European Journal of Combinatorics 73 (2018) 138–152

Contents lists available at ScienceDirect

European Journal of Combinatorics

journal homepage: www.elsevier.com/locate/ejc

How unproportional must a graph be?

Humberto Naves [a], Oleg Pikhurko [b], Alex Scott [c]

[a] Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55455, USA
[b] Mathematics Institute and DIMAP, University of Warwick, Coventry CV4 7AL, UK
[c] Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK

ARTICLE INFO

Article history:
Received 18 January 2017
Accepted 29 May 2018

ABSTRACT

Let $u_k(G, p)$ be the maximum over all k-vertex graphs F of by how much the number of induced copies of F in G differs from its expectation in the binomial random graph with the same number of vertices as G and with edge probability p. This may be viewed as a measure of how close G is to being p-quasirandom. For a positive integer n and $0 < p < 1$, let $D(n, p)$ be the distance from $p\binom{n}{2}$ to the nearest integer. Our main result is that, for fixed $k \geq 4$ and for n large, the minimum of $u_k(G, p)$ over n-vertex graphs has order of magnitude $\Theta\left(\max\{D(n, p), p(1-p)\}n^{k-2}\right)$ provided that $p(1-p)n^{1/2} \to \infty$.

1. Introduction

An important result of Erdős and Spencer [11] states that every graph G of order n contains a set $S \subseteq V(G)$ such that $e(G[S])$, the number of edges in the subgraph induced by S, differs from $\frac{1}{2}\binom{|S|}{2}$ by at least $\Omega(n^{3/2})$; an earlier observation of Erdős [9] shows that this lower bound is tight up to the constant. More generally, it was shown in [10] that for graphs with density $p \in (\frac{2}{n-1}, 1-\frac{2}{n-1})$, there is some subset where the number of edges differs from expectation by at least $c\sqrt{p(1-p)}n^{3/2}$ (see [4–6] for further results and discussion).

When p is constant, the above results can be equivalently reformulated in the language of graph limits as that the smallest cut-distance from the constant-p graphon to an order-n graph G is $\Theta(n^{-1/2})$. Instead of defining all terms here (which can be found in Lovász' book [21]), we observe that the cut-distance in this special case is equal, within some multiplicative constant, to the maximum over $S \subseteq V(G)$ of $\frac{1}{n^2}\left|2e(G[S]) - p|S|^2\right|$.

E-mail addresses: hnaves@ima.umn.edu (H. Naves), O.Pikhurko@warwick.ac.uk (O. Pikhurko), scott@maths.ox.ac.uk (A. Scott).

https://doi.org/10.1016/j.ejc.2018.05.007
0195-6698/

There are other measures of how close a graph G is to the constant-p graphon, which means measuring how close G is to being p-quasirandom. Here we consider two possibilities, subgraph statistics and graph norms, as follows.

For graphs G and H, we denote by $N(H, G)$ the number of induced subgraphs of G that are isomorphic to H. For example, if $v(H) = k \leq n$, then the expected number of H-subgraphs in the binomial random graph $\mathbb{G}_{n,p}$ (where each pair on the vertex set $[n] := \{1, \ldots, n\}$ is independently included as an edge with probability p) is

$$\mathbf{E}[N(H, \mathbb{G}_{n,p})] = \frac{n(n-1)\ldots(n-k+1)}{|\mathrm{Aut}(H)|}\, p^{e(H)}(1-p)^{\binom{k}{2}-e(H)},$$

where $\mathrm{Aut}(H)$ is the group of automorphisms of H.

Let $k \geq 2$ be a fixed integer parameter. For any graph G on n vertices and a real $0 < p < 1$, let

$$u_k(G, p) := \max\Big\{ \big| N(F, G) - \mathbf{E}[N(F, \mathbb{G}_{n,p})] \big| \ : v(F) = k \Big\}, \tag{1.1}$$

where the maximum is taken over all (non-isomorphic) graphs F on k vertices. The quantity $u_k(G, p)$ measures how far the graph G is away from the random graph $\mathbb{G}_{n,p}$ in terms of k-vertex induced subgraph counts. For example, $u_k(G, p)/n^k$ is within a constant factor (that depends on k only) from the total variational distance between $\mathbb{G}_{k,p}$ and a random k-vertex subgraph of G.

We are interested in estimating

$$u_k(n, p) := \min\{u_k(G, p) : v(G) = n\}, \tag{1.2}$$

the minimum value of $u_k(G, p)$ that a graph G of order n can have. Informally speaking, we ask how p-quasirandom a graph of order n can be.

Clearly, $u_2(n, p) < 1$ and $u_2(n, p) = 0$ if $p\binom{n}{2}$ is integer. In fact, if we denote by $D(n, p)$ the distance from $p\binom{n}{2}$ to the nearest integer, then $u_2(n, p) = D(n, p)$. The problem of constructing pairs (F, p) with $u_3(F, p) = 0$ (such graphs F were called *p-proportional*) received some attention because the Central Limit Theorem fails for the random variable $N(F, \mathbb{G}_{n,p})$ for such F, see [2,13,17]. Apart from sporadic examples, infinitely many such pairs were constructed by Janson and Kratochvil [16] for $p = 1/2$ and by Janson and Spencer [18] for every fixed rational p; see Kärrman [19] for a different proof of the last result.

The main contribution of this paper is the following.

Theorem 1.1. (a) *Let $k \geq 3$ be fixed and $p = p(n) \in (0, 1)$ with $\frac{1}{p(1-p)} = o(n^{1/2})$. Then*

$$u_k(n, p) = O\big(\max\{D(n, p),\ p(1-p)\}n^{k-2}\big).$$

(b) *Let $k \geq 4$ be fixed and $p = p(n) \in (0, 1)$. Then*

$$u_k(n, p) = \Omega\big(\max\{D(n, p),\ p(1-p)\}n^{k-2}\big).$$

Note that the existence of proportional graphs shows that the lower bound of Theorem 1.1 does not extend in general to $k = 3$.

Another measure of graph similarity is the $2k$th Shatten norm $\|G-p\|_{C_{2k}}$. Lemma 8.12 in [21] shows that the 4th Shatten norm defines the same topology as the cut-norm. Again, we define it only for the special case when we want to measure how p-quasirandom an n-vertex graph G is, where we allow loops. Here, we take the (normalised) ℓ_{2k}-norm of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $M = A - pJ$, where A is the adjacency matrix of G and J is the all-1 matrix:

$$\|G - p\|_{C_{2k}} := \frac{(\lambda_1^{2k} + \cdots + \lambda_n^{2k})^{1/2k}}{n}.$$

We remark that when G has a loop, the corresponding diagonal entry in the matrix A is 1. An equivalent and more combinatorial definition of the $2k$th Shatten norm is to take $\|G - p\|_{C_{2k}} = t(C_{2k}, M)^{1/2k}$, where C_{2k} is the $2k$-cycle and $t(F, M)$ denotes the *homomorphism density* of a graph F, which is the expected value of $\prod_{ij\in E(F)} M_{f(i),f(j)}$, where $f : V(F) \to [n]$ is a uniformly chosen random function, see [21, Chapter 5]. In other words,

$$\|G-p\|_{C_{2k}} = \left(n^{-2k} \sum_{f:\mathbb{Z}/2k\mathbb{Z}\to[n]} \prod_{i\in\mathbb{Z}/2k\mathbb{Z}} (A_{f(i),f(i+1)} - p) \right)^{1/2k}, \tag{1.3}$$

where the sum is over all n^{2k} maps $f : \mathbb{Z}/2k\mathbb{Z} \to [n]$, from the integer residues modulo $2k$ to $\{1,\dots,n\}$.

We can show the following result.

Theorem 1.2. *Let $k \ge 2$ be a fixed integer. The minimum of $\|G-p\|_{C_{2k}}$ over all n-vertex graphs G with loops allowed is*

$$\Theta\left(\min\left\{p(1-p),\ p^{1/2}(1-p)^{1/2}n^{-(k-1)/2k}\right\}\right).$$

Hatami [12] studied which graphs other than even cycles produce a norm when we use the appropriate analogue of (1.3). He showed, among other things, that complete bipartite graphs with both parts of even size do. We also prove a version of Theorem 1.2 for this norm, see Theorem 4.1 of Section 4.

The rest of this paper is organised as follows. In Section 2 we prove the lower bound from Theorem 1.1. In Section 3 we prove the upper bound. We consider graph norms in Section 4, in particular proving Theorem 1.2 there. The final section contains some open questions and concluding remarks. Throughout the paper, we adopt the convention that k is a fixed constant and all asymptotic notation symbols (Ω, O, o and Θ) are with respect to the variable n. To simplify the presentation, we often omit floor and ceiling signs whenever these are not crucial and make no attempts to optimise the absolute constants involved.

2. Lower bound for $u_k(n,p)$ in the range $k \ge 4$

The goal of this section is to prove that $u_k(n,p) = \Omega\left(\max\{D(n,p),\ p(1-p)\}n^{k-2}\right)$. More precisely, we will show that there exists a constant $\varepsilon = \varepsilon(k) > 0$ such that $u_k(G,p) \ge \varepsilon \max\{D(n,p),\ p(1-p)\}n^{k-2}$, for all graphs G on $n \ge k$ vertices and for all $0 < p < 1$. The following lemma shows that it is enough to prove the lower bound for $k = 4$ only.

Lemma 2.1. *For every $k \ge 2$ there is $c_k > 0$ such that $u_{k+1}(G,p) \ge c_k n \cdot u_k(G,p)$ for every graph G of order $n \ge k+1$ and for all $0 < p < 1$.*

Proof. Define

$$u_F(G,p) := \left|N(F,G) - \mathbf{E}[N(F,\mathbb{G}_{n,p})]\right|.$$

Take a graph F of order k with $u_F(G,p) = u_k(G,p)$. Let $f(G)$ be the number of pairs (A,x) where a k-set A induces F in G and $x \in V(G)\setminus A$. Then $f(G) = (n-k)N(F,G)$ and $\mathbf{E}[f(\mathbb{G}_{n,p})] = (n-k)\mathbf{E}[N(F,\mathbb{G}_{n,p})]$; thus these two parameters differ (in absolute value) by exactly $(n-k)u_k(G,p)$. On the other hand, $f(G)$ can be written as $\sum_{F'} N(F,F')N(F',G)$ where the sum is over non-isomorphic $(k+1)$-vertex graphs F'. The expectation of $f(\mathbb{G}_{n,p})$ obeys the same linear identity:

$$\mathbf{E}[f(\mathbb{G}_{n,p})] = \sum_{v(F')=k+1} N(F,F')\,\mathbf{E}[N(F',\mathbb{G}_{n,p})].$$

We conclude that

$$\begin{aligned}\frac{n}{k+1}\,u_k(G,p) &\le (n-k)\,u_k(G,p) = \left|f(G) - \mathbf{E}[f(\mathbb{G}_{n,p})]\right| \\ &\le \sum_{v(F')=k+1} N(F,F')\,u_{F'}(G,p) \le 2^{\binom{k+1}{2}}\cdot(k+1)\cdot u_{k+1}(G,p).\end{aligned}$$

Thus the lemma holds with $c_k = 2^{-\binom{k+1}{2}}(k+1)^{-2}$. $\square$

In the next lemma we prove one of the bounds for $u_4(n,p)$. We remark that it was implicitly proven in [16, Proposition 3.7].

Lemma 2.2. *There exists an absolute constant $\varepsilon > 0$ such that, for every $0 < p < 1$ and for all graphs G on $n \geq 4$ vertices, the inequality $u_4(G,p) \geq \varepsilon p(1-p)n^2$ holds.*

Proof. Let $\varepsilon > 0$ be a sufficiently small constant. Suppose that there is a graph G of order $n \geq 4$ satisfying $u_4(G,p) < \varepsilon p(1-p)n^2$. By applying Lemma 2.1 twice, we conclude that $u_2(G,p) < \varepsilon_1 p(1-p)$, where we set $\varepsilon_1 := \varepsilon/(c_2c_3)$ with the constants c_i given by the lemma. This implies that

$$\begin{aligned}\left|e(G)^2 - \mathbf{E}\big[e(\mathbb{G}_{n,p})\big]^2\right| &\leq \left|e(G) - \mathbf{E}\left[e(\mathbb{G}_{n,p})\right]\right| \cdot \left(2p\binom{n}{2} + \varepsilon_1 p(1-p)\right)\\ &< \varepsilon_1 p(1-p)\cdot 3p\binom{n}{2} = 3\varepsilon_1 p^2(1-p)\binom{n}{2}.\end{aligned} \tag{2.1}$$

For every graph G, we can write $e(G)^2$ as

$$e(G)^2 = \sum_{2\leq v(F)\leq 4} \alpha_F N(F,G), \tag{2.2}$$

where F in the summation ranges over non-isomorphic graphs satisfying $2 \leq v(F) \leq 4$, and $\alpha_F \geq 0$ is a constant depending on F only. Indeed, split ordered pairs $(e,e') \in E(G)^2$ according to the isomorphism type F of $G[e \cup e']$. The number α_F of times that a given F-subgraph in G is counted equals the number of ways to pick an ordered pair of edges from $E(F)$ whose union is the whole vertex set $V(F)$. For example, if F is an edge then $\alpha_F = 1$, and if $v(F) = 4$ then α_F is the number of ordered pairs of disjoint edges in F.

Since $\mathbf{E}\left[e(\mathbb{G}_{n,p})^2\right] - \mathbf{E}\big[e(\mathbb{G}_{n,p})\big]^2 = \mathbf{Var}[e(\mathbb{G}_{n,p})] = p(1-p)\binom{n}{2}$ is the variance of $e(\mathbb{G}_{n,p})$, we have by (2.1) and the Triangle Inequality that

$$\left|e(G)^2 - \mathbf{E}\left[e(\mathbb{G}_{n,p})^2\right]\right| > p(1-p)\binom{n}{2} - 3\varepsilon_1 p^2(1-p)\binom{n}{2} > \frac{p(1-p)}{2}\binom{n}{2}. \tag{2.3}$$

Moreover, the identity (2.2) implies that $\mathbf{E}\left[e(\mathbb{G}_{n,p})^2\right] = \sum_{2\leq v(F)\leq 4}\alpha_F\, \mathbf{E}[N(F,\mathbb{G}_{n,p})]$. Thus, by (2.3),

$$\begin{aligned}\sum_{k=2}^{4}\sum_{v(F)=k}\alpha_F\, u_k(G,p) &\geq \sum_{k=2}^{4}\sum_{v(F)=k}\alpha_F\left|N(F,G) - \mathbf{E}[N(F,\mathbb{G}_{n,p})]\right|\\ &\geq \left|\sum_{k=2}^{4}\sum_{v(F)=k}\alpha_F\Big(N(F,G) - \mathbf{E}[N(F,\mathbb{G}_{n,p})]\Big)\right|\\ &= \left|e(G)^2 - \mathbf{E}\left[e(\mathbb{G}_{n,p})^2\right]\right| > \frac{p(1-p)}{2}\binom{n}{2}.\end{aligned}$$

Thus for some $k \in \{2,3,4\}$, we have $u_k(G,p) \geq \varepsilon p(1-p)n^2$. Lemma 2.1 implies that $u_4(G,p) > \varepsilon p(1-p)n^2$, contradicting our assumption and proving the lemma. □

The previous two lemmas give that $u_k(n,p) = \Omega(p(1-p)n^{k-2})$ for $k \geq 4$. Thus, in order to finish the proof of the lower bound, we need to show that $u_k(n,p) = \Omega(D(n,p)n^{k-2})$. The latter bound is a consequence of $u_2(n,p) = D(n,p)$ together with Lemma 2.1, thereby concluding the proof of Theorem 1.1(b).

3. Upper bound for $k \geq 3$

In this section, we prove that $u_k(n,p) = O(\max\{D(n,p), p(1-p)\}n^{k-2})$ for fixed $k \geq 3$ and for all $p = p(n)$ such that $\frac{1}{p(1-p)} = o(n^{1/2})$. We can assume, without loss of generality, that $p \leq \frac{1}{2}$. Indeed, if $\overline{G}$ denotes the complement of G then $u_k(G,p) = u_k(\overline{G}, 1-p)$, which implies that $u_k(n,p) = u_k(n,1-p)$.

Thus our assumption can be made because the bound $O(\max\{D(n,p), p(1-p)\}n^{k-2})$ is symmetric with respect to p and $1-p$. (Recall that $D(n,p) = u_2(n,p) = u_2(n,1-p) = D(n,1-p)$.) In addition, note that in the range $p \leq \frac{1}{2}$, it suffices to show that $u_k(n,p) = O(\max\{D(n,p), p\}n^{k-2})$.

To prove the upper bound, we borrow some definitions, results, and proof ideas from [18]. Following their notation, one can count the number of induced subgraphs of G that are isomorphic to H using the following identity

$$N(H,G) = \sum_{H'} \prod_{e \in E(H')} I_G(e) \prod_{e \in E(\overline{H'})} (1 - I_G(e)), \tag{3.1}$$

where we sum over all H' isomorphic to H with $V(H') \subseteq V(G)$, $I_G(e)$ is the indicator function that e is an edge in G and $\overline{H'}$ denotes the complement of the graph H'. Observe that the range of H' taken in the outermost sum in (3.1) depends on $V(G)$ but not on $E(G)$; this will be useful when comparing H-counts in different graphs on the same vertex set. We define a related sum over the same range of H':

$$S(H,G) = S^{(p)}(H,G) := \sum_{H'} \prod_{e \in E(H')} (I_G(e) - p), \tag{3.2}$$

where p is as before. Rewriting (3.1) by replacing each factor $I_G(e)$ by $(I_G(e) - p) + p$ and each factor $1 - I_G(e)$ by $(1-p) - (I_G(e) - p)$ and expanding, we obtain a linear combination of products $\prod_{e \in X}(I_G(e) - p)$, with each X being some subset of unordered pairs of $V(G)$ involving at most $v(H)$ different vertices. All sets X that are isomorphic to the same graph F get the same coefficient, which we denote $a_{F,H}(n,p)$. The coefficient for $X = \emptyset$ (i.e. the constant term) is obtained by summing the same quantity $p^{e(H')}(1-p)^{e(\overline{H'})}$ over all summands H'; thus it is equal to the expected number of H-subgraphs in $\mathbb{G}_{n,p}$. We separate this special term and re-write (3.1) as

$$N(H,G) = \mathbf{E}[N(H, \mathbb{G}_{n,p})] + \sum_{F \in \mathcal{F}_k} a_{F,H}(n,p) S(F,G), \tag{3.3}$$

where $k = v(H)$ and $\mathcal{F}_k$ denotes the family of all graphs F without isolated vertices satisfying $2 \leq v(F) \leq k$. Also, note that $a_{F,H}(n,p)$ does not depend on G and is bounded from above by $O(n^{v(H)-v(F)})$. In fact, one can show that $a_{F,H}(n,p) = O(p^{e(H)-\alpha} n^{v(H)-v(F)})$, where α is the maximum number of edges that a common subgraph of both H and F can have, but we will not need such an estimate.

Thus, in order to prove that there exists a graph G on n vertices such that $u_k(G,p) = O(\max\{D(n,p), p\}n^{k-2})$, it suffices to show that there exists G such that

$$S(F,G) = \begin{cases} O(pn^{v(F)-2}), & \text{for all } F \in \mathcal{F}_k \setminus \{K_2\}, \\ O(D(n,p)), & \text{if } F = K_2. \end{cases} \tag{3.4}$$

(Note that one cannot hope for $S(K_2, G) = O(p)$ in general; this is why we need two terms in the asymptotic formula for $u_k(n,p)$.) A natural candidate for G in (3.4) is the random graph $G \sim \mathbb{G}_{n,p}$. Unfortunately, G does not work "out of the box"; namely, (3.4) typically fails for $F \in \mathcal{F}_k$ with $v(F) \leq 3$. However, by changing the adjacencies of carefully chosen pairs we can steer these parameters to have the desired order of magnitude.

The next lemma yields some bounds for $S(F, \mathbb{G}_{n,p})$.

Lemma 3.1. *Let $G \sim \mathbb{G}_{n,p}$. For all $F \in \mathcal{F}_k$, we have*

$$\mathbf{E}[S(F,G)] = 0 \quad \textit{and} \quad \mathbf{E}[S(F,G)^2] \leq p^{e(F)} n^{v(F)}.$$

Proof. By (3.2), we have

$$\mathbf{E}[S(F,G)] = \sum_{F'} \mathbf{E}\left[\prod_{e \in E(F')} (I_G(e) - p)\right],$$

where the sum is over all F' isomorphic to F with $V(F') \subseteq V(G)$. Each expectation on the right-hand side vanishes, by independence and since $\mathbf{E}[I_G(e)] = p$. Thus $\mathbf{E}[S(F,G)] = 0$.

We similarly write

$$\mathbf{E}[S(F,G)^2] = \sum_{F',F''} \mathbf{E}\left[\prod_{e\in E(F')}(I_G(e)-p)\prod_{e\in E(F'')}(I_G(e)-p)\right].$$

where the sum is over all pairs (F', F'') of graphs isomorphic to F with $V(F') \cup V(F'') \subseteq V(G)$. The expectation term in the above sum vanishes when $F' \neq F''$ and it is equal to $(p-p^2)^{e(F)} \leq p^{e(F)}$ when $F' = F''$. Since the number of possible choices for F' is at most $\binom{n}{f}\cdot f! \leq n^f$, where $f = v(F)$, we conclude that $\mathbf{E}[S(F,G)^2] \leq p^{e(F)} n^{v(F)}$. □

Using Chebyshev's inequality (see, e.g., [1, Theorem 4.1.1]), we have that, for all $\lambda > 0$,

$$\mathbf{Pr}\left[\,|S(F,\mathbb{G}_{n,p})| \geq \lambda \cdot p^{e(F)/2} n^{v(F)/2}\,\right] \leq \lambda^{-2}. \tag{3.5}$$

By the union bound combined with (3.5), the random graph $G \sim \mathbb{G}_{n,p}$ satisfies the following property with probability at least 0.96.

Property A. $|S(F,G)| \leq 5|\mathcal{F}_k|^{1/2} p^{e(F)/2} n^{v(F)/2}$ *for all graphs* $F \in \mathcal{F}_k$.

The inequality $p^{e(F)/2} n^{v(F)/2} \leq pn^{v(F)-2}$ holds whenever $v(F) \geq 4$. This is because every graph on 4 or more vertices in $\mathcal{F}_k$ has at least 2 edges, since no vertex is isolated. In order to find a graph satisfying the conditions expressed in (3.4), we just need to adjust G so that $S(K_2, G) = O(D(n,p))$ and $S(F,G) = O(pn^{v(F)-2})$ when $F \in \mathcal{F}_3 \setminus \{K_2\}$. The family $\mathcal{F}_3 \setminus \{K_2\}$ consists of two graphs: the triangle K_3 and the 2-path P_2, the unique graph on three vertices having exactly two edges. So, we just need to adjust $S(K_2, G)$, $S(K_3, G)$ and $S(P_2, G)$. This must be performed carefully, to prevent $S(F,G)$ from changing too much for graphs $F \in \mathcal{F}_k$ with $v(F) \geq 4$.

Let us investigate what happens to $S(F,G)$ when we add or remove an edge. Note that by "edges", we generally mean edges in the complete graph, i.e., all pairs ij with $i,j \in V(G)$, and not only the pairs that happen to be selected as the edges of G. For each pair ij with $i,j \in V(G)$, let

$$S_{ij}(F,G) := S(F, G\cup\{ij\}) - S(F, G\setminus\{ij\}), \tag{3.6}$$

where $G \cup \{ij\}$ and $G \setminus \{ij\}$ represent the graphs obtained from G by adding and removing the edge ij, respectively. By expanding each of the two terms in (3.6) using (3.2), we can write $S_{ij}(F,G)$ as the sum of $\prod_{e\in E(F')}(I_{G\cup\{ij\}}(e)-p) - \prod_{e\in E(F')}(I_{G\setminus\{ij\}}(e)-p)$ over all F-subgraphs F' inside $V(G)$. If $E(F')$ does not contain ij, then both products are identical. Thus we have that

$$S_{ij}(F,G) = \sum_{F'}\big((1-p)-(-p)\big)\prod_{e\in E(F')\setminus\{ij\}}(I_G(e)-p) = \sum_{F'}\prod_{e\in E(F')\setminus\{ij\}}(I_G(e)-p), \tag{3.7}$$

where we sum over all F' isomorphic to F with $V(F') \subseteq V(G)$ and $ij \in E(F')$.

The next lemma gives a bound for the expectation and the variance of $S_{ij}(F, \mathbb{G}_{n,p})$.

Lemma 3.2. *Let* $G \sim \mathbb{G}_{n,p}$. *For all* $F \in \mathcal{F}_k$ *with* $v(F) \geq 3$ *and all pairs* $1 \leq i < j \leq n$, *we have*

$$\mathbf{E}[S_{ij}(F,G)] = 0 \quad \text{and} \quad \mathbf{E}[S_{ij}(F,G)^2] \leq k^2 p^{e(F)-1} n^{v(F)-2}.$$

Proof. The proof is similar to that of Lemma 3.1.

We have $\mathbf{E}[S_{ij}(F,G)] = 0$ by (3.7), the independence of the random variables $I_G(e)$ and the linearity of expectation.

For the second part of the lemma, we write

$$\mathbf{E}[S_{ij}(F,G)^2] = \sum_{F',F''} \mathbf{E}\left[\prod_{e\in E(F')\setminus\{ij\}}(I_G(e)-p)\prod_{e\in E(F'')\setminus\{ij\}}(I_G(e)-p)\right].$$

where the sum is over all pairs (F', F'') of graphs isomorphic to F with $V(F') \cup V(F'') \subseteq V(G)$ and $\{i, j\} \in E(F') \cap E(F'')$. The expectation term in the above sum vanishes when $F' \neq F''$ and it is upper bounded by $p^{e(F)-1}$ when $F' = F''$. Since the number of possible choices for F' is at most $k^2 n^{v(F)-2}$, we conclude that $\mathbf{E}[S_{ij}(F, G)^2] \leq k^2 p^{e(F)-1} n^{v(F)-2}$, as desired. □

Take a pair ij of vertices. For $0 \leq s \leq 2$, let $Z_s = Z_s(ij)$ denote the number of vertices $z \in V(G) \setminus \{i, j\}$ such that exactly s of the pairs iz and jz belong to $E(G)$. Let us express

$$\begin{aligned} Y_1 &= Y_1(ij) := S_{ij}(P_2, G), \\ Y_2 &= Y_2(ij) := S_{ij}(K_3, G), \end{aligned}$$

in terms of the random variables Z_0 and Z_2. When we compute Y_1 using (3.7), we have to sum over all 2-paths containing the edge ij. Denoting the third vertex of the path by z, we get

$$Y_1 = \sum_{z \in V \setminus \{i,j\}} (I_G(iz) + I_G(jz) - 2p) = 2(1-p)Z_2 + (1-2p)Z_1 - 2pZ_0.$$

Using that $\mathbf{E}[Z_0] = (1-p)^2(n-2)$ and $\mathbf{E}[Z_2] = p^2(n-2)$ (or that $\mathbf{E}[Y_1] = 0$), we derive that

$$\begin{aligned} Y_1 &= 2(1-p)Z_2 + (1-2p)(n-2-Z_0-Z_2) - 2pZ_0 \\ &= (Z_2 - \mathbf{E}[Z_2]) - (Z_0 - \mathbf{E}[Z_0]). \end{aligned} \tag{3.8}$$

Likewise, we obtain

$$\begin{aligned} Y_2 &= \sum_{z \in V \setminus \{i,j\}} (I_G(iz) - p)(I_G(jz) - p) \ = \ (1-p)^2 Z_2 - p(1-p)Z_1 + p^2 Z_0 \\ &= (1-p)(Z_2 - \mathbf{E}[Z_2]) + p(Z_0 - \mathbf{E}[Z_0]). \end{aligned} \tag{3.9}$$

The triple (Z_0, Z_1, Z_2) has a multinomial distribution for $G \sim \mathbb{G}_{n,p}$. In the next lemma we show that for any fixed rectangle $R \subseteq \mathbb{R}^2$ of positive area, there exists $\eta = \eta(R) > 0$ such that $\left(\frac{Y_1}{\sqrt{pn}}, \frac{Y_2}{p\sqrt{n}}\right) \in R$ with probability at least η. Recall that we have assumed that $p \leq 1/2$ and $p^2 n \to \infty$.

Lemma 3.3. *For fixed reals $\alpha_1 < \alpha_2$ and $\beta_1 < \beta_2$ there exists $\eta = \eta(\alpha_1, \alpha_2, \beta_1, \beta_2) > 0$ such that, for all large n, the probability of*

$$\alpha_1 \leq \frac{Y_1}{\sqrt{pn}} \leq \alpha_2 \quad \textit{and} \quad \beta_1 \leq \frac{Y_2}{p\sqrt{n}} \leq \beta_2 \tag{3.10}$$

is at least η.

Proof. Define

$$\begin{aligned} c &:= \frac{1}{2} \min\{\alpha_2 - \alpha_1,\ \beta_2 - \beta_1\}, \\ C &:= 2\max\left\{ |\alpha_1|, |\alpha_2|, |\beta_1|, |\beta_2| \right\}, \\ \delta &:= \frac{c}{8\pi}\, \mathrm{e}^{-2C^2} \ > \ 0. \end{aligned}$$

Let us show that $\eta := \delta^2$ works in the lemma. Consider the following 2×2-matrix and its inverse:

$$A := \begin{bmatrix} -1 & \sqrt{p} \\ \sqrt{p} & 1-p \end{bmatrix} \quad \text{with} \quad A^{-1} = \begin{bmatrix} -1+p & \sqrt{p} \\ \sqrt{p} & 1 \end{bmatrix}.$$

Note that each entry of A and A^{-1} has absolute value at most 1, so the linear maps given by these matrices are 2-Lipschitz in the ℓ_∞-distance. Thus if we let $S = S(n)$ be the square of side length c with centre $(\alpha_0, \beta_0)^T := A^{-1}(\frac{\alpha_1+\alpha_2}{2}, \frac{\beta_1+\beta_2}{2})^T$, then the image of S under A lies inside the rectangle $R := [\alpha_1, \alpha_2] \times [\beta_1, \beta_2]$ while S itself is a subset of $A^{-1}R \subseteq [-C, C]^2$. (Here $(\alpha, \beta)^T$ means the column vector with entries (α, β).)

The matrix A was chosen to encode the linear relations (3.8) and (3.9) between (Y_1, Y_2) and (Z_0, Z_2), with an appropriate normalisation applied to each random variable. Specifically, it holds that

$$A\left(\frac{Z_0 - \mathbf{E}[Z_0]}{\sqrt{pn}}, \frac{Z_2 - \mathbf{E}[Z_2]}{p\sqrt{n}}\right)^T = \left(\frac{Y_1}{\sqrt{pn}}, \frac{Y_2}{p\sqrt{n}}\right)^T. \tag{3.11}$$

By (3.11) it is enough to show, that with probability at least η, we have

$$\alpha_0 - \frac{c}{2} \leq \frac{Z_0 - \mathbf{E}[Z_0]}{\sqrt{pn}} \leq \alpha_0 + \frac{c}{2}, \tag{3.12}$$

$$\beta_0 - \frac{c}{2} \leq \frac{Z_2 - \mathbf{E}[Z_2]}{p\sqrt{n}} \leq \beta_0 + \frac{c}{2}. \tag{3.13}$$

A version of de Moivre–Laplace theorem (see e.g. [3, Theorem 1.6(i)]) states that, for any function $p = p(n) \in (0, 1)$ with $p(1-p)n \to \infty$ and any reals $a < b$, if X_n has the binomial distribution with parameters (n, p), then

$$\lim_{n\to\infty} \mathbf{Pr}\left[a \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq b\right] = \frac{1}{2\pi}\int_a^b e^{-x^2/2} dx. \tag{3.14}$$

Let n be large. We begin by sampling Z_2. We know that Z_2 is distributed according to the binomial distribution: $Z_2 \sim \mathrm{Bin}(n-2, p^2)$. Its variance is $\mathbf{Var}[Z_2] = p^2(1-p^2)(n-2)$. Let $Z_2^* := (Z_2 - \mathbf{E}[Z_2])/\sqrt{\mathbf{Var}[Z_2]}$ be the normalised version of Z_2. Note that the constraint (3.13) is satisfied if and only if Z_2^* belongs to $\gamma_n \cdot [\beta_0 - \frac{c}{2}, \beta_0 + \frac{c}{2}]$, where $\gamma_n := p\sqrt{n}/\sqrt{\mathbf{Var}[Z_2]}$ and $y \cdot X := \{y \cdot x : x \in X\}$ denotes the dilation of a set X by a scalar y. De Moivre–Laplace theorem (3.14) applies to Z_2 since we assumed that $p^2 n \to \infty$ and $p \leq 1/2$. Using $p \leq 1/2$ again, we have that γ_n is between, for example, 1 and 2. Note that the normal distribution assigns probability at least 2δ to every interval of length c inside $[-2C, 2C]$ by the definition of δ.

Let us show that the probability of (3.13) is at least δ. If this is false, then by passing to a subsequence of counterexamples n we can further assume that γ_n and $\beta_0 = \beta_0(n)$ converge to some γ and β respectively (with $\gamma \in [1, 2]$ and $|\beta| \leq C - c/2$). Let $I = [a, b]$ be the interval with centre at $\frac{a+b}{2} = \gamma\beta$ such that de Moivre–Laplace theorem predicts the limiting probability $\frac{3}{2}\,\delta$ for it. Its length $a - b$ is strictly smaller than γc because, as we have already observed, the probability that the normal variable hits $\gamma \cdot [\beta - \frac{c}{2}, \beta + \frac{c}{2}]$ is at least 2δ. Thus, for all large n from our subsequence, I is a subset of $\gamma_n \cdot [\beta_0(n) - \frac{c}{2}, \beta_0(n) + \frac{c}{2}]$. However, our assumption states that each of the latter intervals is hit with probability less than δ by Z_2^*, contradicting de Moivre–Laplace theorem when applied to the constant interval I.

Let $\alpha \in \{0, \ldots, n-2\}$ be such that $|\beta - \beta_0| \leq c/2$, where we set $\beta := (\alpha - (n-2)p^2)/(p\sqrt{n})$. Let X_α be Z_0 conditioned on $Z_2 = \alpha$. The random variable X_α has the binomial distribution with parameters $(1-p^2)(n-2) - \beta p\sqrt{n}$ and $\frac{(1-p)^2}{1-p^2} = \frac{1-p}{1+p}$. By our assumption $p^2 n \to \infty$, the term $\beta p\sqrt{n} = O(p\sqrt{n})$ is negligible when compared to $p^2 n$. We have

$$\mathbf{E}[X_\alpha] = (1-p)^2(n-2) - \frac{1-p}{1+p} \cdot \beta p\sqrt{n},$$

$$\mathbf{Var}[X_\alpha] = (1+o(1))\frac{1-p}{1+p} \cdot \frac{2p}{1+p} \cdot (1-p^2)n = (2+o(1))\frac{p(1-p)^2 n}{1+p}.$$

We see that $\mathbf{Var}[X_\alpha]$ lies between, for example, $np/4$ and $4np$. As before, a compactness argument based on de Moivre–Laplace theorem shows that the infimum over all intervals $I \subseteq [-2C, 2C]$ of length $c/2$ of the probability that $(X_\alpha - \mathbf{E}[X_\alpha])/\sqrt{\mathbf{Var}[X_\alpha]}$ belongs to I is at least δ for all large n.

We see that, when conditioned on any value α of Z_2 that satisfies (3.13), the probability that (3.12) holds is at least δ. Therefore, the probability that (3.12) and (3.13) hold simultaneously is at least $\eta = \delta^2$, which concludes the proof. $\square$

Next, we put a pair $e \subseteq V(G)$ in at most one of sets $E_1, \ldots, E_5$ as follows:

$$E_1 := \{e : e \in E(G),\ \sqrt{pn} < Y_1(e) \text{ and } p\sqrt{n} < Y_2(e)\},$$
$$E_2 := \{e : e \in E(G),\ \sqrt{pn} < Y_1(e) \text{ and } Y_2(e) < -p\sqrt{n}\},$$
$$E_3 := \{e : e \in E(G),\ Y_1(e) < -\sqrt{pn} \text{ and } p\sqrt{n} < Y_2(e)\},$$
$$E_4 := \{e : e \in E(G),\ Y_1(e) < -\sqrt{pn} \text{ and } Y_2(e) < -p\sqrt{n}\},$$
$$E_5 := \{e : e \notin E(G),\ |Y_1(e)| < 0.1\sqrt{pn} \text{ and } |Y_2(e)| < 0.1p\sqrt{n}\}.$$

Also, let E^* denote the set of pairs ij, where $i, j \in V(G)$ are distinct vertices such that

$$|S_{ij}(F, G)| > 4k \cdot \varepsilon^{-1/2}|\mathcal{F}_k|^{1/2}p^{(e(F)-1)/2}n^{v(F)/2-1} \tag{3.15}$$

for at least one $F \in \mathcal{F}_k$.

Informally speaking, the rest of the proof proceeds as follows. First, by using Lemma 3.3 we show that, with reasonably high probability, the set $E_i \setminus E^*$ is "large" for each $i \in [5]$. Then, by applying a simple greedy algorithm, Corollary 3.5 gives a bounded degree graph H' consisting of $\Omega(n)$ edges from each $E_i \setminus E^*$. We will modify the random graph G to satisfy (3.4) by flipping some pairs, all restricted to H'. First, by flipping the appropriate number of pairs inside either E_1 or E_5, we can make $|S(K_2, G)|$ to be equal to $D(n, p)$, the smallest possible value, thus satisfying one of the constraints in (3.4). Next, by adding an edge from E_5 to $E(G)$ and removing an edge in E_i from $E(G)$, we do not change $S(K_2, G)$ while we can steer each of $S(K_3, G)$ and $S(P_2, G)$ in the right direction by having the freedom to choose $i \in [4]$. The latter claim can be justified using the fact that all flipped pairs come from a bounded degree graph H', so the updated values of $Y_1(e)$ and $Y_2(e)$ stay close to the initial values for every pair $e \subseteq V(G)$. Furthermore, since H' is disjoint from E^*, the effect on $S(F, G)$ of every H'-flip is small for each $F \in \mathcal{F}_k$. Thus we make (3.4) hold for $F \in \mathcal{F}_3$ without violating it for the graphs in $\mathcal{F}_k \setminus \mathcal{F}_3$.

Let us provide all the details. Let $\varepsilon > 0$ be sufficiently small, in particular so that $\eta = \varepsilon$ satisfies Lemma 3.3 for any choice of $\alpha_1 < \alpha_2$ and $\beta_1 < \beta_2$ from $\{\pm 0.1, \pm 1, \pm 2\}$.

First, let us show that $|E_1| \geq \varepsilon pn^2/4$ asymptotically almost surely. Recall that E_1 consists of those pairs $e \subseteq V(G)$ for which

$$e \in E(G), \quad \sqrt{pn} < Y_1(e) \quad \text{and} \quad p\sqrt{n} < Y_2(e). \tag{3.16}$$

Let $I_1(e)$ be the indicator random variable for E_1. For the random graph $G \sim \mathbb{G}_{n,p}$, the first condition $e \in E(G)$ for e to be in E_1 is independent of the other two conditions. Thus, by the choice of ε, we can assume that $\mathbf{E}[I_1(e)] \geq \varepsilon p$. We have $|E_1| = \sum_e I_1(e)$, hence $\mathbf{E}[\,|E_1|\,] \geq \varepsilon p\binom{n}{2}$. We re-write the variance of $|E_1|$ as the sum of pairwise covariances of its components: with $\mathbf{Cov}[X, Y] := \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]$ we have

$$\mathbf{Var}[\,|E_1|\,] = \sum_{e \cap e' = \emptyset} \mathbf{Cov}[I_1(e), I_1(e')] + \sum_{e \cap e' \neq \emptyset} \mathbf{Cov}[I_1(e), I_1(e')], \tag{3.17}$$

Take any pairs $e = xy$ and $e' = x'y'$ that have no common vertices. Let us show that $\mathbf{Cov}[I_1(e), I_1(e')] = o(p^2)$. Informally speaking, $I_1(e)$ can only influence $I_1(e')$ through the four edges that connect e to e', while the probability that Y_1 or Y_2 is so close to the cut-off values in (3.16) as to be affected by these four edges is $o(1)$ by de Moivre–Laplace theorem. A bit more formally, we first expose all edges between the set $A := e \cup e'$ and its complement $V(G) \setminus A$, and compute the "current" values Y_1' and Y_2' on e and e' where, for example,

$$Y_1'(e) := \sum_{z \in V(G)\setminus A} (I_G(xz) + I_G(yz) - 2p)$$

takes into account those 2-paths on $V(G)$ that contain $e = xy$ as an edge but are vertex–disjoint from the other pair e'. The values of Y_1 and Y_2 on e and e' can be computed from Y_1' and Y_2' by adding the contribution from the four edges connecting e to e'. By (3.8) and (3.9), each of these increments is at most 8. If $Y_1'(e), Y_1'(e') \not\in \sqrt{pn} \pm 8$ and $Y_2'(e), Y_2'(e') \not\in p\sqrt{n} \pm 8$, then the validity of the requirements on Y_1 and Y_2 in (3.16) does not depend on the four edges between e and e'; thus the corresponding contribution to $\mathbf{Cov}[I_1(e), I_1(e')]$ is zero. The complementary event, that at least one of Y_1' and Y_2' is within additive constant 8 from the corresponding cut-off value, has probability $o(1)$ by an application of de Moivre–Laplace theorem. Furthermore, the constraints $e, e' \in E(G)$ in (3.16), that are independent of everything else, contribute $O(p^2)$ to the covariance of $I_1(e)$ and $I_1(e')$. Thus indeed $\mathbf{Cov}[I_1(e), I_1(e')] = o(p^2)$.

We see that the first sum in (3.17) has $O(n^4)$ terms, each $o(p^2)$. Since the second sum has $O(n^3)$ terms, each at most p^2 and $\binom{n}{2}$ terms, each at most p, the variance of $|E_1|$ is $o(n^4p^2)$. By Chebyshev's inequality,

$$\mathbf{Pr}[\,|E_1| < \varepsilon pn^2/4\,] \;\le\; \mathbf{Pr}[\,|E_1 - \mathbf{E}[E_1]| > \varepsilon pn^2/5\,] \;=\; o(1),$$

proving the required.

The argument above implies that asymptotically almost surely $|E_i| \ge \varepsilon pn^2/4$ for all $i = 1, \dots, 4$. Similarly, one can show that $|E_5| \ge \varepsilon n^2/4$ asymptotically almost surely. (Note that E_5 might be much "denser" than the other sets because we dropped the requirement $e \in E(G)$.) Finally, using the standard Chernoff estimates one can show that asymptotically almost surely $\Delta(G) \le 2np$ for $G \sim \mathbb{G}_{n,p}$. In particular, the following property is satisfied with probability at least 0.99 when n is large.

Property B. $|E_i| \ge \varepsilon pn^2/4$ *for* $i = 1, \dots, 4$. *Moreover,* $|E_5| \ge \varepsilon n^2/4$ *and* $\Delta(G) \le 2pn$.

Next, we would like to show that the set E^* that was defined by (3.15) is small. Chebyshev's inequality together with Lemma 3.2 implies that $\mathbf{Pr}[ij \in E^*] \le \varepsilon/16$. Hence $\mathbf{E}[\,|E^*|\,] \le \varepsilon n^2/32$. By Markov's inequality, $\mathbf{Pr}[\,|E^*| > \varepsilon n^2/8\,] < \frac{1}{4}$. Similarly, $\mathbf{Pr}[\,|E^* \cap E(G)| > \varepsilon pn^2/8\,] < \frac{1}{4}$. Thus by the union bound, $G \sim \mathbb{G}_{n,p}$ satisfies the following property with probability at least 0.5.

Property C. E^* *has size at most* $\varepsilon n^2/8$. *Moreover,* $|E^* \cap E(G)| \le \varepsilon pn^2/8$.

Also, we state and prove the following simple result that asserts the existence of large matchings in relatively dense graphs.

Proposition 3.4. *Let* H *be a graph and let* $\Delta := \Delta(H)$. *There exists a matching in* H *of size at least* $\frac{e(H)}{2\Delta}$. *In particular, if* $m < \Delta$ *then* H *contains a subgraph* H' *with maximal degree* $\Delta(H') \le m$ *and* $e(H') \ge \frac{m}{4\Delta} e(H)$.

Proof. Let M be a maximal matching in H, and assume M has $k < \frac{e(H)}{2\Delta}$ pairs. All the edges of H have at least one endpoint in $V(M)$. Hence

$$e(H) \le |V(M)| \cdot \Delta = 2k \cdot \Delta < e(H),$$

a contradiction. We remark that the bound $\frac{e(H)}{2\Delta}$ is not tight but it suffices for our purposes.

To construct H', we start with the empty graph. At each step of the construction, we apply the first assertion of the proposition to the graph $H \setminus H'$, in order to obtain a matching M having exactly $\lceil \frac{e(H)}{4\Delta} \rceil$ edges. We then add all the edges from M to H'. We repeat this step exactly m times. Since we always have $e(H') \le m \cdot \lceil \frac{e(H)}{4\Delta} \rceil < \frac{e(H)}{2}$, and thus $e(H \setminus H') > \frac{e(H)}{2}$, it is always possible to find such M, in all the steps of the process. □

An important corollary of Proposition 3.4 is as follows.

Corollary 3.5. *Let* $C > 0$ *be fixed. If Properties B and C simultaneously hold for a graph* G *and* n *is sufficiently large, then there exists a graph* H' *having at least* Cn *edges from each* $E_i \setminus E^*$, $i = 1, \dots, 5$, *such that* $\Delta(H') \le 320C/\varepsilon$.

Proof. Because of Property C, we have $|E^* \cap E(G)| \le \varepsilon pn^2/8$ and $|E^*| \le \varepsilon n^2/8$, which, together with Property B, implies that $|E_i \setminus E^*| \ge \varepsilon pn^2/8$ for $i = 1, \dots, 4$, and $|E_5 \setminus E^*| \ge \varepsilon n^2/8$. Let H_i be the graph on $V(G)$ having edge set $E_i \setminus E^*$. We have $\Delta(H_i) \le \Delta(G) \le 2np$ for $i = 1, \dots, 4$ and $\Delta(H_5) \le n$. Hence $\frac{e(H_i)}{\Delta(H_i)} \ge \frac{\varepsilon n}{16}$ for all $i = 1, \dots, 5$. By Proposition 3.4 applied with $m = 64C/\varepsilon < \min\{\Delta(H_i) : i = 1, \dots, 5\}$, each H_i contains a subgraph H_i' having at least $\frac{m}{4} \cdot \frac{e(H_i)}{\Delta(H_i)} \ge Cn$ edges such that $\Delta(H_i') \le m$.

Let $H' = \bigcup_{i=1}^{5} H'_i$. Clearly $\Delta(H') \le 5m = 320C/\varepsilon$ and H' contains at least Cn edges from each $E_i \setminus E^*$, thereby proving the corollary. $\square$

Proof of the upper bound in Theorem 1.1. Given $p \in (0, 1/2]$ and $k \ge 3$, choose small $\varepsilon > 0$ and then sufficiently large C. Let $n \to \infty$. By the union bound, $G \sim \mathbb{G}_{n,p}$ satisfies Properties A–C with probability at least 0.4. Hence there exists a graph G on n vertices satisfying the three properties simultaneously. Fix such G.

From Corollary 3.5, there exists a graph H' having at least Cn edges from each $E_i \setminus E^*$, such that $\Delta := \Delta(H') \le 320C/\varepsilon$. Let $E' = E(H')$.

In what follows, we change $E(G)$ on pairs, all of which will belong to E'. Note that at any intermediate step, the effect of (for instance) removing an edge $ij \in E' \cap E_1$ from $E(G)$ on $S(P_2, G)$ and $S(K_3, G)$ is not quite given by the initial values of $Y_1(ij)$ and $Y_2(ij)$, since certain edges iw, jw might have been changed. But E' was defined in such a way that there are most $2\Delta = o(\sqrt{pn})$ changed edges which affect either Y_1 or Y_2. So, the removal of $ij \in E_1 \setminus E^*$ from $E(G)$ at any intermediate stage, still decreases $S(P_2, G)$ by an amount between $\sqrt{pn} - 2\Delta$ and $4k\varepsilon^{-1/2}|\mathcal{F}_k|^{1/2}\sqrt{pn} + 2\Delta < \varepsilon^{-1}\sqrt{pn}$. Similarly, because $\Delta = o(p\sqrt{n})$, the same operation decreases $S(K_3, G)$ by an amount between $p\sqrt{n} - 2\Delta$ and $4k\varepsilon^{-1/2}|\mathcal{F}_k|^{1/2}p\sqrt{n} + 2\Delta < \varepsilon^{-1}p\sqrt{n}$.

By Property A, we know that

$$|S(K_2, G)| \le 5|\mathcal{F}_k|^{1/2}p^{1/2}n =: \tau.$$

If $S(K_2, G) \ge 1$, we can pick an $e \in E' \setminus E_5$ and remove it from G. This has the effect of reducing $S(K_2, G)$ by 1. If $S(K_2, G) \le -1$, then we can pick an $e \in E' \cap E_5$ and add it to G. This new edge increases the value of $S(K_2, G)$ by 1. Iterate this process at most τ times to obtain a graph G such that $|S(K_2, G)| = D(n, p)$, always using a different edge e. This is possible because there are at least Cn edges from $E' \cap E_i$, for each i.

Since we have flipped at most τ edges, all belonging to H', and each flip changes $S(K_3, G)$ (reps. $S(P_2, G)$) by at most $\varepsilon^{-1}p\sqrt{n}$ (resp. $\varepsilon^{-1}\sqrt{pn}$) in absolute value, the current graph satisfies $|S(K_3, G)| \le pS_0$ and $|S(P_2, G)| \le p^{1/2}S_0$, where

$$S_0 = 5|\mathcal{F}_k|^{1/2}p^{1/2}n^{3/2} + \tau \cdot \varepsilon^{-1}\sqrt{n}.$$

Our next goal is to make both $|S(K_3, G)|$ and $|S(P_2, G)|$ small without changing $S(K_2, G)$. We repeat the following step $Cp^{1/2}n - \tau$ times. Consider the current graph G. There are four cases depending on whether each of $S(K_3, G)$ and $S(P_2, G)$ is positive or not. First suppose that they are both positive. Pick previously unused edges $e \in E' \cap E_1$ and $e' \in E' \cap E_5$, and replace e with e' in G. This operation preserves the value of $S(K_2, G)$, and has the effect of reducing both $S(K_3, G)$ and $S(P_2, G)$. It reduces $S(K_3, G)$ by an amount between $(1 - 0.1)p\sqrt{n} - 4\Delta \ge 0.8p\sqrt{n}$ and $2\varepsilon^{-1}p\sqrt{n} < pn$. Thus if (initially) $S(K_3, G) \ge pn$, then this value is lowered by at least $0.8p\sqrt{n}$. Regarding $S(P_2, G)$, the operation reduces it by between $0.8\sqrt{pn}$ and $2\varepsilon^{-1}\sqrt{pn} < pn$. Likewise, if $S(K_3, G) < 0$ and $S(P_2, G) > 0$, we replace an $e \in E' \cap E_2$ by an $e' \in E' \cap E_5$, and similarly in the other two cases. We iterate this process, always using edges e and e' that have not been used before. This is possible since E' contains at least Cn edges from each E_i. Also, once one of $|S(K_3, G)|$ or $|S(P_2, G)|$ becomes less than pn, it stays so for the rest of the process. Since $(Cp^{1/2}n - \tau) \cdot 0.8\sqrt{n} > S_0$, we have that $\max\{|S(K_3, G)|, |S(P_2, G)|\} < pn$ at the end.

The iterative process might change the value of $S(F, G)$ for $F \in \mathcal{F}_k$ with at least 4 vertices. Take any such F and let $f = v(F)$. Initially, $|S(F, G)|$ was at most $5|\mathcal{F}_k|^{1/2}p^{e(F)/2}n^{f/2}$ by Property A. If we add to it $Cp^{1/2}n$, an upper bound on the number of the changed edges, multiplied by $4k\varepsilon^{-1/2}|\mathcal{F}_k|^{1/2}p^{(e(F)-1)/2}n^{f/2-1}$, then this accounts for every copy of F inside the vertex set $V(G)$ except perhaps those that contain at least two of the changed edges. (This estimate used the fact that none of the changed edges is in E^*.) A pair of two disjoint changed edges is trivially in at most f^4n^{f-4} copies of F. It remains to consider the case when xy and xz are two changed intersecting edges. Note that there are at most $Cp^{1/2}n \cdot 2\Delta$ choices of (xy, xz). Consider a copy F' of F with vertex set $X \supseteq \{x, y, z\}$. If none of the pairs $e \subseteq X$ with $e \not\subseteq \{x, y, z\}$ is an element of $E(G)$ or a changed edge, then this F' contributes at most p in absolute value to the sum in (3.2) that defines $S(F, G)$. (Indeed, as F has at least 4 non-isolated vertices, at least one edge of F' has to intersect $X \setminus \{x, y, z\}$; thus the F'-term in (3.2) contains at least one factor $-p$.) Otherwise, X has to contain a changed edge or an edge from $E(G)$ that is not inside $\{x, y, z\}$. The number of such subgraphs for any given triple $\{x, y, z\}$ can be bounded by

$$3(\Delta + 2pn)f^4 n^{f-4} + (Cp^{1/2}n + pn^2)f^5 n^{f-5} \leq 2f^5 pn^{f-3}.$$

Putting all together we obtain that, at the end of the process,

$$\begin{aligned} |S(F, G)| \leq\ & 5|\mathcal{F}_k|^{1/2} p^{e(F)/2} n^{f/2} + Cp^{1/2}n \cdot 4k\varepsilon^{-1/2}|\mathcal{F}_k|^{1/2} p^{(e(F)-1)/2} n^{f/2-1} \\ & + (Cp^{1/2}n)^2 f^4 n^{f-4} + Cp^{1/2}n \cdot 2\Delta \cdot (p \cdot f^3 n^{f-3} + 2f^5 pn^{f-3}). \end{aligned}$$

This is $O(pn^{f-2})$ since F has $f \geq 4$ vertices and $e(F) \geq 2$ edges.

We conclude that the final graph G satisfies $S(F, G) = O(pn^{v(F)-2})$ for all $F \in \mathcal{F}_k \setminus \{K_2\}$ and $S(K_2, G) = O(D(n, p))$. That is, we satisfied (3.4), which implies the required upper bound on $u_k(G, p)$. $\square$

4. Shatten norms and other related norms

Note that the graphs in this section are allowed to have loops. When we define the complement $\overline{G}$ of a graph G, loopless vertices are mapped to loops and vice versa. For a graph G on $[n]$ and a function $p = p(n)$, let $M = A - pJ$ denote the shifted adjacency matrix of G, that is,

$$M_{ij} = \begin{cases} 1 - p, & \text{if } ij \in E(G), \\ -p, & \text{otherwise,} \end{cases} \qquad 1 \leq i, j \leq n. \tag{4.1}$$

In order to make some forthcoming formulas shorter, we define $\epsilon(G) := \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}$. In other words, $\epsilon(G)$ is the number of loops plus twice the number of non-loop edges in G. For example, $\epsilon(G) + \epsilon(\overline{G}) = n^2$.

Let us prove Theorem 1.2.

Proof of Theorem 1.2. Let $s = 2k$ and let G be a graph (possibly with loops) on $[n]$, where $n \to \infty$. Without loss of generality we may assume that $p \leq \frac{1}{2}$. This is because $\|G - p\|_{C_s}^s = \|\overline{G} - (1 - p)\|_{C_s}^s$ and the expression in the statement we have to prove is symmetric with respect to p and $1 - p$.

The matrix M in (4.1) is a symmetric real matrix so it has real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. For an even integer $s \geq 4$, we have

$$n^s \, \|G - p\|_{C_s}^s = \sum_{i=1}^{n} \lambda_i^s = \operatorname{tr}(M^s) = \sum_{i=1}^{n} (M^s)_{ii},$$

where tr denotes the trace of a matrix.

From now on we split the analysis of the lower bound for $\|G - p\|_{C_s}^s$ into two cases.

In the first case, we assume that $\epsilon(G) \geq \frac{p}{2}\, n^2$. This (together with $p \leq \frac{1}{2}$) implies that

$$\sum_{i=1}^{n} \lambda_i^2 = \sum_{i,j=1}^{n} M_{ij}^2 = (1-p)^2 \epsilon(G) + p^2 \epsilon(\overline{G}) \geq \left((1-p)^2 \frac{p}{2} + p^2 \left(1 - \frac{p}{2}\right)\right) n^2 = \frac{p}{2}\, n^2. \tag{4.2}$$

By the inequality between the arithmetic and kth power means for $k \geq 2$ applied to non-negative numbers $\lambda_1^2, \ldots, \lambda_n^2$ (or just by the convexity of $x \mapsto x^k$ for $x \geq 0$), we conclude that

$$\left(\frac{\lambda_1^{2k} + \cdots + \lambda_n^{2k}}{n}\right)^{1/k} \geq \frac{\lambda_1^2 + \cdots + \lambda_n^2}{n} \geq \frac{pn}{2}.$$

Thus $n^{2k}\|p - G\|_{C_{2k}}^{2k} = \sum_{i=1}^{n} \lambda_i^{2k} = \Omega(p^k n^{k+1})$, giving the required lower bound in the first case.

In the second case, we assume that $\epsilon(G) < \frac{p}{2}\, n^2$. Since λ_n is the smallest eigenvalue of M, we have $\lambda_n = \min\{\langle Mv, v\rangle : \|v\|_2 = 1\}$. So if we choose $v = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in \mathbb{R}^n$, we obtain

$$\lambda_n \leq \langle Mv, v\rangle = \frac{(1-p)\epsilon(G) - p\epsilon(\overline{G})}{n} \leq \left((1-p)\frac{p}{2} - p(1 - \frac{p}{2})\right) n = -\frac{pn}{2}. \tag{4.3}$$

This implies that $\sum_{i=1}^{n}\lambda_i^{2k} \geq \lambda_n^{2k} = \Omega(p^{2k}n^{2k})$, thereby proving the lower bound in the second case.

On the other hand, for the upper bound we have two constructions. Again we assume that $p \leq \frac{1}{2}$. The first construction is very simple: the empty graph. If G is empty, a straightforward computation shows that $\|G-p\|_{C_{2k}} = p$, and this proves the upper bound whenever $p \leq n^{-(k-1)/k}$. For the second construction, we consider $G \sim \mathbb{G}_{n,p}^{\text{loop}}$ to be a random graph with loops, where every possible pair or loop belongs to $E(G)$ independently with probability p. Here we assume that $p > n^{-(k-1)/k}$. Let $X = n^{2k}\|G-p\|_{C_{2k}}^{2k}$. By (1.3), we have $X = \sum_{f:\mathbb{Z}/2k\mathbb{Z}\to V(G)} X_f$, where $X_f = \prod_{i\in\mathbb{Z}/2k\mathbb{Z}} M_{f(i),f(i+1)}$ and $M = A - pJ$ is as before. Then the expectation of X_f is 0 unless for every i there is $j \neq i$ with $\{f(j), f(j+1)\} = \{f(i), f(i+1)\}$, that is, every edge of C_{2k} is glued with some other edge. If f is a map with $\mathbf{E}[X_f] \neq 0$ then the image under f of the edge set of C_{2k} is a connected multi-graph where every edge (or loop) appears with even multiplicity, so it contains at most $k+1$ vertices. Since the number of maps f for which the image of C_{2k} contains at most e distinct edges (ignoring multiplicity) is $O(n^{e+1})$, we have

$$\mathbf{E}[X] = O\left(\sum_{e=1}^{k} n^{e+1}p^e\right) = O(n^{k+1}p^k),$$

since $p > n^{-1}$. Now take an outcome G such that the value of X is at most its expected value. This finishes the proof of the theorem. □

A related result of Hatami [12] shows that a complete bipartite graph $F = K_{2k,2m}$, with even part sizes $2k$ and $2m$, also gives a norm by a version of (1.3). If G is a graph on $[n]$, then this norm, for $G - p$, is

$$\|G-p\|_F := t(F, M)^{1/(2k+2m)} = n^{-1}X^{1/(2k+2m)},$$

where M is as in (4.1),

$$X := \sum_{f:A\cup B\to V(G)} \prod_{a\in A}\prod_{b\in B} M_{f(a),f(b)},$$

and A, B are fixed disjoint sets of sizes $2k$ and $2m$ respectively.

Theorem 4.1. *Let $F = K_{2k,2m}$ with $1 \leq k \leq m$. The minimum of $\|G-p\|_F$ over n-vertex graphs G with loops allowed is*

$$\Theta\left(\min\left\{p^{4km}(1-p)^{4km},\ p^{2km}(1-p)^{2km}n^{-k}\right\}^{1/(2m+2k)}\right).$$

Proof. For the same reasons stated in the beginning of the proof of Theorem 1.2 we may assume, without loss of generality, that $p \leq \frac{1}{2}$. We begin with the lower bound. We rewrite X by grouping all maps $f : A \cup B \to V(G)$ by the restriction of f to A. For every fixed $h : A \to V(G)$, we have

$$\sum_{g:B\to V(G)} \prod_{a\in A}\prod_{b\in B} M_{h(a),g(b)} = \left(\sum_{u\in V(G)}\prod_{a\in A} M_{h(a),u}\right)^{2m} \geq 0.$$

As in the proof of Theorem 1.2, we divide the analysis into two cases.

In the first case, we assume that $\epsilon(G) \geq \frac{p}{2}n^2$. Let $\mathcal{H}$ be the set of all $h : A \to V(G)$ such that $h(2i-1) = h(2i)$ for all $i \in [k]$, where we assumed that $A = [2k]$. Note that $|\mathcal{H}| = n^k$. If $h \in \mathcal{H}$ we have

$$\sum_{u\in V(G)}\prod_{a\in A} M_{h(a),u} = \sum_{u\in V(G)}\prod_{i\in[k]} M_{h(2i),u}^2.$$

Thus by the convexity of $x \mapsto x^{2m}$ for $x \in \mathbb{R}$, the convexity of $x \mapsto x^k$ for $x \geq 0$, and the calculation in (4.2), we have that

$$X = \sum_{h:A\to V(G)}\left(\sum_{u\in V(G)}\prod_{a\in A} M_{h(a),u}\right)^{2m} \geq \sum_{h\in\mathcal{H}}\left(\sum_{u\in V(G)}\prod_{i\in[k]} M_{h(2i),u}^2\right)^{2m}$$

$$\geq n^k\left(\frac{1}{n^k}\sum_{h\in\mathcal{H}}\sum_{u\in V(G)}\prod_{i\in[k]}M^2_{h(2i),u}\right)^{2m} = n^k\left(\frac{1}{n^k}\sum_{u\in V(G)}\left[\sum_{v\in V(G)}M^2_{v,u}\right]^k\right)^{2m}$$

$$\geq n^k\left(\frac{1}{n^{k-1}}\left[\frac{1}{n}\sum_{u\in V(G)}\sum_{v\in V(G)}M^2_{v,u}\right]^k\right)^{2m} = n^k\left(\frac{1}{n^{k-1}}\left[\frac{(1-p)^2\epsilon(G)+p^2\epsilon(\overline{G})}{n}\right]^k\right)^{2m}$$

$$\geq n^k\left(\frac{1}{n^{k-1}}\left[\frac{pn}{2}\right]^k\right)^{2m} = \Omega\left(p^{2km}n^{k+2m}\right),$$

which proves the lower bound in the first case.

In the second case, we assume that $\epsilon(G) < \frac{p}{2}n^2$. By the convexity of $x \mapsto x^{2m}$ and $x \mapsto x^{2k}$ for all $x \in \mathbb{R}$ and by the calculation in (4.3), we have that

$$X = \sum_{h:A\to V(G)}\left(\sum_{u\in V(G)}\prod_{a\in A}M_{h(a),u}\right)^{2m} \geq n^{2k}\left(\frac{1}{n^{2k}}\sum_{h:A\to V(G)}\sum_{u\in V(G)}\prod_{i\in[2k]}M_{h(i),u}\right)^{2m}$$

$$= n^{2k}\left(\frac{1}{n^{2k}}\sum_{u\in V(G)}\left[\sum_{v\in V(G)}M_{v,u}\right]^{2k}\right)^{2m} \geq n^{2k}\left(\frac{1}{n^{2k-1}}\left[\frac{1}{n}\sum_{u\in V(G)}\sum_{v\in V(G)}M_{v,u}\right]^{2k}\right)^{2m}$$

$$= n^{2k}\left(\frac{1}{n^{2k-1}}\left[\frac{(1-p)\epsilon(G)-p\epsilon(\overline{G})}{n}\right]^{2k}\right)^{2m} = \Omega\left(p^{4km}n^{2k+2m}\right),$$

which proves the lower bound in the second case.

We turn to the upper bound. We need two constructions. The first one is again the empty graph. If G is empty then

$$\|G-p\|_F = p^{2km/(k+m)},$$

and this proves the upper bound whenever $p \leq n^{-1/(2m)}$. The second construction is the random graph $G \sim \mathbb{G}^{\text{loop}}_{n,p}$. Write X as the sum of X_f over $f : A\cup B \to V(G)$. Each f with $\mathbf{E}[X_f] \neq 0$ maps $E(K_{2k,2m})$ into a connected multi-graph where every edge appears with even multiplicity. Consider the equivalence relation on $A\cup B$ given by one such f, where two vertices in $A\cup B$ are equivalent if their images under f coincide. If non-trivial classes (i.e., those containing more than one vertex) miss some $a \in A$ and some $b \in B$, then $\{f(a), f(b)\}$ is a singly-covered edge, a contradiction. Thus, non-trivial classes have to cover at least one of A or B entirely, so the number of identifications is at least $\min\{|A|, |B|\}/2 = k$. It follows that the image of F under f has at most $k + 2m$ vertices. In fact, if the image of F under f contains exactly $2k+2m-t$ vertices (where $t \geq k$), the number of distinct edges in the image of F by f is at least $4km - 2mt$. This is because every "identification" of vertices under the same equivalence class of f can "destroy" at most $2m$ edges. Therefore

$$\mathbf{E}[X] = O\left(\sum_{t=k}^{2k+2m-1} n^{2k+2m-t}p^{4km-2mt}\right) = O(n^{k+2m}p^{2km}),$$

since $p > n^{-1/(2m)}$. Now take an outcome G such that the value of X is at most its expected value. This finishes the proof of the theorem. □

5. Concluding remarks and open questions

Observe that the result of Chung, Graham, Wilson [7] implies that there cannot be a graph G with $t(K_2, A) = p$ and $t(C_4, A) = p^4$ where $0 < p < 1$ and A is the adjacency matrix of G. (Indeed, otherwise the uniform blow-ups of G would form a quasirandom sequence, which is a contradiction.) This argument does not work with the subgraph count function $N(F, G)$. We do not know if the fact that $u_k(n, p)$ can be zero infinitely often for $k = 3$ (when p is rational) but not for $k = 4$ can directly be related to the fact that quasirandomness is forced by 4-vertex densities.

Let $\mathbb{G}_{n,m}$ be the random graph on $[n]$ with m edges, where all $\binom{\binom{n}{2}}{m}$ outcomes are equally likely. Janson [14] completely classified the cases when the random variable $N(F, \mathbb{G}_{n,m})$ satisfies the Central Limit Theorem where $n \to \infty$ and $m = \lfloor p\binom{n}{2}\rfloor$. He showed that the exceptional F are precisely those graphs for which $S^{(p)}(H, F) = 0$ for every H from the following set: connected graphs with 5 vertices and graphs without isolated vertices with 3 or 4 vertices. It is an open question if at least one such pair (F, p) with $p \neq 0, 1$ exists, see, e.g., [14, Page 65] and [15, Page 350]. Note that nothing is stipulated about $S^{(p)}(K_2, F)$. In fact, it has to be non-zero e.g. by Theorem 1.1; moreover, [14, Theorem 4] shows that, for given $v(F)$ and p, the number of edges in such hypothetical F is uniquely determined. This indicates that the problem of understanding possible joint behaviour of the S-statistics is difficult already for very small graphs.

It would be interesting to extend Theorem 1.1 to a wider range of p, or to other structures such as, for example, r-uniform hypergraphs with respect to different notions of quasirandomness (see [8,20,22]).

Acknowledgements

We thank the anonymous referees for the careful reading of the manuscript and helpful comments. The first author was supported in part by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation. The second author was supported in part by ERC grant 306493 and EPSRC grant EP/K012045/1.

References

[1] N. Alon, J.H. Spencer, The Probabilistic Method, fourth ed., in: Wiley Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., Hoboken, NJ, 2016, p. xiv+375.
[2] A.D. Barbour, M. Karoński, A. Ruciński, A central limit theorem for decomposable random variables with applications to random graphs, J. Combin. Theory (B) 47 (1989) 125–145.
[3] B. Bollobás, Random Graphs, second ed., Cambridge Univ. Press, 2001.
[4] B. Bollobás, A.D. Scott, Discrepancy in graphs and hypergraphs, in: More Sets, Graphs and Numbers, in: Bolyai Soc. Math. Stud., vol. 15, Springer, Berlin, 2006, pp. 33–56.
[5] B. Bollobás, A.D. Scott, Intersections of graphs, J. Graph Theory 66 (2011) 261–282.
[6] B. Bollobás, A.D. Scott, Intersections of hypergraphs, J. Combin. Theory (B) 110 (2015) 180–208.
[7] F.R.K. Chung, R.L. Graham, R.M. Wilson, Quasi-random graphs, Combinatorica 9 (1989) 345–362.
[8] D. Conlon, H. Hàn, Y. Person, M. Schacht, Weak quasi-randomness for uniform hypergraphs, Random Structures Algorithms 40 (2012) 1–38.
[9] P. Erdős, On combinatorial questions connected with a theorem of Ramsey and van der Waerden, Mat. Lapok 14 (1963) 29–37.
[10] P. Erdős, M. Goldberg, J. Pach, J. Spencer, Cutting a graph into two dissimilar halves, J. Graph Theory 12 (1988) 121–131.
[11] P. Erdős, J. Spencer, Imbalances in k-colorations, Networks 1 (1971/72) 379–385.
[12] H. Hatami, Graph norms and Sidorenko's conjecture, Israel J. Math. 175 (2010) 125–150.
[13] S. Janson, A functional limit theorem for random graphs with applications to subgraph count statistics, Random Structures Algorithms 1 (1990) 15–37.
[14] S. Janson, Orthogonal decompositions and functional limit theorems for random graph statistics, Mem. Amer. Math. Soc. 111 (534) (1994) vi+78.
[15] S. Janson, A graph fourier transform and proportional graphs, Random Structures Algorithms 6 (1995) 341–351.
[16] S. Janson, J. Kratochvíl, Proportional graphs, Random Structures Algorithms 2 (1991) 209–224.
[17] S. Janson, K. Nowicki, The asymptotic distributions of generalized U-statistics with applications to random graphs, Probab. Theory Related Fields 90 (1991) 341–375.
[18] S. Janson, J. Spencer, Probabilistic construction of proportional graphs, Random Structures Algorithms 3 (1992) 127–137.
[19] J. Kärrman, Existence of proportional graphs, J. Graph Theory 17 (1993) 207–220.
[20] J. Lenz, D. Mubayi, The poset of hypergraph quasirandomness, Random Structures Algorithms 46 (2015) 762–800.
[21] L. Lovász, Large Networks and Graph Limits, in: Colloquium Publications, Amer. Math. Soc., 2012.
[22] H. Towsner, σ-algebras for quasirandom hypergraphs, Random Structures Algorithms 50 (2017) 114–139.