How unproportional must a graph be?

Humberto Naves* Oleg Pikhurko†

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 Erdős-Rényi random graph with edge probability \( p \). This may be viewed as a measure of how close \( G \) is to being \( p \)-quasirandom. Our main result is that, for fixed \( k \geq 4 \) and \( 0 < p < 1 \), the minimum of \( u_k(G, p) \) over \( n \)-vertex graphs has order of magnitude \( \Theta(n^{k-2}) \).

1 Introduction

An important result of Erdős and Spencer [6] 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 \( \frac{1}{2} |S|^2 \) by at least \( \Omega(n^{3/2}) \). The method in [6] extends to give the same order of magnitude when the target edge density \( 1/2 \) is replaced by an arbitrary fixed real \( p \in (0, 1) \), see [5]. Also, an earlier observation of Erdős [4] shows that this lower bound is tight.

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 [16]), 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 \( \frac{1}{n^2} |2e(G[S]) - p|S|^{2}| \).

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 Erdős-Rényi random graph \( G_{n,p} \) is

\[
\mathbb{E}[N(H, G_{n,p})] = \frac{n(n-1)\ldots(n-k+1)}{|\text{Aut}(H)|} p^{e(H)} (1-p)^{(k^2-e(H))},
\]

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

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

\[
u_k(G, p) := \max_F |N(F, G) - \mathbb{E}[N(F, G_{n,p})]|,
\]

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 \( G_{n,p} \) in terms of

\*Department of Mathematics - ETH, 8092 Zurich, Switzerland and Department of Mathematics - UCLA, Los Angeles, CA 90095 USA. Email: hnaves@math.ucla.edu.

†Mathematics Institute and DIMAP, University of Warwick, Coventry CV4 7AL, UK. Supported by ERC grant 306493 and EPSRC grant EP/K012045/1.
Let $k \geq 3$ and $p \in (0, 1)$ be fixed. We have $u_k(n, p) = O(n^{k-2})$. Moreover, $u_k(n, p) = \Omega(n^{k-2})$ for $k \geq 4$.

Another measure of graph similarity is the $2k$-th Shatten norm. Lemma 8.12 in [10] shows that the 4-th 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 a graph $G$ of order $n$ is. Here, we take the $\ell_{2k}$-norm of the eigenvalues of $M = A - pJ$, where $A$ is the adjacency matrix of $G$ and $J$ is the all-1 matrix. 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 [16, Chapter 5]. In other words,

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

(1.3)

where the sum is over all (not necessarily injective) maps $f : \mathbb{Z}/2k\mathbb{Z} \to V(G)$ from the residues modulo $2k$ to the vertex set of $G$.

Proposition 1.2. Let $p \in (0, 1)$ and integer $k \geq 2$ be fixed. Then the minimum of $\|G - p\|_{C_{2k}}$ over all $n$-vertex graphs $G$ is $\Theta(n^{-(k-1)/2k})$.

Hatami [8] 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 Proposition [12] for this norm.

The rest of this paper is organized as follows. In Section 2 we prove the lower bound $u_k(n, p) = \Omega(n^{k-2})$ for $k \geq 4$. In Section 4 we prove the upper bound $u_k(n, p) = O(n^{k-2})$ for all $k \geq 3$. We consider graphs norms in Section 4 in particular proving Proposition [12] there. The final section contains some open questions and concluding remarks. Throughout the paper, we adopt the convention that $k$ and $p$ are fixed constants and all asymptotic notation symbols ($\Omega$, $O$, $o$ and $\Theta$) 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 optimize the absolute constants involved.
2 Lower bound for $u_k(n, p)$ in the range $k \geq 4$

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

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

**Proof.** Let $u_F(G, p) := |N(F, G) - \mathbb{E}[N(F, \mathcal{G}_{n, p})]|$. 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 $A$ induces $F$ in $G$ and $x \in V(G) \setminus A$. Then $f(G) = (n-k)N(F, G)$ and $\mathbb{E}[f(G_{n, p})] = (n-k)\mathbb{E}[N(F, \mathcal{G}_{n, p})]$; 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 $\sum N(F, F')N(F', G)$ where the sum is over $(k + 1)$-vertex graphs $F'$. The expectation of $f(G_{n, p})$ obeys the same linear identity:

$$\mathbb{E}[f(G_{n, p})] = \sum N(F, F') \mathbb{E}[N(F', \mathcal{G}_{n, p})].$$

Thus there is a $(k + 1)$-vertex graph $F'$ satisfying $u_{F'}(G, p) \geq c_k n \cdot u_k(G, p)$, where $c_k$ is a constant depending solely on $k$. \hfill $\square$

In the next lemma we prove the desired lower bound for $u_4(n, p)$. We remark that it was implicitly proven in [12], Proposition 3.7.

**Lemma 2.2.** For every $0 < p < 1$, there exists a constant $\varepsilon = \varepsilon(p) > 0$ such that $u_4(G, p) \geq \varepsilon n^2$, for all graphs $G$ on $n \geq 4$ vertices.

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

$$e(G)^2 = \sum_F \alpha_F N(F, G),$$

where in the summation $F$ ranges over graphs satisfying $2 \leq v(F) \leq 4$, and $\alpha_F$ is a constant depending on $F$ only. For example, if $F$ is an edge then $\alpha_F = 1$ and if $v(F) = 4$, then $\alpha_F$ is the number of ordered pairs of disjoint edges in $F$.

Let $\varepsilon > 0$ be a sufficiently small constant depending on $p$. Suppose that there is a graph $G$ of order $n \geq 4$ satisfying $u_4(G, p) < \varepsilon n^2$. By applying Lemma 2.1 twice, we conclude that $u_2(G, p) < \varepsilon / c_2 c_3$, where $c_i$'s are given by the lemma. This implies that

$$\left| e(G)^2 - \mathbb{E}[e(G_{n, p})]^2 \right| < \left| e(G) - \mathbb{E}[e(G_{n, p})] \right| \cdot \left( 2p \binom{n}{2} + \frac{\varepsilon}{c_2 c_3} \right) < 3p \frac{\varepsilon}{c_2 c_3} \binom{n}{2}.$$ 

Since $\mathbb{E}[e(G_{n, p})]^2 - \mathbb{E}[e(G_{n, p})]^2 = \text{Var}[e(G_{n, p})] = p(1 - p) \binom{n}{2}$, we have that

$$\left| e(G)^2 - \mathbb{E}[e(G_{n, p})]^2 \right| > \frac{p(1 - p)}{2} \binom{n}{2}. \quad (2.2)$$

Moreover, the identity (2.1) implies that $\mathbb{E}[e(G_{n, p})]^2 = \sum_F \alpha_F \mathbb{E}[N(F, G_{n, p})]$. Thus

$$\sum_F \alpha_F u_4(G, p) \geq \sum_F \alpha_F \left| N(F, G) - \mathbb{E}[N(F, G_{n, p})] \right| > \frac{p(1 - p)}{2} \binom{n}{2}.$$ 

This implies that $u_4(G, p) > \varepsilon n^2$, contradicting our assumption and proving the lemma. \hfill $\square$
3 Upper bound for $k \geq 3$

In this section, we prove that $u_k(n,p) = O(n^{k-2})$ for fixed $k \geq 3$ and $p \in (0,1)$. For that purpose, we borrow some definitions, results, and proof ideas from [14]. Following their notation, one can count the number of induced subgraphs of $G$ that are isomorphic to $H$ using the following identity

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

(3.1)

where we sum over all $H'$ isomorphic to $H$ with $V(H') \subseteq V(G)$, $I_G(e)$ is the indicator that $e$ is an edge in $G$ and $\overline{H'}$ denotes the complement of the graph $H'$. Observe that $H'$ in the sum is not restricted to subgraphs of $G$. We define a related sum over the same range of $H'$:

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

(3.2)

where $p$ is as before. Replacing $I_G(e)$ by $(I_G(e) - p) + p$ and $1 - I_G(e)$ by $(1 - p) - (I_G(e) - p)$ in (3.1), we obtain

$$N(H, G) = \sum_{F \in \mathcal{F}_k} a_{F,H}(n,p)S(F, G),$$

(3.3)

where $k = v(H)$, $\mathcal{F}_k$ denotes the family of all graphs $F$ without isolated vertices satisfying $2 \leq v(F) \leq k$, and $a_{F,H}(n,p) = O(n^{v(H)-v(F)})$ is a coefficient that does not depend on $G$. To prove that there exists a graph $G$ on $n$ vertices such that $u_k(G, p) = O(n^{k-2})$, it suffices to show that there exists $G$ such that

$$S(F, G) = O(n^{v(F)-2})$$

(3.4)

for all $F \in \mathcal{F}_k$. A natural candidate for $G$ in (3.4) is the random graph $\mathcal{G}_{n,p}$. The next lemma yields some bounds for $S(F, \mathcal{G}_{n,p})$.

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

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

**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 pair $(F', F'')$ 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 bounded by 1 when $F' = F''$. Since the number of possible choices for $F'$ is at most $\binom{n}{t} \cdot t! \leq n^t$, where $t = v(F)$, we conclude that $\mathbf{E}[S(F, G)^2] \leq n^{v(F)}$. \hfill \qed
Using Chebyshev’s inequality (see, e.g., [1, Theorem 4.1.1]), we have that, for all \( \lambda > 0 \),

\[
\Pr \left[ |S(F, G_{n,p})| \geq \lambda \cdot n^{v(F)/2} \right] \leq \lambda^{-2}.
\]  

(3.5)

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

Property A. \( |S(F,G)| \leq 5 |F_k|^{1/2} n^{v(F)/2} \) for all graphs \( F \in F_k \).

Clearly \( n^{v(F)/2} \leq n^{v(F)−2} \) when \( v(F) \geq 4 \). In order to find a graph satisfying the conditions expressed in (3.4), we just need to adjust \( G \) so that \( S(F,G) = O(n^{v(F)−2}) \) when \( F \in F_3 \). The adjustment must be performed carefully, to prevent \( S(F,G) \) from changing too much for graphs \( F \in F_k \) 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

\[
S_{ij}(F,G) := S(F,G \cup \{ij\}) − S(F,G \setminus \{ij\}),
\]

(3.6)

where \( G \cup \{ij\} \) and \( G \setminus \{ij\} \) represent the graphs obtained from \( G \) by adding and removing the edge \( ij \), respectively. The next lemma gives a bound for the expectation and the variance of \( S_{ij}(F,G_{n,p}) \).

Lemma 3.2. Let \( G \sim G_{n,p} \). For all \( F \in F_k \) with \( v(F) \geq 3 \) and all pairs \( 1 \leq i < j \leq n \), we have

\[
\mathbb{E}[S_{ij}(F,G)] = 0 \quad \text{and} \quad \mathbb{E}[S_{ij}(F,G)^2] \leq k^2 n^{v(F)−2}.
\]

Proof. By expanding the definition in (3.6), we have

\[
S_{ij}(F,G) = \sum_{F'} \prod_{e \in E(F') \setminus \{ij\}} (I_G(e) − p),
\]

where the sum is over all \( F' \) isomorphic to \( F \) with \( V(F') \subseteq V(G) \) and \( \{i,j\} \in E(F') \). By independence and the linearity of expectation, we have \( \mathbb{E}[S_{ij}(F,G)] = 0 \). For the last part of the lemma, we write

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

where the sum is over all pair \((F',F'')\) 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 bounded by 1 when \( F' = F'' \). Since the number of possible choices for \( F' \) is at most \( k^2 n^{v(F)−2} \), we conclude that \( \mathbb{E}[S(F,G)^2] \leq k^2 n^{v(F)−2} \). \( \square \)

The family \( F_3 \) consists of three graph: cliques \( K_2 \) and \( K_3 \) as well as 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 \( S(P_2,G) \). Let \( \varepsilon > 0 \) be sufficiently small, depending on \( k \) and \( p \).

Take a pair \( ij \) of vertices. Let \( Z_s^* \), for \( 0 \leq s \leq 2 \), denote the number of vertices \( w \in V(G) \setminus \{i,j\} \) such that exactly \( s \) of the pairs \( iw \) and \( jw \) belong in \( G \). The triple \((Z_0^*,Z_1^*,Z_2^*)\) has a multinomial distribution for \( G \sim G_{n,p} \). We shift to zero mean by setting

\[
Z_0 := Z_0^* − (n − 2) \cdot (1 − p)^2,
\]

\[
Z_2 := Z_2^* − (n − 2) \cdot p^2.
\]
Note that $Z_0$ and $Z_2$ determine $S_{ij}(F,G)$ for every 3-vertex graph $F$, which is the amount by how much $S(F,G)$ changes when we “flip” $ij$. Namely, using $Z^*_1 = n - 2 - Z^*_0 - Z^*_2$, we obtain

\[
S_{ij}(P_2, G) = 2(1 - p)Z^*_2 + (1 - 2p)Z^*_1 = Z_2 =: Y_1
\]

\[
S_{ij}(K_3, G) = (1 - p)^2Z^*_2 - p(1 - p)Z^*_1 + p^2Z^*_0 = (1 - p)Z_2 + pZ_0 =: Y_2.
\]

For each vertex $w \in V(G) \setminus \{i, j\}$, consider the zero-mean random vector $v_w \in \mathbb{R}^2$ such that its components $v_{w,1}$ and $v_{w,2}$ are the contributions of $w$ to $Y_1$ and $Y_2$. Thus $v_w$ depends only on the events $iw \in E(G)$ and $jw \in E(G)$ and assumes values

\[
\begin{bmatrix}
-2p \\
p^2
\end{bmatrix}, \begin{bmatrix}
1 - 2p \\
-p + p^2
\end{bmatrix} \text{ and } \begin{bmatrix}
2 - 2p \\
(1 - p)^2
\end{bmatrix}
\]

with probabilities respectively $(1 - p)^2$, $2p(1 - p)$ and $p^2$. Since $E[v_w] = 0$, the covariance matrix of $v_w$ is

\[
C := \text{Cov}[v_w] = \begin{bmatrix}
E[v^2_{w,1}] & E[v_{w,1}v_{w,2}] \\
E[v_{w,1}v_{w,2}] & E[v^2_{w,2}]
\end{bmatrix} = \begin{bmatrix}
2p(1 - p) & 0 \\
0 & p^2(1 - p^2)
\end{bmatrix}.
\]

Since $v_w$ are independent and identically distributed when $w \in V(G) \setminus \{i, j\}$, the vector

\[
\frac{1}{\sqrt{n}} \sum_{w \in V(G) \setminus \{i, j\}} v_w = \frac{1}{\sqrt{n}} \begin{bmatrix}
Y_1 \\
Y_2
\end{bmatrix}
\]

converges to a normal random vector having mean zero and covariance matrix $C$ as $n$ tends to infinity, see e.g. [7, Theorem VIII.4.2]. This means that $Y_1$ and $Y_2$ are asymptotically independent. Thus for any fixed reals $\alpha_1 < \alpha_2$ and $\alpha_3 < \alpha_4$ there exists $\varepsilon_1 = \varepsilon_1(p, \alpha_1, \alpha_2, \alpha_3, \alpha_4) > 0$, such that with probability at least $\varepsilon_1$,

\[
\alpha_1 \sqrt{n} < Y_1 < \alpha_2 \sqrt{n} \quad \text{and} \quad \alpha_3 \sqrt{n} < Y_2 < \alpha_4 \sqrt{n}.
\]

Let **Class 1** consist of those pairs $e \in E(G)$ for which

\[
e \in E(G), \quad \sqrt{n} < Y_1(e) \quad \text{and} \quad \sqrt{n} < Y_2(e).
\]

(3.7)

Let $I_1(e)$ be the indicator random variable for Class 1. For the random graph $G \sim \mathbb{G}_{n,p}$, the first condition $e \in E(G)$ for $e$ to be in Class 1 is independent of the other two conditions. Thus we can assume that $E[I_1(e)] \geq \varepsilon$. Let $W$ be the number of $e \in E(G)$ in Class 1. Then $W = \sum_e I_1(e)$, hence $E[W] \geq \varepsilon(n/2)$. We re-write the variance of $W$ as the sum of pairwise covariances of its components:

\[
\text{Var}[W] = \sum_{e \cap e' = \emptyset} \text{Cov}[I_1(e), I_1(e')] + \sum_{e \cap e' \neq \emptyset} \text{Cov}[I_1(e), I_1(e')].
\]

(3.8)

When $e$ and $e'$ have no common vertices, then $I_1(e)$ can only influence $I_1(e')$ through the four common edges with both endpoints in $e \cup e'$. By the normal approximation, the probability that $Y_1$ or $Y_2$ is within 8 from the thresholds in (3.7) is $o(1)$. It follows that the covariance $\text{Cov}[I_1(e), I_1(e')] = o(1)$. Thus the first sum in (3.8) has $O(n^4)$ terms, each $o(1)$. Since the second sum has $O(n^3)$ terms, each at most 1, the variance of $W$ is $o(n^4)$. By Chebyshev’s inequality,

\[
\Pr[W < \varepsilon n^2/4] = o(1).
\]

Next, we put $e$ in one of Classes 2–5 if the following properties hold:
Class 2: \( e \in E(G), \sqrt{n} < Y_1(e) \) and \( Y_2(e) < -\sqrt{n} \);

Class 3: \( e \in E(G), Y_1(e) < -\sqrt{n} \) and \( \sqrt{n} < Y_2(e) \);

Class 4: \( e \in E(G) \) and \( Y_1(e) < -\sqrt{n} \) and \( Y_2(e) < -\sqrt{n} \);

Class 5: \( e \notin E(G) \), \( |Y_1(e)| < 0.1\sqrt{n} \) and \( |Y_2(e)| < 0.1\sqrt{n} \).

For each of these classes the argument above implies that asymptotically almost surely every class has at least \( \varepsilon n^2/4 \) edges. In particular, \( G \sim \mathbb{G}_{n,p} \) satisfies the following property with probability at least 0.99 when \( n \) is large.

Property B. There are at least \( \varepsilon n^2/4 \) edges in each of the five above-named classes.

Let \( E^* \) denote the set of pairs \( ij \), where \( i, j \in V(G) \), such that

\[
|S_{ij}(F, G)| > 4k \cdot \varepsilon^{-1/2} |F_k|^{1/2} n^{v(F)/2-1}
\]

for at least one \( F \in \mathcal{F}_k \). Chebyshev’s inequality together with Lemma 3.2 implies that \( \Pr[ij \in E^*] \leq \varepsilon/16 \). Hence \( \mathbb{E}[|E^*|] \leq \varepsilon n^2/32 \). By Markov’s inequality, \( \Pr[|E^*| > \varepsilon n^2/8] < \frac{1}{7} \). Thus \( G \sim \mathbb{G}_{n,p} \) satisfies the following property with probability at least 0.75.

Property C. \( E^* \) has size at most \( \varepsilon n^2/8 \).

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

**Proposition 3.3.** Let \( H \) be a graph on \( n \) vertices having at least \( \beta \binom{n}{2} \) edges. Then \( H \) contains a matching of size at least \( \frac{\beta n}{8} \). In particular, if \( \Delta < n \) and \( \beta > 16/n \), then \( H \) contains a subgraph \( H' \) with maximal degree \( \Delta(H') \leq \Delta \) and \( e(H') \geq \frac{\beta \Delta n}{16} \).

**Proof.** Let \( M \) be a maximal matching in \( H \), and assume \( M \) has \( k < \frac{\beta n}{8} \) pairs. All the edges of \( H \) have at least one endpoint in \( V(M) \), hence

\[
e(H) \leq \binom{2k}{2} + 2k \cdot (n - 2k) < \binom{\beta n/4}{2} + \frac{\beta n^2}{4} < \beta \binom{n}{2},
\]

a contradiction. We remark that the bound \( \frac{\beta n}{8} \) 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 \( \lceil \frac{\beta n}{16} \rceil \) edges. We then add all the edges from \( M \) to \( H' \). We repeat this step exactly \( \Delta \) times. Since we always have \( e(H') \leq \Delta \cdot \lceil \frac{\beta n}{16} \rceil < \frac{\beta \Delta n}{2} \), and thus \( e(H \setminus H') > \frac{\beta \Delta n}{2} \), it is always possible to find such \( M \), in all the steps of the process.

**Proof of the upper bound in Theorem 1.1.** Given \( p \in (0, 1) \) and \( k \geq 4 \), 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, B and C with probability at least 0.7. Hence there exists a graph \( G \) on \( n \) vertices satisfying the three properties simultaneously. Fix such \( G \).

From Property C, we know that \( |E^*| \leq \varepsilon n^2/8 \). This, together with Property B, implies that each class contains at least \( \varepsilon n^2/8 \) edges not in \( E^* \). By Proposition 3.3, we obtain a set \( E' \) disjoint from \( E^* \) such that \( E' \) contains at least \( Cn \) edges from each class, and every vertex \( v \in V(G) \) is incident to at most \( \Delta := 320C/\varepsilon \) edges in \( E' \).
In what will follow, we will change \( E(G) \) on pairs, all of which will belong to \( E' \). Note that at any intermediate step, the effect of (for instance) removing a Class 1 edge \( ij \in E' \) 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 \( 2\Delta = O(1) \) changed edges which affect either \( Y_1 \) or \( Y_2 \). So, the removal of \( ij \notin E' \) from \( E(G) \) at any intermediate stage, still decreases both \( S(P_2, G) \) and \( S(K_3, G) \) by an amount between \( \sqrt{n} - 2\Delta \) and \( 4k\epsilon^{-1/2}|F_k|^{1/2}n^{1/2} + 2\Delta < \epsilon^{-1}\sqrt{n} \).

By Property A, we know that \( |S(K_2, G)| \leq 5|F_k|^{1/2}n \). If \( S(K_2, G) \geq 1 \), we can pick an \( e \in E' \) in any class but Class 5 and remove it from \( G \). This has the effect of reducing \( S(K_2, G) \) by 1. If \( S(K_2, G) \leq -1 \), then we can pick an \( e \in E' \) in Class 5 and add it to \( G \). This new edge increases the value of \( S(K_2, G) \) by 1. Iterate this process at most \( 5|F_k|^{1/2}n \) times to obtain a graph \( G \) such that \( |S(K_2, G)| < 1 \), always using a different edge \( e \). This is possible because there are at least \( Cn \) edges from \( E' \) in each class.

Our next goal is to make both \( |S(K_3, G)| \) and \( |S(P_2, G)| \) small without changing \( S(K_2, G) \). Since all the operations were performed on a graph disjoint from \( E^* \) of maximum degree at most \( \Delta \), we know that

\[
S := \max\{|S(K_3, G)|, |S(P_2, G)|\},
\]

is currently bounded from above by

\[
S_0 = 5|F_k|^{1/2}n^{3/2} + 5|F_k|^{1/2}n \cdot \epsilon^{-1}\sqrt{n}.
\]

We repeat the following step \((C - 5|F_k|^{1/2})n\) 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 an edge \( e \in E' \) in Class 1 and an \( e' \in E' \) in Class 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 \( S(P_2, G) \) by between \( 0.8\sqrt{n} \) and \( 2\epsilon^{-1}\sqrt{n} < n \). Thus if \( S \geq n \), then the value of \( S \) is lowered by at least \( 0.8\sqrt{n} \). Likewise, if \( S(K_3, G) < 0 \) and \( S(P_2, G) > 0 \), we replace an \( e \in E' \) in Class 2 by an \( e' \in E' \) in Class 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 class. Also, once one of \( |S(K_3, G)| \) or \( |S(P_2, G)| \) becomes less than \( n \), it stays so for the rest of the process. Since \((C - 5|F_k|^{1/2})n \cdot 0.8\sqrt{n} > S_0 \), we have that \( S < n \) at the end.

The iterative process might change the value of \( S(F, G) \) for \( F \in F_k \) with \( v(F) \geq 4 \). However, since all flips were performed on a bounded degree edge-set \( E' \) disjoint from \( E^* \), we have that

\[
|S(F, G)| \leq 5|F_k|^{1/2}n^{v(F)/2} + Cn \cdot 4k\epsilon^{-1/2}|F_k|^{1/2}n^{v(F)/2-1} + 4(C + \Delta)^2(v(F))^4n^{v(F)-2} = O(n^{v(F)-2}),
\]

where the last summand is a generous upper bound on the number of copies of \( F \) that use more than one changed edge.

Therefore \( S(F, G) = O(n^{v(F)-2}) \) for all \( F \in F_k \), finishing the proof of the upper bound. \( \square \)

### 4 Shatten norms and other related norms

We begin with the proof of Proposition \ref{p:shatten_norms}

**Proof of Proposition \ref{p:shatten_norms}**. Fix \( p \in (0, 1) \). Let \( G \) be a graph of order \( n \to \infty \). Let \( M = A - pJ \) be the shifted adjacency matrix of \( G \), that is,

\[
M_{ij} = \begin{cases} 
1 - p, & \text{if } ij \in E(G), \\
-p, & \text{if } i = j \text{ or } ij \in E(G), 
\end{cases} 
\]

\[1 \leq i, j \leq n. \quad (4.1)
\]
It is a symmetric real matrix so it has real eigenvalues \( \lambda_1, \ldots, \lambda_n \in \mathbb{R} \). For an even integer \( s \geq 4 \), we have
\[
\sum_{i=1}^{n} \lambda_i^s = \text{tr}(M^s) = n^s \| p - G \|_{C_s}^s.
\]
Also, \( \sum_{i=1}^{n} \lambda_i^2 = \sum_{i,j=1}^{n} M_{ij}^2 = (1 - p)^2 e(G) + p^2 e(\overline{G}) + p^2 n = \Omega(n^2) \).

By the inequality between the arithmetic and \( k \)-th 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
\[
\left( \frac{\lambda_1^{2k} + \cdots + \lambda_n^{2k}}{n} \right)^{1/k} \geq \frac{\lambda_1^2 + \cdots + \lambda_n^2}{n} = \Omega(n)
\]
Thus \( n^{2k} \| p - G \|_{C_{2k}}^{2k} = \sum_{i=1}^{n} \lambda_i^{2k} = \Omega(n^{k+1}) \), giving the required lower bound.

On the other hand, let \( G \sim \mathbb{G}_{n,p} \) be a random graph. Let \( X = n^{2k} \| G - p \|_{C_{2k}}^k \). Write
\[
X = \sum_{f: \mathbb{Z}/2k\mathbb{Z} \to V(G)} X_f, \quad X_f = \prod_{i \in \mathbb{Z}/2k\mathbb{Z}} M_{f(i), f(i+1)} \quad \text{and} \quad M = A - pJ \text{ is as before.}
\]
Then the expectation of \( X_f \) is 0 unless for every \( i \) we have \( f(i) = f(i + 1) \) or there is \( j \neq i \) with \( \{ f(j), f(j + 1) \} = \{ f(i), f(i + 1) \} \) (that is, every edge of \( C_{2k} \) is mapped into a loop or is glued with some other edge). Thus the number of maps \( f \) with \( \mathbb{E}[X_f] \neq 0 \) is at most \( O(n^{k+1}) \); the image under \( f \) of the edge set of \( C_{2k} \) is a connected graph consisting of at most \( k \) edges (and some loops) so it contains at most \( k + 1 \) vertices. Now take an outcome \( G \) such that the value of \( X \) is at most its expected value. This finishes the proof of the proposition.

A related result of Hatami [8] shows that a complete bipartite graph \( F = K_{2k, 2m} \) with both part sizes, \( 2k \) and \( 2m \), being even also gives a norm by a version of \( (4.1) \). 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.

**Proposition 4.1.** Let \( p \in (0, 1) \) be fixed and \( F = K_{2k, 2m} \) with \( 1 \leq k \leq m \). Then the minimum of \( \| G - p \|_F \) over \( n \)-vertex graphs \( G \) is \( \Theta(n^{-k/(2m+2k)}) \).

**Proof.** We have to show that the minimal value of \( X \) is \( \Theta(n^{k+2m}) \).

Let us show that a typical random graph \( \mathbb{G}_{n,p} \) gives the upper bound. Write \( X \) as the sum of \( X_f \) over \( f: A \cup B \to V(G) \). Each \( f \) with \( \mathbb{E}[X_f] \neq 0 \) maps \( E(K_{2k, 2m}) \) into a connected graph consisting of multiple edges and loops. Consider the equivalence relation on \( A \cup B \) given by such \( f \), where two vertices in \( A \cup B \) are equivalent if their images under \( f \) coincide. If non-trivial classes 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 \( f \) has at most \( k + 2m \) vertices, giving the required bound.

On the other hand, let us 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
\[
\sum_{g:B \to V(G)} \prod_{a \in A} \prod_{b \in B} M_{h(a), g(b)} = \left( \sum_{u \in V(G)} \prod_{a \in A} M_{h(a), u} \right)^{2m} \geq 0
\]
Take any \( h \) such that it identifies pairs of vertices of \( A = [2k] \), say \( h(2i-1) = h(2i) \) for all \( i \in [k] \). Then
\[
\sum_{u \in V(G)} \prod_{a \in A} M_{h(a), u} = \sum_{u \in V(G)} \prod_{i \in [k]} M_{h(2i), u}^2 \geq n \cdot \left( \min\{p^2, (1-p)^2\} \right)^k = \Omega(n).
\]
Since there are $\Omega(n^k)$ such maps $h$, we conclude that $X \geq \Omega(n^k \cdot n^{2m})$, as required.

5 Concluding remarks and open questions

Observe that the result of Chung, Graham, Wilson [3] 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 $G_{n,m}$ be the random graph on $[n]$ with $m$ edges, where all $\binom{n}{2}$ outcomes are equally likely. Janson [10] completely classified the cases when the random variable $N(F, G_{n,m})$ 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., [10, Page 65] and [11, 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 [14] moreover, [10, 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 general question of understanding possible joint behaviour of the $S$-statistics may be difficult.

It would be interesting to extend Theorem [14] to other structures such as, for example, $r$-uniform hypergraphs.

References

[1] N. Alon and J. Spencer, The probabilistic method, 3d ed., Wiley Interscience, 2008.
[2] A. D. Barbour, M. Karoński, and A. Ruciński, A central limit theorem for decomposable random variables with applications to random graphs, J. Combin. Theory (B) 47 (1989), 125–145.
[3] F. R. K. Chung, R. L. Graham, and R. M. Wilson, Quasi-random graphs, Combinatorica 9 (1989), 345–362.
[4] P. Erdős, On combinatorial questions connected with a theorem of Ramsey and van der Waerden, Mat. Lapok 14 (1963), 29–37.
[5] P. Erdős, M. Goldberg, J. Pach, and J. Spencer, Cutting a graph into two dissimilar halves, J. Graph Theory 12 (1988), 121–131.
[6] P. Erdős and J. Spencer, Imbalances in $k$-colorations, Networks 1 (1971/72), 379–385.
[7] W. Feller, An introduction to probability theory and its applications, 3d ed., Wiley-Intersci. Publ., 1968.
[8] H. Hatami, Graph norms and Sidorenko’s conjecture, Israel J. Math. 175 (2010), 125–150.
[9] S. Janson, A functional limit theorem for random graphs with applications to subgraph count statistics, Random Struct. Algorithms 1 (1990), 15–37.
[10] ______, *Orthogonal decompositions and functional limit theorems for random graph statistics*, Mem. Amer. Math. Soc. **111** (1994), no. 534, vi+78.

[11] ______, *A graph Fourier transform and proportional graphs*, Random Struct. Algorithms **6** (1995), 341–351.

[12] S. Janson and J. Kratochvíl, *Proportional graphs*, Random Struct. Algorithms **2** (1991), 209–224.

[13] S. Janson and K. Nowicki, *The asymptotic distributions of generalized U-statistics with applications to random graphs*, Probab. Theory Related Fields **90** (1991), 341–375.

[14] S. Janson and J. Spencer, *Probabilistic construction of proportional graphs*, Random Struct. Algorithms **3** (1992), 127–137.

[15] J. Käärmann, *Existence of proportional graphs*, J. Graph Theory **17** (1993), 207–220.

[16] L. Lovász, *Large networks and graph limits*, Colloquium Publications, Amer. Math. Soc, 2012.