\sigma^4 - 5t\sigma^2 = 5t(3\sigma^4 - \sigma^2) \geq 0.
\]
Case 4. \( h \in [0, 1] \). It is easy to check that function \( f \) restricted to the interval \([0, 1]\) is five times differentiable and its \( k \)-th derivative is 

\[
    f^{(k)}(h) = (-1)^k c_k \left( \mathbb{E}(\eta - h)^{5-k} - \frac{\sigma^2}{2} (1 - h)^{5-k} \right), \quad k = 1, 2, \ldots, 5,
\]

where \( c_k \) are positive constants, and we make a convention \( 0^0 = 1 \).

The following argument is clear if one looks at the graphs of \( f^{(k)} \).

Note that \( f^{(5)}(h) = c_5 \sigma^2/2 - c_5 \mathbb{E}(\eta > h) \), so \( f^{(5)} \) is increasing. By Chebyshev's inequality we have \( \mathbb{P}(\eta > h) \leq \sigma^2/2 \), so \( f^{(5)}(1) \geq 0 \). Consequently, there is a number \( x \in [0, 1] \) such that \( f^{(5)} \leq 0 \) on \([0, x]\) and \( f^{(5)} \geq 0 \) on \([x, 1]\). Therefore \( f^{(3)} \) is concave on \([0, x]\) and \( f^{(3)} \) is convex on \([x, 1]\).

In order to see how the sign of \( f^{(3)} \) varies, we observe that

\[
    f^{(3)}(0) = -c_3 (\mathbb{E}(\eta^+)^2 - \sigma^2/2) = 0, \quad \text{and} \quad f^{(3)}(1) = -c_3 (\mathbb{E}(\eta - 1)^+)^2 < 0.
\]

Consequently, there is some number \( y \in [0, 1] \) such that \( f^{(3)} \geq 0 \) on \([0, y]\) and \( f^{(3)} \leq 0 \) on \([y, 1]\). Therefore, \( f' \) is convex on \([0, y]\) and \( f' \) is concave on \([y, 1]\).

In order to see how the sign of \( f' \) varies we check that

\[
    f'(0) = -5(\mathbb{E}(\eta^+)^4 - \sigma^2/2) = -5(3\sigma^4/2 - \sigma^2/2) \leq 0,
\]

\[
    f'(1) = -5(\mathbb{E}(\eta^+)^4) < 0, \quad \text{and} \quad f''(1) = 20(\mathbb{E}(\eta^+)^3) > 0,
\]

so we see that \( f' \) is negative on \([0, 1]\). Finally, since \( f(1) = \mathbb{E}(\eta - 1)^+ > 0 \), we get that \( f > 0 \) on \([0, 1]\).

**Lemma 3.** Let the random variable \( S_n \) be defined as in 4. Then for any \( h < t \) we have

\[
    \mathbb{E}(S_n - h)^+ \leq \mathbb{E}(Z_n - h)^+,
\]

where \( \eta \) is a centered normal random variable such that \( \text{Var} Z_n = \text{Var} S_n = n\sigma^2 \).

**Proof.** We can write \( Z \) as a sum \( \eta_1 + \cdots + \eta_n \) of i.i.d normal random variables each with mean zero and variance \( \sigma^2 \). We will now use induction on \( n \) to prove (10). For \( n = 1 \) it is equivalent to the combination of Lemmas 1 and 2. Now suppose (10) holds for \( 1, \ldots, n - 1 \). Using the induction hypothesis twice (for \( n - 1 \) and 1), we get

\[
    \mathbb{E}(S_n - h)^+ = \mathbb{E}[\mathbb{E}(Y_1 + \cdots + Y_n - h)^+ | Y_1)]
\]

\[
    \leq \mathbb{E}[\mathbb{E}(Y_1 + \cdots + \eta_n - h)^+ | Y_1)]
\]

\[
    = \mathbb{E}[\mathbb{E}(Y_1 + \cdots + \eta_n - h)^+ | \eta_2, \ldots, \eta_n)]
\]

\[
    \leq \mathbb{E}[\mathbb{E}(\eta_1 + \cdots + \eta_n - h)^+ | \eta_2, \ldots, \eta_n)] = \mathbb{E}(Z_n - h)^+.
\]
References

1. B. Bollobás, I. Leader. Compressions and isoperimetric inequalities. *J. Combin. Theory Ser. A*, 56(1):47–62, 1991.
2. V. Bentkus, N. Kalosha, M. van Zuijlen. On domination of tail probabilities of (super)martingales: explicit bounds. *Liet. Mat. Rink.*, 46(1):3–54, 2006.
3. V. Bentkus. On Hoeffding’s inequalities. *Ann. Probab.*, 32(2):1650–1673, 2004.
4. V. Bentkus, G. Geuze, M. van Zuijlen. Trinomial Laws Dominating Conditionally Symmetric Martingales. Technical report 0514, Department of Mathematics, Radboud University, October 2005.
5. V. Bentkus, G. Geuze, M. van Zuijlen. Unimodality: The Symmetric Case. Technical report 0612, Department of Mathematics, Radboud University, November 2006.

REZIUMĖ

M. Šileikis. Apie mato koncentraciją grafų sandaugoje

Bollobás ir Leader [1] parodė, jog tarp n jungių k-osių eilės grafų sandaugų didžiausią mato koncentraciją turi n-mates gardečės grafas. Jei aibė A turi pusę grafų sandaugos viršūnų, tai viršūnų, esančių nuo A ne arčiau kaip per t, skaičius yra apraštas tikimybe \( \mathbb{P}(X_1 + \cdots + X_n \geq t) \), kur \( X_i \) – tam tikri paprasti n.v.p., atsitiktiniai dydžiai. Bollobás ir Leader naudodami momentų generuojančią funkciją gavo eksponentini įverti. Naudodami kiek subtilesnę techniką (plg. [3]), mes pageriname įvertį eilę, įterpdami trūkstamą daugiklį.

Raktiniais žodžiais: grafų sandauga, diskrečios izoperimetrinės nelygūnės, mato koncentracija, neprikluso-somą atsitiktinių dydžių sumos, uodegų tikimybės, didieji nuokrypiai.