J. Appl. Probab. **61,** 401–419 (2024)
*doi:*10.1017/jpr.2023.48

NORMAL AND STABLE APPROXIMATION TO SUBGRAPH COUNTS IN SUPERPOSITIONS OF BERNOULLI RANDOM GRAPHS

MINDAUGAS BLOZNELIS,* *Vilnius University*
JOONA KARJALAINEN,** *** AND
LASSE LESKELÄ,** **** *Aalto University*

Abstract

Real networks often exhibit clustering, the tendency to form relatively small groups of nodes with high edge densities. This clustering property can cause large numbers of small and dense subgraphs to emerge in otherwise sparse networks. Subgraph counts are an important and commonly used source of information about the network structure and function. We study probability distributions of subgraph counts in a community affiliation graph. This is a random graph generated as an overlay of m partly overlapping independent Bernoulli random graphs (layers) $G_1, \ldots, G_m$ with variable sizes and densities. The model is parameterised by a joint distribution of layer sizes and densities. When m grows linearly in the number of nodes n, the model generates sparse random graphs with a rich statistical structure, admitting a nonvanishing clustering coefficient and a power-law limiting degree distribution. In this paper we establish the normal and α-stable approximations to the numbers of small cliques, cycles, and more general 2-connected subgraphs of a community affiliation graph.

Keywords: Subgraph count; overlapping communities; Erdös–Rényi graph; normal approximation; clustering; power law; complex network; affiliation network; random intersection graph

2020 Mathematics Subject Classification: Primary 60F05; 05C80
Secondary 05C82; 91C20

1. Introduction and results

Mathematical modeling of complex networks aims to explain and reproduce characteristic properties of large real-world networks, such as power-law degree distributions and clustering. By clustering we refer to the tendency of nodes to cluster together by forming relatively small groups with a high density of ties within a group. Locally, in the vicinity of a vertex v, clustering can be measured by the probability that two randomly selected neighbors of v are adjacent. The average of these probabilities defines the local clustering coefficient of a network. Globally, the fraction of wedges (paths of length 2) that induce triangles defines the

Received 16 March 2022; accepted 2 June 2023.
* Postal address: Institute of Computer Science, Vilnius University, Didlaukio 47, LT-08303, Vilnius, Lithuania. Email: mindaugas.bloznelis@mif.vu.lt
** Postal address: Department of Mathematics and Systems Analysis, School of Science, Aalto University, Otakaari 1, FI-02150 Espoo, Finland.
*** Email address: joona.h.karjalainen@jyu.fi
**** Email address: lasse.leskela@aalto.fi

global clustering coefficient, which represents the probability that endpoints of a randomly selected wedge (friends of a friend) are adjacent. Clearly, nonvanishing clustering coefficients are connected to the abundance of triangles and other small and dense subgraphs. A natural and interesting question is to trace the relation between the clustering characteristics and the frequencies of various network motifs. We address this question by determining the distributional asymptotics of motif (subgraph) counts in a particular network model (community affiliation graph) that possesses the clustering property and a power-law degree distribution.

Another motivation for studying distributions of motif counts in complex network models comes from network science and its applications, where motif frequencies are used for parameter estimation [1, 16] and model evaluation [8]. Moreover, motif frequencies tell of the structure, function, and similarities of real-world networks [2, 13, 19, 20, 24]. In these contexts, it is important to understand the variability of the empirical statistics used in the methods. For example, the approach taken in [25, 27] was to compare empirical statistics from various datasets to their theoretical bounds. Here the knowledge of (asymptotic) distributions of respective motif counts facilitates statistical inference.

In the present paper we establish the normal and α-stable approximations of the numbers of k-cliques, k-cycles, and more general 2-connected subgraphs in a sparse network model defined by a superposition of Bernoulli random graphs [6, 28, 29].

To the best of our knowledge this is the first systematic study of an α-stable approximation to subgraph counts in a theoretical model of a sparse affiliation network. We note that in the network model considered, the clustering property and the power-law degree distribution, the two basic properties of complex networks, are essential for an α-stable limit to emerge.

1.1. Network model

We start with the description of individual layers $G_1, \dots, G_m$. Let (X, Q) be a random vector with values in $\{0, 1, 2, \dots\} \times [0, 1]$, and let $\mathcal{G} = \{G(x, p)\colon x \in \{1, 2 \dots\},\ p \in [0, 1]\}$ be a family of Bernoulli random graphs independent of (X, Q). We set $[x] = \{1, 2, \dots, x\}$ to be the vertex set of $G(x, p)$. Recall that in $G(x, p)$ every pair of vertices $\{i, j\} \subset [x]$ is declared adjacent independently at random with probability p. For notational convenience we introduce the empty graph $G_\emptyset$ having no vertices and set $G(0, p) = G_\emptyset$ for any $p \in [0, 1]$. We define the mixture of Bernoulli random graphs $\mathrm{G}(X, Q)$ in a natural way: we first generate a random vector (X, Q) and then, given the instance (X, Q), we generate a Bernoulli random graph on X vertices with edge density Q. The individual layers $G_1, \dots, G_m$ are independent copies of $G(X, Q)$.

In the next step we map the vertex sets of the layers $G_1, \dots, G_m$ to the set $V = \{1, \dots, n\}$ independently and uniformly at random. The union of mapped layers represents the community affiliation graph, which we denote by $G_{[n,m]}$. More rigorously, let $(X_1, Q_1), (X_2, Q_2), \dots$ be a sequence of independent copies of (X, Q), and let $\mathcal{G}_i = \{G_i(x, p)\colon x \in \mathbb{N},\ p \in [0, 1]\}$, $i = 1, 2, \dots$, be independent copies of $\mathcal{G}$. Given $X_1, \dots, X_m$, let $\mathcal{V}_{n,i} = \mathcal{V}_{n,i}(X_i)$, $1 \le i \le m$, be independent random subsets of $[n]$ defined as follows. For $X_i \le n$ we select $\mathcal{V}_{n,i}$ uniformly at random from the class of subsets of $[n]$ of size X_i. For $X_i > n$ we set $\mathcal{V}_{n,i} = [n]$. We write $\tilde{X}_i = |\mathcal{V}_{n,i}| = X_i \wedge n$. Let $G_{n,i}$, $1 \le i \le m$, be independent random graphs with vertex sets $\mathcal{V}_{n,i}$ defined as follows. We obtain $G_{n,i}$ by a one-to-one mapping of vertices of $G_i(\tilde{X}_i, Q_i)$ to the elements of $\mathcal{V}_{n,i}$ and by retaining the adjacency relations of $G_i(\tilde{X}_i, Q_i)$. We denote by $\mathcal{E}_{n,i}$ the edge set of $G_{n,i}$. Finally, let $G_{[n,m]} = (V, \mathcal{E})$ be the random graph with the vertex set $V = [n]$ and edge set $\mathcal{E} = \mathcal{E}_{n,1} \cup \dots \cup \mathcal{E}_{n,m}$. Therefore, $G_{[n,m]}$ is the superposition of the layers (communities) $G_{n,1}, \dots, G_{n,m}$.

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

The random graph $G_{[n,m]}$ represents a null model of the community affiliation graph model (AGM) introduced in [28, 29], which has attracted considerable attention in the literature. It is worth mentioning that community memberships (i.e. the vertex sets of respective overlapping communities) in the AGM [28, 29] are defined by a design that features non-negligible overlaps, whereas the null model $G_{[n,m]}$ assumes that $\mathcal{V}_{n,1}, \dots, \mathcal{V}_{n,m}$ are located at random and, therefore, their overlaps are typically small. (In particular, for $\mathbb{E}\, X < \infty$ and $m = \Theta(n)$ the expected number of overlaps is linear in m as $n, m \to +\infty$. Moreover, most of the overlaps are one-element sets.) We also mention that in the particular case where $Q \equiv 1$ the random graph $G_{[n,m]}$ reduces to a union of randomly located cliques of variable sizes $\tilde{X}_1, \dots, \tilde{X}_m$. This model has been studied in the literature under the name 'passive' random intersection graph; see, e.g., [10].

In the parameter regime $m = \Theta(n)$ as $m, n \to +\infty$ the random graph $G_{[n,m]}$ admits a power-law degree distribution with tunable power-law exponent, a nonvanishing global clustering coefficient, and a tunable clustering spectrum [6]. Moreover, it admits a limiting bidegree distribution with (stochastically dependent) power-law marginals, as shown in [7]. The present paper continues the study of the random graph $G_{[n,m]}$ and focuses on the asymptotic distributions of (dense) subgraph counts.

1.2. Results

Let $F = (\mathcal{V}_F, \mathcal{E}_F)$ be a graph with vertex set $\mathcal{V}_F$ and edge set $\mathcal{E}_F$. We write $v_F = |\mathcal{V}_F|$ and $e_F = |\mathcal{E}_F|$. We assume in what follows that F is 2-connected. That is, F is connected and, moreover, it stays connected even if we remove any one of its vertices. We call F balanced if $e_F/v_F = \max\{e_H/v_H \colon H \subset F \text{ with } e_H \geq 1\}$. For example, the cycle $\mathcal{C}_k$ and clique $\mathcal{K}_k$ (where k stands for the number of vertices) are 2-connected and balanced. Let N_F be the number of copies of F in $G(X, Q)$. By a copy of F we mean a graph isomorphic to F. Denote by $\sigma_F^2 = \operatorname{Var} N_F$ the variance of N_F. We write $\sigma_F^2 < \infty$ if the variance is finite and $\sigma_F^2 = \infty$ otherwise. We use the shorthand notation $N_F^* := \mathbb{E}(N_F \mid X, Q) = a_F \binom{X}{v_F} Q^{e_F}$, where a_F stands for the number of distinct copies of F in the complete graph on v_F vertices. We have, for example, that $N_{\mathcal{C}_k}^* = (X)_k Q^k/(2k)$ and $N_{\mathcal{K}_k}^* = (X)_k Q^{\binom{k}{2}}/k!$. Here and below $(x)_k = x(x-1)\cdots(x-k+1)$ denotes the falling factorial. Furthermore, we have $\mathbb{E}\, N_F = \mathbb{E}\, N_F^* = a_F \mathbb{E}\big(\binom{X}{v_F} Q^{e_F}\big)$.

In Theorems 1 and 2 and Remark 4 below we consider a sequence of random graphs $\{G_{[n,m]}, n = 1, 2, \dots\}$, where $m = m_n$ satisfies $m_n = \Theta(n)$ (i.e. both relations $m_n = O(n)$ and $n = O(m_n)$ hold) as $n \to +\infty$. We often suppress the subscript n for notational simplicity.

Let $\mathcal{N}_F$ be the number of copies of F in $G_{[n,m]}$. Our first result, Theorem 1, establishes the asymptotic normality of $\mathcal{N}_F$.

Theorem 1. *Let $m, n \to +\infty$ and assume that $m = \Theta(n)$. Let F be a 2-connected graph with $v_F \geq 3$ vertices. Assume that $\mathbb{E}\, X < \infty$ and $0 < \sigma_F^2 < \infty$. Assume, in addition, that*

$$\mathbb{E}\big(X^{1+s(1-1/2e_F)} Q^s\big) < \infty \qquad \text{for each } s = 1, 2, \dots, v_F - 1. \tag{1}$$

Then $(\mathcal{N}_F - m\mathbb{E}N_F)/(\sigma_F\sqrt{m})$ converges in distribution to the standard normal distribution.

Remark 1. For a balanced graph F, the finite variance condition $\sigma_F^2 < \infty$ is equivalent to the second moment condition $\mathbb{E}\big(N_F^*\big)^2 < \infty$. In particular, we have $\sigma_F^2 < \infty \Leftrightarrow \mathbb{E}(X^{2v_F} Q^{2e_F}) < \infty$.

Remark 2. In the special case where F is a clique on $k \geq 3$ vertices ($F = \mathcal{K}_k$), condition (1) can be replaced by

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

$$\mathbb{E}\big(X^{r-\hat{r}/(k(k-1))}Q^{\hat{r}}\big) < \infty \qquad \text{for each} \quad r = 2, \dots, k, \tag{2}$$

where we write $\hat{r} := \binom{r-1}{2} + 1$. Note that the moment condition (2) can be weaker than (1) for large k.

The proofs of Theorem 1 and Remarks 1 and 2 are presented in Section 2. Let us briefly explain the result and conditions of Theorem 1. Let $N_{F,i}$ be the number of copies of F in $G(X_i, Q_i)$, and define $S_F = N_{F,1} + \cdots + N_{F,m}$. The first moment condition $\mathbb{E}\, X < \infty$ and the assumption $m = \Theta(n)$ ensure that, with high probability, $\tilde{X}_i = X_i$, $1 \le i \le m$, i.e. the layer sizes do not need to be truncated. Next, from the fact that the typical overlap of two layers is either empty or a single-element set, we can deduce that (for 2-connected F) the principal contribution to the subgraph count $\mathcal{N}_F$ comes from the subgraph counts $N_{F,i}$ of individual layers. Therefore we have $\mathcal{N}_F \approx S_F$. To make this approximation rigorous we introduce conditions (1) and (2) aimed at controlling the number of overlaps of different copies of F in $G_{[n,m]}$. The combinatorial origin of (1) and (2) is explained in Lemmas 1–4. Finally, the asymptotic normality of $\mathcal{N}_F$ follows from the asymptotic normality of S_F. The latter is guaranteed by the second moment condition $\sigma_F^2 < \infty$.

In the case where F is balanced and the random variable N_F^* has an infinite second moment, we can obtain an α-stable limiting distribution for the subgraph count $\mathcal{N}_F$. In Theorem 2 we assume that, for some $a > 0$ and $0 < \alpha < 2$, we have

$$\mathbb{P}\big\{N_F^* > t\big\} = (a + o(1))t^{-\alpha} \qquad \text{as} \quad t \to +\infty. \tag{3}$$

Let $N_{F,i}^* = \mathbb{E}(N_{F,i} \mid X_i, Q_i)$, $1 \le i \le m$, be independent and identically distributed (i.i.d.) copies of N_F^*, and put $S_F^* = N_{F,1}^* + \cdots + N_{F,m}^*$. It is well known [9, Theorem 2, §35] that the distribution of $m^{-1/\alpha}(S_F^* - B_m)$ converges to a stable distribution, say $G_{\alpha,a}$, which is defined by a and α. Here, $B_m = m\mathbb{E}\, N_F^* = \mathbb{E}\, N_F$ for $1 < \alpha < 2$ and $B_m \equiv 0$ for $0 < \alpha < 1$. For $\alpha = 1$ we have $B_m = c_{\alpha,a}^{\star} \ln m$, where the constant $c_{\alpha,a}^{\star} > 0$ depends on a and α.

Our second result establishes an α-stable approximation to the distribution of $\mathcal{N}_F$.

Theorem 2. *Let $n, m \to +\infty$ and assume that $m = \Theta(n)$. Let F be a balanced and 2-connected graph with $v_F \ge 3$ vertices. Let $a > 0$ and $0 < \alpha < 2$. Assume that $\mathbb{E}X < \infty$ and that (3) holds. Assume, in addition, that*

$$\mathbb{E}\big(X^{1+s(1-1/\alpha e_F)}Q^{s}\big) < \infty \qquad \textit{for each} \quad s = 1, \dots, v_F - 1. \tag{4}$$

Then $(\mathcal{N}_F - B_m)/m^{1/\alpha}$ converges in distribution to $G_{\alpha,a}$.

Remark 3. In the special case where F is a clique on $k \ge 3$ vertices ($F = \mathcal{K}_k$), condition (4) can be replaced by

$$\mathbb{E}\big(X^{r-\hat{r}(2/(\alpha k(k-1)))}Q^{\hat{r}}\big) < \infty \qquad \text{for each} \quad r = 2, \dots, k, \tag{5}$$

where $\hat{r} = \binom{r-1}{2} + 1$.

The result of Theorem 2 is obtained by the approximations $\mathcal{N}_F \approx S_F$ and $S_F \approx S_F^*$. To make the latter approximation rigorous we apply exponential large-deviation bounds [15] combined with Janson's inequality [14, Theorem 2.14] to individual subgraph counts $N_{F,i}$ conditionally given (X_i, Q_i); see Lemma 5. (At this step we use the assumption that F is balanced.) The α-stable limit of S_F^* is now guaranteed by condition (3) and [9, Theorem 2, §35].

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

We briefly comment on the technical conditions (1), (2), (4), and (5). The mixed moments defined there appear in our upper bounds on the expected number of overlaps of different copies of F in $G_{[n,m]}$; see Lemmas 1 and 4 and inequality (10) in the proof below. We note that, for particular graphs F, the moment conditions (1), (2), (4), and (5) can be relaxed. For example, in the simplest case where $F = \mathcal{K}_2$ such conditions are not needed at all.

Remark 4. Let $F = \mathcal{K}_2$. Let $n, m \to +\infty$. Assume that $m = \Theta(n)$ and $\mathbb{E}\, X < \infty$.

(i) Assume that $0 < \sigma_{\mathcal{K}_2} < \infty$. Then $(\mathcal{N}_F - m\mathbb{E}N_F)/(\sigma_F\sqrt{m})$ converges in distribution to the standard normal distribution. Here, $\sigma_F^2 = \mathrm{Var}\binom{X}{2}Q) + \mathbb{E}\binom{X}{2}Q(1-Q)) < \infty$ whenever $\mathbb{E}(X^4Q^2) < \infty$.

(ii) Assume that, for some $a > 0$ and $0.5 < \alpha < 2$, condition (3) holds. Then $(\mathcal{N}_F - B_m)/m^{1/\alpha}$ converges in distribution to $G_{\alpha,a}$. We note that $\mathbb{E}\, X < \infty$ implies $\alpha > 0.5$.

Let us examine Theorems 1 and 2 in the special case where the marginals X, Q of (X, Q) are independent and $\mathbb{P}\{Q > 0\} > 0$. We first consider Theorem 1. The finite variance condition $\sigma_F^2 < \infty$ of Theorem 1 reduces to the moment condition $\mathbb{E}\, X^{2v_F} < \infty$. Indeed, by the simple inequality $N_F \le (X)_{v_F}$, we have that $\mathbb{E}\, X^{2v_F} < \infty \Rightarrow \mathbb{E}\, N_F^2 < \infty \Rightarrow \sigma_F^2 < \infty$. On the other hand, by the variance identity $\mathrm{Var}\, N_F = \mathrm{Var}\, N_F^* + \mathbb{E}(\mathrm{Var}(N_F \mid X, Q))$, we have that $\sigma_F^2 < \infty \Rightarrow \mathbb{E}\big(N_F^*\big)^2 < \infty$, where the latter inequality (for independent X and Q) implies $\mathbb{E}\, X^{2v_F} < \infty$. Moreover, the moment condition $\mathbb{E}\, X^{2v_F} < \infty$ implies (1). Therefore, Theorem 1 establishes the asymptotic normality under the minimal second moment condition $\sigma_F^2 < \infty$.

We now turn to Theorem 2. For independent X and Q condition (3) of Theorem 2 is equivalent to the condition

$$\mathbb{P}\{X > t\} = (b + o(1))t^{-\gamma} \qquad \text{as} \quad t \to +\infty, \tag{6}$$

where $\gamma = \alpha v_F$ and where b solves the equation $a = b(a_F/v_F!)^{\gamma/v_F}\mathbb{E}\, Q^{\gamma e_F/v_F}$. Note that $\mathbb{E}\, X < \infty$ implies $\gamma > 1$. Furthermore, the inequality $v_F \le e_F$ (which holds for any 2-connected F with $v_F \ge 3$) combined with $\gamma > 1$ implies $\alpha e_F > 1$. Observe that, for $\alpha e_F > 1$, condition (4) reads as $\mathbb{E}\, X^{1+(v_F-1)(1-1/\alpha e_F)} < \infty$. In view of (6), the latter expectation is finite whenever

$$1 + (v_F - 1)\left(1 - \frac{1}{\alpha e_F}\right) < \gamma. \tag{7}$$

We have arrived at the following corollary.

Corollary 1. *Let $n, m \to +\infty$ and assume that $m = \Theta(n)$. Let F be a 2-connected graph with $v_F \ge 3$ vertices. Assume that X and Q are independent and $\mathbb{P}\{Q > 0\} > 0$.*

(i) *If $\mathbb{E}\, X^{2v_F} < \infty$ then $(\mathcal{N}_F - \mathbb{E}\,\mathcal{N}_F)/(\sigma_F\sqrt{m})$ converges in distribution to the standard normal distribution.*

(ii) *Let $b > 0$ and $1 < \gamma < 2v_F$. Assume that (6) holds. Assume, in addition, that F is balanced and (7) holds, where $\alpha = \gamma/v_F$. Then $(\mathcal{N}_F - B_m)/m^{1/\alpha}$ converges in distribution to $G_{\alpha,a}$. Here, B_m and $G_{\alpha,a}$ are the same as in Theorem 2, with $a = b(a_F/v_F!)^{\gamma/v_F}\mathbb{E}\, Q^{\gamma e_F/v_F}$.*

It is relevant to mention that the moment condition $\mathbb{E}\, X < \infty$ together with the assumption $m = vn + o(n)$ for some $v > 0$ (which is stronger than $m = \Theta(n)$), imply the existence of

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

an asymptotic degree distribution of $G_{[n,m]}$ as $n, m \to +\infty$. An asymptotic power-law degree distribution is obtained if we choose an appropriate distribution for the layer type (X, Q). Furthermore, under an additional moment condition $\mathbb{E}\, X^3 Q^2 < \infty$, the random graph $G_{[n,m]}$ has a nonvanishing global clustering coefficient; see [6]. Therefore, Theorems 1 and 2 establish the limit distributions of subgraph counts in a highly clustered complex network.

Finally, we discuss an important question about the relation between the community size X and strength Q. In Theorems 1 and 2, no assumption has been made about the stochastic dependence between the marginals X and Q of the bivariate random vector (X, Q) defining the random graph $G_{[n,m]}$. Although we can simplify the model by assuming that X and Q are independent (as in Corollary 1), for network modeling purposes, various types of dependence between X and Q are of interest. For example, a negative correlation between X and Q would emphasise small strong communities and large weak communities, a pattern likely to occur in real networks with overlapping communities. Assuming that Q is proportional to a negative power of X, for example, $Q = \min\{1, bX^{-\beta}\}$ for some $\beta \geq 0$ and $b > 0$ (cf. [28, 29]), and we obtain a mathematically tractable network model admitting tunable power-law degree and bidegree distributions and a rich clustering spectrum [6, 7].

1.3. Related work

Asymptotic distributions of subgraph counts in Bernoulli random graphs is a well-established area of research, see, e.g., [14, 23] and references therein. For a recent development we refer to [3, 12, 17, 21, 22, 30]. A significant difference between the sparse Bernoulli random graphs and complex networks is that the former have no or very few copies of a triangle or a larger clique, while the latter often have abundant numbers of them. Since the global and local clustering coefficients are expressed in terms of counts of triangles and wedges, a rigorous asymptotic analysis of clustering coefficients reduces to that of the triangle counts and wedge counts. In particular, the bivariate asymptotic normality for triangle and wedge counts in a related sparse random intersection graph was shown in [4], and related α-stable limits were established in [5]. Another line of research pursued in [11, 16] addresses the concentration of subgraph counts in $G_{[n,m]}$. We also mention related work on local weak limits and subgraph counts: the results of [18, 26] imply the linear growth in n of the numbers of small dense subgraphs for a large class of sparse affiliation network models. Establishing the distributional asymptotics here is an interesting problem for future research. Another interesting question is about revoking the 2-connectivity and balancedness conditions on F in Theorems 1 and 2.

The rest of the paper is organised as follows. In Section 2 we formulate and prove Theorems 1 and 2 and Remarks 1–4. We mention that the combinatorial Lemmas 2 and 3 and inequality (17) may be of independent interest.

2. Proofs

2.1. Notation

Before the proof we introduce some notation. Let $\mathcal{K}$ be the complete graph with vertex set $V = [n]$ so that $G_{[n,m]} \subset \mathcal{K}$. By $\mathbb{E}^*(\,\cdot\,) = \mathbb{E}(\cdot \mid X, X_1, \ldots, X_m, Q, Q_1, \ldots, Q_m)$ we denote the conditional expectation given $X, X_1, \ldots, X_m, Q, Q_1, \ldots, Q_m$. Given F, for any positive sequences $\{a_n\}$ and $\{b_n\}$ we write $a_n \asymp b_n$ (respectively $a_n \prec b_n$) whenever, for sufficiently large n, we have $c_1 \leq a_n/b_n \leq c_2$ (respectively $a_n \leq c_2 b_n$), where constants $0 < c_1 < c_2$ may

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

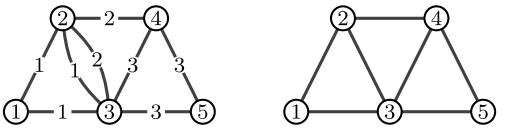

FIGURE 1. Multigraph $G^{\star}_{[5,3]}$ and overlay graph $G_{[5,3]}$.

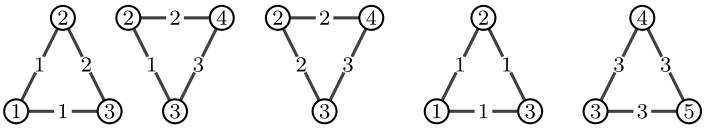

FIGURE 2. Three polychromatic and two monochromatic copies of $\mathcal{K}_3$ in $G^{\star}_{[5,3]}$.

only depend on F. For a sequence of random variables $\{Y_n\}$ we write $Y_n = o_P(a_n)$ whenever $\lim_{n\to\infty} \mathbb{P}\{|Y_n| < \varepsilon |a_n|\} = 1$ for any $\varepsilon > 0$; and $Y_n = O_P(a_n)$ if, for every $\varepsilon > 0$, there exists a constant $c_\varepsilon > 0$ such that $\lim_{n\to\infty} \mathbb{P}\{|Y_n| < c_\varepsilon |a_n|\} > 1 - \varepsilon$.

Recall that N_F and $N_{F,i}$ denote the numbers of copies of F in $G(X, Q)$ and $G(X_i, Q_i)$, respectively. Furthermore, $N^*_F = \mathbb{E}(N_F \mid X, Q)$, $N^*_{F,i} = \mathbb{E}(N_{F,i} \mid X_i, Q_i)$, and $S_F = N_{F,1} + \cdots + N_{F,m}$, $S^*_F = N^*_{F,1} + \cdots + N^*_{F,m}$. Note that $N^*_{F,i} = \mathbb{E}^*(N_{F,i})$ and $S^*_F = \mathbb{E}^*(S_F)$. Finally, let $\tilde{N}_{F,i}$ be the number of copies of F in $G_{n,i}$, and let $\tilde{S}_F = \tilde{N}_{F,1} + \cdots + \tilde{N}_{F,m}$.

We can identify the indices $1 \le i \le m$ with colours, and assign (the edges of) each $G_{n,i}$ the colour i. The coloured graph is denoted by $G^{\star}_{n,i}$. The union of coloured graphs $G^{\star}_{n,1} \cup \cdots \cup G^{\star}_{n,m}$ defines a multigraph, denoted by $G^{\star}_{[n,m]}$, where parallel edges have different colours. Furthermore, each edge $u \sim v$ of $G_{[n,m]}$ is assigned the set of colours that correspond to parallel edges of $G^{\star}_{[n,m]}$ connecting u and v.

A subgraph $H \subset G_{[n,m]}$ is called monochromatic if it is a subgraph of some $G_{n,i}$ and none of the edges of H are assigned more than one colour. Otherwise H is called polychromatic. $\mathcal{N}_{F,M}$ and $\mathcal{N}_{F,P}$ stand for the numbers of monochromatic and polychromatic copies of F in $G_{[n,m]}$. A subgraph $H^{\star} \subset G^{\star}_{[n,m]}$ is called monochromatic if it is a subgraph of some $G^{\star}_{n,i}$. It is called polychromatic if it contains edges of different colours. Let $\mathcal{N}^{\star}_{F,P}$ be the number of polychromatic copies of F in $G^{\star}_{[n,m]}$.

Figure 1 depicts an instance of the overlay graph $G_{[5,3]}$ and respective multigraph $G^{\star}_{[5,3]} = G^{\star}_{5,1} \cup G^{\star}_{5,2} \cup G^{\star}_{5,3}$, where $G^{\star}_{5,i}$ has vertex set $\mathcal{V}_{5,i} = \{i, i+1, i+2\}$ and edges labelled (coloured) i. $G^{\star}_{[5,3]}$ has three polychromatic and two monochromatic copies of $\mathcal{K}_3$ (Figure 2), while $G_{[5,3]}$ has two polychromatic copies of $\mathcal{K}_3$ (induced by $\{1, 2, 3\}$ and $\{2, 3, 4\}$) and one monochromatic copy of $\mathcal{K}_3$ (induced by $\{3, 4, 5\}$).

Given $H^{\star} = (\mathcal{V}_{H^{\star}}, \mathcal{E}_{H^{\star}}) \subset G^{\star}_{[n,m]}$, let $H_0 \subset \mathcal{K}$ be the graph on the vertex set $\mathcal{V}_{H^{\star}}$ obtained from $H^{\star}$ as follows: two vertices of H_0 are adjacent whenever they are joined by an edge in $H^{\star}$. We call H_0 the projection of $H^{\star}$. Note that there can be several monochromatic and/or polychromatic copies of F in $G^{\star}_{[n,m]}$ sharing the same projection F_0. We fix a copy F_0 of F in $\mathcal{K}$ and denote by h_F the expected number of polychromatic copies of F in $G^{\star}_{[n,m]}$ whose projection is F_0. By the symmetry of the random graph model $G^{\star}_{[n,m]}$, the quantity h_F does not depend on the location of F_0 in $\mathcal{K}$. An expression of h_F in terms of mixed moments $\mathbb{E}((\tilde{X}_1)_s Q_1^t)$ is given in (11) and (12).

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

2.2. Proofs

We first prove Theorems 1 and 2, and Remarks 2 and 3. Afterwards we prove Remarks 4 and 1.

We start with an outline of the proof of Theorems 1 and 2. We approximate $\mathcal{N}_F \approx \tilde{S}_F$ and $\tilde{S}_F \approx S_F$. In the case where $\mathbb{E}\, N_F^2 < \infty$ we deduce the normal approximation to the sum S_F (of i.i.d. random variables) by the standard central limit theorem. In the case where N_F has an infinite variance we further approximate $S_F \approx S_F^*$ and deduce the α-stable approximation by the generalised central limit theorem [9, Theorem 2, §35].

Approximation $\mathcal{N}_F \approx \tilde{S}_F$ The approximation follows from the simple observation that

$$\mathcal{N}_F = \mathcal{N}_{F,M} + \mathcal{N}_{F,P}, \qquad \mathcal{N}_{F,M} \le \tilde{S}_F \le \mathcal{N}_{F,M} + \mathcal{N}^\star_{F,P}, \qquad \mathcal{N}_{F,P} \le \mathcal{N}^\star_{F,P}. \tag{8}$$

The inequalities $\mathcal{N}_{F,M} \le \tilde{S}_F$ and $\mathcal{N}_{F,P} \le \mathcal{N}^\star_{F,P}$ are easy. To see why the inequality $\tilde{S}_F \le \mathcal{N}_{F,M} + \mathcal{N}^\star_{F,P}$ holds true, let us inspect a pair $F_i \in G_{n,i}$ and $F_j \in G_{n,j}$ of copies of $F = (\mathcal{V}_F, \mathcal{E}_F)$ that share $t := |\mathcal{E}_{F_i} \cap \mathcal{E}_{F_j}| \ge 1$ edges. Note that both copies F_i and F_j contribute to the sum $\tilde{S}_F$, and neither contributes to the sum $\mathcal{N}_{F,M}$. In the case where $t < |\mathcal{E}_F|$ the pair gives rise to $2 \cdot 2^t - 2 \ge 2$ polychromatic copies of F in $G^\star_{[n,m]}$. In the case where $t = |\mathcal{E}_F|$ (now $t \ge 3$) the pair gives rise to $2^t - 2$ polychromatic copies of F in $G^\star_{[n,m]}$. Hence, $\tilde{S}_F \le \mathcal{N}_{F,M} + \mathcal{N}^\star_{F,P}$. From (8) we conclude that

$$|\tilde{S}_F - \mathcal{N}_F| \le \mathcal{N}^\star_{F,P}. \tag{9}$$

In order to assess the accuracy of the approximation $\mathcal{N}_F \approx \tilde{S}_F$ we evaluate the expected value of $\mathcal{N}^\star_{F,P}$. We fix a copy of F in $\mathcal{K}$, denoted $F_0 = (\mathcal{V}_0, \mathcal{E}_0) \subset \mathcal{K}$, with vertex set $\mathcal{V}_0 = \{1, \dots, v_F\}$. Recall that h_F denotes the expected number of polychromatic copies of F in $G^\star_{[n,m]}$ whose projection is F_0. We have, by symmetry,

$$\mathbb{E}\, \mathcal{N}^\star_{F,P} = \binom{n}{v_F} a_F h_F. \tag{10}$$

Note that every polychromatic copy of F in $G^\star_{[n,m]}$ (say, $F^\star \subset G^\star_{[n,m]}$) whose projection is F_0 is defined by a partition of the edge set $\mathcal{E}_0$ into non-empty colour classes, say, $B_1 \cup \dots \cup B_r = \mathcal{E}_0$, and a vector of distinct colours $(i_1, \dots, i_r) \in [m]^r$ such that all the edges in B_j are of colour i_j (edges of B_j belong to $G^\star_{n,i_j}$). Denote by $\tilde{B} = (B_1, \dots, B_r)$ and $\tilde{\imath} = (i_1, \dots, i_r)$ the partition and its colouring. The polychromatic subgraph $F^\star$ defined by the pair $(\tilde{B}, \tilde{\imath})$ is denoted $F(\tilde{B}, \tilde{\imath})$. The probability that such a subgraph is present in $G^\star_{[n,m]}$ is

$$h(\tilde{B}, \tilde{\imath}) := \mathbb{P}\Big\{F(\tilde{B}, \tilde{\imath}) \subset G^\star_{[n,m]}\Big\} = \prod_{j=1}^{r} \frac{1}{(n)_{v_j}} \mathbb{E}\Big((\tilde{X}_{i_j})_{v_j} Q_{i_j}^{b_j}\Big). \tag{11}$$

Here, $b_j := |B_j|$, and v_j is the number of distinct vertices incident to edges from B_j. We have

$$h_F = \mathbb{E}\left(\sum_{(\tilde{B},\tilde{\imath})} \mathbf{1}_{\left\{F(\tilde{B},\tilde{\imath}) \in G^*_{[n,m]}\right\}}\right) = \sum_{(\tilde{B},\tilde{\imath})} h(\tilde{B}, \tilde{\imath}). \tag{12}$$

Here, the sum runs over all possible polychromatic copies $F^\star$ of F whose projection is F_0. We upper bound the quantity on the right of (12) in Lemmas 1 and 4 below.

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

Approximation $\tilde{S}_F \approx S_F$ For $1 \le i \le m$ we couple $G(\tilde{X}_i, Q_i) \subset G(X_i, Q_i)$ and $\tilde{N}_{F,i} \le N_{F,i}$ so that $G(\tilde{X}_i, Q_i) \ne G(X_i, Q_i)$ and $\tilde{N}_{F,i} \ne N_{F,i}$ whenever $X_i > n$. For $m = O(n)$, the event $\mathcal{A}_n := \{\max_{1\le i\le m} X_i > n\}$ has probability

$$\mathbb{P}\{\mathcal{A}_n\} \le \sum_{i=1}^{m} \mathbb{P}\{X_i > n\} \le \frac{m}{n}\mathbb{E}\big(X_1 \mathbf{1}_{\{X_1>n\}}\big) = o(1). \tag{13}$$

Hence, $\mathbb{P}\{\tilde{S}_F \ne S_F\} = o(1)$. In (13) we used the fact that $\mathbb{E}\,X_1 < \infty \Rightarrow \mathbb{E}\big(X_1 \mathbf{1}_{\{X_1>n\}}\big) = o(1)$.

Proof of Theorem 1 and Remark 2. By Lemma 1 (respectively, Lemma 4), we have $h_F = o(n^{0.5-v_F})$. Invoking this bound in (10), we obtain $\mathcal{N}^{\star}_{F,P} = o_P(\sqrt{m})$. Next, from (9) we obtain that $(\mathcal{N}_F - \tilde{S}_F) = o_P(\sqrt{m})$. Then, an application of (13) shows that $(\mathcal{N}_F - S_F) = o_P(\sqrt{m})$. Finally, we apply the classical central limit theorem to the sum of i.i.d. random variables S_F to get the asymptotic normality of $(\mathcal{N}_F - m\mathbb{E}N_F)/(\sigma_F\sqrt{m})$. □

Proof of Theorem 2 and Remark 3. By Lemma 1 (respectively, Lemma 4), we have $h_F = o(n^{(1/\alpha)-v_F})$. Using this bound and proceeding as in the proof of Theorem 1, we obtain $\mathcal{N}_F = S_F + o_P(m^{1/\alpha})$. Next, from the fact that the random variables $N_{F,1}, N_{F,2}, \ldots$ obey the power law (28) (see Lemma 5), we conclude that $(S_F - B_m)/m^{1/\alpha}$ converges in distribution to $G_{\alpha,a}$ [9, Theorem 2, §35]. Hence, $(\mathcal{N}_F - B_m)/m^{1/\alpha}$ converges in distribution to $G_{\alpha,a}$. □

Proof of Remark 4. We have $\mathcal{N}_{\mathcal{K}_2} = |\mathcal{E}| = |\mathcal{E}_{n,1} \cup \cdots \cup \mathcal{E}_{n,m}|$ and $\tilde{S}_{\mathcal{K}_2} = \sum_{i=1}^{n} |\mathcal{E}_{n,i}|$. By the inclusion–exclusion principle,

$$0 \le \sum_{i=1}^{m} |\mathcal{E}_{n,i}| - |\mathcal{E}| \le \sum_{\{i,j\}\subset[m]} |\mathcal{E}_{n,i} \cap \mathcal{E}_{n,j}|. \tag{14}$$

We write $|\mathcal{E}_{n,i} \cap \mathcal{E}_{n,j}| = \sum_{\{u,v\}\subset V} \mathbf{1}_{\{\{u,v\}\in\mathcal{E}_{n,i}\}} \mathbf{1}_{\{\{u,v\}\in\mathcal{E}_{n,j}\}}$ and evaluate the conditional expectation

$$\mathbb{E}^*|\mathcal{E}_{n,i} \cap \mathcal{E}_{n,j}| = \binom{n}{2} \frac{(\tilde{X}_i)_2 Q_i}{(n)_2} \frac{(\tilde{X}_j)_2 Q_j}{(n)_2}.$$

To prove (i), in view of the identity $\sigma^2_{\mathcal{K}_2} = \mathrm{Var}\big(\binom{X}{2}Q\big) + \mathbb{E}\big(\binom{X}{2}Q(1-Q)\big)$ we have $\sigma^2_{\mathcal{K}_2} < \infty \Leftrightarrow \mathbb{E}\big(X^4Q^2\big) < \infty$. Hence, $\sigma^2_{\mathcal{K}_2} < \infty$ implies $\infty > \mathbb{E}(X^4Q^2) \ge (\mathbb{E}(X^2Q))^2$, by Cauchy–Schwarz. Consequently, the expected value of the quantity on the right of (14) is

$$\mathbb{E}\left(\sum_{\{i,j\}\subset[m]} \left(\frac{(\tilde{X}_i)_2 Q_i (\tilde{X}_j)_2 Q_j}{2(n)_2}\right)\right) = \binom{m}{2}\left(\frac{(\mathbb{E}((\tilde{X}_1)_2 Q_1))^2}{2(n)_2}\right) = O(1).$$

Now, (14) implies $\mathcal{N}_{\mathcal{K}_2} = \tilde{S}_{\mathcal{K}_2} + O_P(1)$. Next, (13) implies $(\mathcal{N}_{\mathcal{K}_2} - S_{\mathcal{K}_2})/(\sigma_{\mathcal{K}_2}\sqrt{m}) = o_P(1)$. Finally, the asymptotic normality of $(\mathcal{N}_{\mathcal{K}_2} - m\mathbb{E}N_{\mathcal{K}_2})/(\sigma_{\mathcal{K}_2}\sqrt{m})$ follows by the classical central limit theorem applied to the sum $S_{\mathcal{K}_2} = \sum_{i\in[m]} N_{\mathcal{K}_2,i}$.

To prove (ii), we have $N^*_{\mathcal{K}_2} = \binom{X}{2}Q$. Observing that (3) implies $\mathbb{P}\{X^2 > t\} \ge \mathbb{P}\{N^*_{\mathcal{K}_2} > t\} = (a + o(1))t^{-\alpha}$, we obtain from the first moment condition $\mathbb{E}\,X < \infty$ that $\alpha > 0.5$.

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

Let R denote the quantity on the right of (14), and put $R^* = \mathbb{E}^* R$. We first show that $R = o_P(m^{1/\alpha})$. Note that $R^* \le 4m^{2/\alpha} n^{-2} T_*^2$, where $T_* := m^{-1/\alpha} \sum_{i\in[m]} N^*_{\mathcal{K}_2,i}$. Given $\varepsilon \in (0,1)$, we have, for $A = \varepsilon m^{1/\alpha}$ and $B = \varepsilon A$,

$$\mathbb{P}\{R > \varepsilon m^{1/\alpha}\} = \mathbb{P}\{R > A\} \le \mathbb{P}\{R > A, R^* \le B\} + \mathbb{P}\{R^* > B\} \le \varepsilon + o(1). \tag{15}$$

Indeed, $\mathbb{P}\{R^* > B\} \le \mathbb{P}\{4m^{2/\alpha} n^{-2} T_*^2 > B\} = \mathbb{P}\{4T_*^2 > m^{-1/\alpha} n^2 \varepsilon^2\} = o(1)$, since $m^{-1/\alpha} n^2 \varepsilon^2 \to +\infty$ for $\alpha > 0.5$ and $T_* = O_P(1)$ by (3). Furthermore, by Markov's inequality,

$$\mathbb{P}\{R > A, R^* \le B\} = \mathbb{E}\big(\mathbb{E}^*(\mathbf{1}_{\{R>A\}} \mathbf{1}_{\{R^* \le B\}})\big) \le \mathbb{E}\left(\frac{R^*}{A} \mathbf{1}_{\{R^* \le B\}}\right) \le \frac{B}{A} = \varepsilon.$$

Clearly, (15) implies the bound $R = o_P(m^{1/\alpha})$. Now, (14) implies $\mathcal{N}_{\mathcal{K}_2} = \tilde{S}_{\mathcal{K}_2} + o_P(m^{1/\alpha})$. Next, (13) implies $(\mathcal{N}_{\mathcal{K}_2} - S_{\mathcal{K}_2}) m^{-1/\alpha} = o_P(1)$. In the last step of the proof we show that $(S_{\mathcal{K}_2} - B_m)/m^{1/\alpha}$ converges in distribution to $G_{\alpha,a}$ using the same argument as in the proof of Theorem 2. □

Proof of Remark 1. We have $\sigma_F^2 = \operatorname{Var} N_F = \operatorname{Var} N_F^* + \mathbb{E}\big(\Delta_F^*\big)^2$, where $\Delta_F^* := N_F - N_F^*$. Therefore, $\sigma_F^2 < \infty \Rightarrow \operatorname{Var} N_F^* < \infty \Rightarrow \mathbb{E}\big(N_F^*\big)^2 < \infty$. To prove that $\mathbb{E}\big(N_F^*\big)^2 < \infty \Rightarrow \sigma_F^2 < \infty$, it suffices to show that $\mathbb{E}\big(\Delta_F^*\big)^2 < \infty$. By [14, Lemma 3.5], we have $\mathbb{E}^*\big(\Delta_F^*\big)^2 \prec \big(N_F^*\big)^2 / \Phi_F(X, Q)$, where $\Phi_F(X, Q) = \min_{H\subset F} X^{v_H} Q^{e_H}$. Furthermore, from the inequality in (27), which holds for balanced F, we obtain

$$\mathbb{E}^*\big(\Delta_F^*\big)^2 \prec \frac{\big(N_F^*\big)^2}{\min\left\{\big(N_F^*\big)^{2/v_F}, N_F^*\right\}} = \max\left\{\big(N_F^*\big)^{2-2/v_F}, N_F^*\right\} \le \max\left\{1, \big(N_F^*\big)^2\right\}.$$

Hence, $\mathbb{E}\big(N_F^*\big)^2 < \infty$ implies $\mathbb{E}\big(\Delta_F^*\big)^2 = \mathbb{E}\Big(\mathbb{E}^*\big(\Delta_F^*\big)^2\Big) < \infty$. □

2.3. Auxiliary lemmas

In Lemmas 1 and 4 we upper bound the quantities h_F for 2-connected F and for $F = \mathcal{K}_k$, respectively. Clearly, the result of Lemma 1 applies to $F = \mathcal{K}_k$ as well, but the bound of Lemma 4 is tighter for large k.

Lemma 1. *Let F be a 2-connected graph with $v_F \ge 3$ vertices. Let $n, m \to +\infty$. Assume that $m = O(n)$.*

(i) *Assume that (1) holds. Then $h_F = o\big(n^{0.5-v_F}\big)$.*

(ii) *Assume that $0 < \alpha < 2$, and that (4) holds. Then $h_F = o\big(n^{(1/\alpha)-v_F}\big)$.*

In the proof we use the simple fact that, for any $s, t, \tau > 0$, the moment condition $\mathbb{E}(X^s Q^t) < \infty$ implies

$$\mathbb{E}\big((\min\{X, n\})^{s+\tau} Q^t\big) = o(n^\tau). \tag{16}$$

Write $\tilde{X} := \min\{X, n\}$. To see why (16) holds, choose $0 < \delta < \tau/(s+\tau)$ and split the expectation:

$$\mathbb{E}(\tilde{X}^{s+\tau} Q^t) = \mathbb{E}\big(\tilde{X}^{s+\tau} Q^t \mathbf{1}_{\{X < n^\delta\}}\big) + \mathbb{E}\big(\tilde{X}^{s+\tau} Q^t \mathbf{1}_{\{X \ge n^\delta\}}\big) =: I_1 + I_2.$$

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

The inequalities $\tilde{X} \le n$ and $\mathbb{E}(X^s Q^t) < \infty$ imply $I_2 \le n^\tau \mathbb{E}\big(X^s Q^t \mathbf{1}_{\{X \ge n^\delta\}}\big) = n^\tau \cdot o(1)$, and the inequality $\tilde{X} \le X$ implies $I_1 \le n^{\delta(s+\tau)} = o(n^\tau)$.

Proof of Lemma 1. The proofs of statements (i) and (ii) are identical. Therefore we only prove statement (i).

We start by establishing an auxiliary inequality, (17), which may be interesting in itself. Let $r \ge 2$. Given a partition $\tilde{B} = (B_1, \dots, B_r)$ of the edge set $\mathcal{E}_0$ of the graph $F_0 = (\mathcal{V}_0, \mathcal{E}_0)$, and given $i \in [r]$, let V_i be the set of vertices incident to the edges from B_i. Let ρ_i be the number of (connected) components of the graph $Z_i = (V_i, B_i)$, and put $v_i = |V_i|$. We claim that

$$v_1 + \dots + v_r \ge v_F + \rho_1 + \dots + \rho_r. \tag{17}$$

To establish the claim we consider the list $H_1, H_2, \dots, H_t$ of components of $Z_1, \dots, Z_r$ arranged in an arbitrary order. Here, $t := \rho_1 + \dots + \rho_r$. Therefore, each graph H_i is a component of some Z_j, and their union $H_1 \cup \dots \cup H_t = Z_1 \cup \dots \cup Z_r = F_0$. Let us consider the sequence of graphs $\bar{H}_j := H_1 \cup \dots \cup H_j$ for $j = 1, \dots, t-1$. Let $\bar{\rho}_j$ and $\bar{v}_j$ denote the number of components and the number of vertices of $\bar{H}_j$. Let v'_j denote the number of vertices of H_j. We use the observation that

$$\bar{v}_j \le \bar{v}_{j-1} + v'_j + \bar{\rho}_j - \bar{\rho}_{j-1} - 1, \qquad j = 2, \dots t-1. \tag{18}$$

Indeed, $\bar{\rho}_{j-1} = \bar{\rho}_j$ means that the vertex set of (the connected graph) H_j intersects with exactly one component of $\bar{H}_{j-1}$. Consequently, H_j and $\bar{H}_{j-1}$ have at least one common vertex and therefore (18) holds. Similarly, $\bar{\rho}_{j-1} - \bar{\rho}_j = y > 0$ means that the vertex set of H_j intersects with exactly $y+1$ different components of $\bar{H}_{j-1}$. Consequently, H_j and $\bar{H}_{j-1}$ have at least $y+1$ common vertices and (18) holds again. The remaining case, $\bar{\rho}_{j-1} - \bar{\rho}_j = -1$, is realised by the configuration where the vertex sets of H_j and $\bar{H}_{j-1}$ have no common elements. In this case (18) follows from the identity $\bar{v}_j = \bar{v}_{j-1} + v'_j$.

By summing the inequalities in (18), we obtain, using $\bar{\rho}_1 = 1$, that $\bar{v}_{t-1} \le v'_1 + \dots + v'_{t-1} + \bar{\rho}_{t-1} - t + 1$. Note that, given $\bar{H}_{t-1}$ with $\bar{\rho}_{t-1}$ components, the vertex set of H_t must intersect with each component in two or more points in order to make the union $\bar{H}_{t-1} \cup H_t = F_0$ 2-connected. Consequently, we have $\bar{v}_t \le \bar{v}_{t-1} + v'_t - 2\bar{\rho}_{t-1}$. Finally, we obtain $v_F = \bar{v}_t \le v'_1 + \dots + v'_t - \bar{\rho}_{t-1} - t + 1$. The claim follows from the identity $v'_1 + \dots + v'_t = v_1 + \dots + v_r$ and the inequality $\bar{\rho}_{t-1} \ge 1$.

To prove statement (i), given $(\tilde{B}, \tilde{\imath})$, we obtain from (11) and (17) (recall the notation $b_j = |B_j|$) that

$$h(\tilde{B}, \tilde{\imath}) \le \frac{1}{n^{v_1 + \dots + v_r}} \prod_{j=1}^{r} \mathbb{E}\big(\tilde{X}^{v_j} Q^{b_j}\big) \le \frac{1}{n^{v_F + \rho_1 + \dots + \rho_r}} \prod_{j=1}^{r} \mathbb{E}\big(\tilde{X}^{v_j} Q^{b_j}\big).$$

Given $\tilde{B} = (B_1, \dots, B_r)$, we estimate the sum over all possible colourings (there are $(m)_r$ of them):

$$\sum_{\tilde{\imath}} h(\tilde{B}, \tilde{\imath}) \prec \frac{(m)_r}{n^{v_F + \rho_1 + \dots + \rho_r}} \prod_{j=1}^{r} \mathbb{E}\big(\tilde{X}^{v_j} Q^{b_j}\big) \asymp n^{-v_F} \prod_{j=1}^{r} \frac{\mathbb{E}\big(\tilde{X}^{v_j} Q^{b_j}\big)}{n^{\rho_j - 1}}$$

$$= n^{0.5 - v_F} \prod_{j=1}^{r} \frac{\mathbb{E}\big(\tilde{X}^{v_j} Q^{b_j}\big)}{n^{\rho_j - 1 + (b_j/(2e_F))}} = o\big(n^{0.5 - v_F}\big).$$

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

In the second-last identity we used $b_1 + \cdots + b_r = e_F$, while the last bound follows by the chain of inequalities

$$n^{1-\rho_j}\mathbb{E}\big(\tilde{X}^{v_j}Q^{b_j}\big) \le \mathbb{E}\big(\tilde{X}^{v_j+1-\rho_j}Q^{b_j}\big) \le \mathbb{E}\big(\tilde{X}^{v_j+1-\rho_j}Q^{v_j-\rho_j}\big) = o\big(n^{(v_j-\rho_j)/(2e_F)}\big) = o\big(n^{b_j/(2e_F)}\big).$$

Here, in the first step we used $\tilde{X} \le n$; in the second step we used $Q \le 1$ and $b_j \ge v_j - \rho_j$ (the latter inequality is based on the observation that any graph with v_j vertices and ρ_j components has at least $v_j - \rho_j$ edges); the third step follows by (16) from the moment condition (1) applied to $s = v_j - \rho_j$; and the last step follows from the inequality $b_j \ge v_j - \rho_j$.

Finally, we conclude that

$$h_F = \sum_{\tilde{B}} \sum_{\tilde{i}} h(\tilde{B}, \tilde{i}) = o\big(n^{0.5-v_F}\big), \tag{19}$$

because the number of partitions $\tilde{B}$ of the edge set of a given graph F is always finite. □

Before showing an upper bound for h_F, $F = \mathcal{K}_k$, we introduce some notation. Given an integer $b \ge 1$, let $b^\star$ be the minimal number of vertices that a graph with b edges may have. Let H_b be such a graph. It has a simple structure described below. Let $k_b \ge 2$ be the largest integer satisfying $b \ge \binom{k_b}{2}$. Then $b = \binom{k_b}{2} + \Delta_b$ for some integer $0 \le \Delta_b \le k_b - 1$. For $\Delta_b = 0$ we have $b^\star = k_b$ and $H_b = \mathcal{K}_{b^\star}$ (clique on $b^\star = k_b$ vertices). For $\Delta_b > 0$, graph H_b is a union of $\mathcal{K}_{k_b}$ and a star $\mathcal{K}_{1,\Delta_b}$ such that all the vertices of the star except for the central vertex belong to the vertex set of $\mathcal{K}_{k_b}$. In this case, $b^\star = k_b + 1$. In other words, we obtain H_b from $\mathcal{K}_{k_b+1}$ by deleting $k_b - \Delta_b$ edges sharing a common endpoint. The next two lemmas establish useful properties of the function $b \to b^\star$.

Lemma 2. *For integers $s \ge t \ge 1$,*

$$s^\star + t^\star \ge (s + t - 1)^\star + 2. \tag{20}$$

Proof. We consider graphs H_s and H_t that have disjoint vertex sets so that the union $H_s \cup H_t$ has $s^\star + t^\star$ vertices.

Note that for $t = 1$ both sides of (20) are equal. In order to show (20) for $s \ge t \ge 2$ we consider the chain of neighbouring pairs

$$(s, t) \to (s + 1, t - 1) \to \cdots \to (s + t - 1, 1). \tag{21}$$

In a step $(x, y) \to (x + 1, y - 1)$ we remove an edge from H_y and add it to H_x. A simple analysis of the step $(H_x, H_y) \to (H_{x+1}, H_{y-1})$ shows that

$$(x+1)^\star + (y-1)^\star = x^\star + y^\star + 1 \quad \text{whenever } \Delta_x = 0,\ \Delta_y \ne 1; \tag{22}$$

$$(x+1)^\star + (y-1)^\star = x^\star + y^\star - 1 \quad \text{whenever } \Delta_x \ne 0,\ \Delta_y = 1; \tag{23}$$

$$(x+1)^\star + (y-1)^\star = x^\star + y^\star \quad \text{in the remaining cases.} \tag{24}$$

We call a step $(x, y) \to (x + 1, y - 1)$ positive (respectively negative or neutral) if (23) (respectively (22) or (24)) holds. Therefore, as we move in (21) from left to right, every positive (negative) step decreases (increases) the total number of vertices in the union $H_x \cup H_y$.

Let us now traverse (21) from right to left. We observe that the first non-neutral step encountered is positive (if we encounter a non-neutral step at all). Furthermore, after a negative step

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

the first non-neutral step encountered is positive. Note that it may happen that the last non-neutral step encountered is negative. Therefore, the total number of positive steps is at least as large as the number of negative ones. This proves (20). □

Lemma 3. *Let $k \geq 3$ and $r \geq 2$. Let $B_1 \cup \cdots \cup B_r$ be a partition of the edge set of the clique $\mathcal{K}_k$. Write $b_i = |B_i|$, $1 \leq i \leq r$, and $\varkappa = \binom{k}{2}$. Then*

$$b_1^\star + \cdots + b_r^\star \geq (\varkappa - (r-1))^\star + 2(r-1) \geq k + r.$$

Proof. The first inequality follows from (20) and the identity $b_1 + \cdots + b_r = \varkappa$. The second inequality is simple. Indeed, for $r \geq k$ the inequality follows from $2(r-1) \geq k + r - 2$ and $(\varkappa - (r-1))^\star \geq 2$. For $r \leq k-1$ we have $\varkappa - (r-1) \geq \binom{k-1}{2} + 1$ and therefore $(\varkappa - (r-1))^\star \geq k$. □

Now we are ready to bound h_F for $F = \mathcal{K}_k$.

Lemma 4. *Let $k \geq 3$, $0 < \alpha \leq 2$, and $A > 0$. Let $n, m \to +\infty$. Assume that $m \leq An$. Let $F = \mathcal{K}_k$. Then (5) implies the bound $h_F = o\big(n^{(1/\alpha)-k}\big)$. Note that for $\alpha = 2$ condition (5) is the same as (2).*

Proof. For $F = \mathcal{K}_k$ we have $e_F = \binom{k}{2}$. We observe that (5) implies

$$\mathbb{E}\big(X^{b^\star - b/(\alpha\, e_F)} Q^b\big) < \infty \qquad \text{for each } 1 \leq b < \binom{k}{2}. \tag{25}$$

Note that $\hat{s} = \binom{s-1}{2} + 1$ is the smallest integer t such that $t^\star = s$. In particular, for any b with $b^\star = s$ we have $b \geq \hat{s}$. Therefore, given $2 \leq s \leq k$, the moment condition $\mathbb{E}\big(X^{s-\hat{s}/(\alpha\, e_F)} Q^{\hat{s}}\big) < \infty$ implies $\mathbb{E}\big(X^{s-b/(\alpha\, e_F)} Q^b\big) < \infty$ for any b satisfying $b^\star = s$. In this way, (5) yields (25).

Let us bound $h_{\mathcal{K}_k}$ from above. Given a partition $\tilde{B} = (B_1, \ldots, B_r)$ of the edge set $\mathcal{E}_0$ of $\mathcal{K}_k = ([k], \mathcal{E}_0)$, let v_j be the number of vertices incident to the edges from B_j and let $b_j = |B_j|$. For any vector $\tilde{\imath} = (i_1, \ldots, i_r)$ of distinct colours,

$$h(\tilde{B}, \tilde{\imath}) \leq \prod_{j=1}^r \frac{\mathbb{E}\big(\tilde{X}^{v_j} Q^{b_j}\big)}{n^{v_j}} \leq \prod_{j=1}^r \frac{\mathbb{E}\big(\tilde{X}^{b_j^\star} Q^{b_j}\big)}{n^{b_j^\star}} \leq \frac{1}{n^{k+r}} \prod_{j=1}^r \mathbb{E}\big(\tilde{X}^{b_j^\star} Q^{b_j}\big).$$

Here, the first inequality follows from $(\tilde{X})_t/(n)_t \leq \tilde{X}^t/n^t$, since $\tilde{X} \leq n$. The second inequality follows from the obvious inequality $b_j^\star \leq v_j$ and the fact that $\tilde{X} \leq n$. The last inequality follows from the inequality $b_1^\star + \cdots + b_r^\star \geq k + r$ of Lemma 3.

For each r-partition $\tilde{B}$ as above we bound the sum over all possible colourings $\tilde{\imath}$ (there are $(m)_r$ of them):

$$\sum_{\tilde{\imath}} h(\tilde{B}, \tilde{\imath}) \leq \frac{(m)_r}{n^{k+r}} \prod_{j=1}^r \mathbb{E}\big(\tilde{X}^{b_j^\star} Q^{b_j}\big) \leq \frac{A^r}{n^k} \prod_{j=1}^r \mathbb{E}\big(\tilde{X}^{b_j^\star} Q^{b_j}\big) = o(n^{(1/\alpha)-k}). \tag{26}$$

In the very last step, with $e_F = b_1 + \cdots + b_r = \binom{k}{2}$, we used the bounds $\mathbb{E}\big(\tilde{X}^{b_j^\star} Q^{b_j}\big) = o\big(n^{b_j/\alpha e_F}\big)$ that follow from the moment conditions $\mathbb{E}\big(X^{b_j^\star - (b_j/\alpha\, e_F)} Q^{b_j}\big) < \infty$; see (25), via (16). Finally, proceeding as in (19), we obtain the desired bound $h_F = o\big(n^{(1/\alpha)-k}\big)$ from (26). □

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

2.4. Power-law tails

Recall that, given a graph $F=(\mathcal{V}_F, \mathcal{E}_F)$, we denote by $v_F=|\mathcal{V}_F|$ the number of vertices and by $e_F=|\mathcal{E}_F|$ the number of edges. Let $\Psi_F=\Psi_F(n,p)=n^{v_F}p^{e_F}$, and define $\Phi_F=\Phi_F(n,p)=\min_{H\subset F,\, e_H\geq 1}\Psi_H$, $m_F=\max_{H\subset F,\, e_H\geq 1}(e_H/v_H)$. Here, the minimum/maximum is taken over all subgraphs $H\subset F$ with $e_H\geq 1$. Recall that F is called balanced if $m_F=e_F/v_F$. For a balanced F we have, for any $H\subset F$ with $e_H\geq 1$, $\Psi_H=\left(np^{e_H/v_H}\right)^{v_H}\geq\left(np^{e_F/v_F}\right)^{v_H}=\Psi_F^{v_H/v_F}$. Hence,

$$\Phi_F\geq\min\left\{\Psi_F^{2/v_F},\,\Psi_F\right\}. \tag{27}$$

Lemma 5. *Let $a>0$ and $0<\alpha<2$. Assume that F is balanced, connected, and $v_F\geq 2$. Assume that (3) holds. Then*

$$\mathbb{P}\{N_F>t\}=(a+o(1))t^{-\alpha}\qquad as\quad t\to+\infty. \tag{28}$$

We remark that, for $0<\alpha<2$, the tail asymptotics (28) implies that N_F belongs to the domain of attraction of an α-stable distribution. Indeed, the left tail of N_F vanishes since $\mathbb{P}\{N_F\geq 0\}=1$. Therefore, the conditions of [9, Theorem 2, §35, Chapter 7] are satisfied.

Proof. With a little abuse of notation we shall denote the conditional expectation and probability given (X, Q) by $\mathbb{E}^*$ and $\mathbb{P}^*$. Furthermore, we write $k=v_F$ and $\Delta_F^*=N_F-N_F^*$.

To prove (28), we show that the contribution of Δ_F^* to the sum $N_F=N_F^*+\Delta_F^*$ is negligible compared to N_F^* and, therefore, the tail asymptotic (28) is determined by (3). For this purpose we apply exponential large-deviation bounds for subgraph counts in Bernoulli random graphs [14, 15] (for $F=\mathcal{K}_2$, we can apply Chernoff's bounds).

Given large $t>0$ and small $\varepsilon>0$, introduce the event $\mathcal{H}=\left\{-\varepsilon N_F^*\leq\Delta_F^*\leq\varepsilon t\right\}$ and split $\mathbb{P}\{N_F>t\}$:

$$\begin{aligned}\mathbb{P}\{N_F>t\}&=\mathbb{P}\{N_F>t,\,\mathcal{H}\}+\mathbb{P}\left\{N_F>t,\,\Delta_F^*<-\varepsilon N_F^*\right\}+\mathbb{P}\left\{N_F>t,\,\Delta_F^*>\varepsilon t\right\}\\&=:P_1+P_2+P_3.\end{aligned} \tag{29}$$

We first consider P_1. Replacing Δ_F^* by its extreme values (on $\mathcal{H}$) yields the inequalities

$$\mathbb{P}\left\{(1-\varepsilon)N_F^*>t,\,\mathcal{H}\right\}\leq P_1\leq\mathbb{P}\left\{N_F^*>t(1-\varepsilon),\,\mathcal{H}\right\}.$$

We note that the right-hand side of this is at most $\mathbb{P}\left\{N_F^*>t(1-\varepsilon)\right\}$, and the left-hand side is at least $\mathbb{P}\left\{(1-\varepsilon)N_F^*>t\right\}-P_2'-P_3'$, where

$$P_2':=\mathbb{P}\left\{(1-\varepsilon)N_F^*>t,\,\Delta_F^*<-\varepsilon N_F^*\right\},\qquad P_3':=\mathbb{P}\left\{(1-\varepsilon)N_F^*>t,\,\Delta_F^*>\varepsilon t\right\}.$$

Hence, we have

$$\mathbb{P}\left\{(1-\varepsilon)N_F^*>t\right\}-P_2'-P_3'\leq P_1\leq\mathbb{P}\left\{N_F^*>t(1-\varepsilon)\right\}. \tag{30}$$

Invoking the simple inequalities $P_2\leq P_2'$ and $P_3'\leq P_3$, we obtain, from (29) and (30), that

$$\mathbb{P}\left\{(1-\varepsilon)N_F^*>t\right\}-P_2'\leq\mathbb{P}\{N_F>t\}\leq\mathbb{P}\left\{N_F^*>t(1-\varepsilon)\right\}+P_2'+P_3. \tag{31}$$

We show below that, for any $0<\varepsilon<1$,

$$P_2'=o(t^{-\alpha})\text{ and }P_3=o(t^{-\alpha})\qquad\text{as}\quad t\to+\infty. \tag{32}$$

Note that (3) and (31) together with (32) imply (28). It remains to show (32).

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

To illustrate the argument for doing so, we first examine the simplest case, where $F = \mathcal{K}_2$. We apply Chernoff's inequalities [14, (2.5), (2.6)] to Δ_F^* conditionally given (X, Q). We have

$$\begin{aligned} P_2' &= \mathbb{E}\Big(\mathbf{1}_{\{(1-\varepsilon)N_F^* > t\}}\mathbb{P}^*\big\{\Delta_F^* < -\varepsilon N_F^*\big\}\Big) \le \mathbb{E}\Big(\mathbf{1}_{\{(1-\varepsilon)N_F^* > t\}}\mathrm{e}^{-(\varepsilon^2/2)N_F^*}\Big) \\ &\le \exp\left\{-\frac{1}{2}\frac{\varepsilon^2}{1-\varepsilon}t\right\} = o(t^{-\alpha}); \\ P_3 &\le \mathbb{E}\big(\mathbb{P}^*\big\{\Delta_F^* > \varepsilon t\big\}\big) \le \mathbb{E}\,\exp\left\{-\frac{\varepsilon^2 t^2}{2\big(N_F^* + \varepsilon t/3\big)}\right\} \\ &\le \mathbb{P}\Big\{N_F^* > t^{3/2}\Big\} + \exp\left\{-\frac{\varepsilon^2 t^2}{2(t^{3/2} + \varepsilon t/3)}\right\} = o(t^{-\alpha}). \end{aligned}$$

In the last inequality we used the fact that

$$\exp\left\{-\frac{\varepsilon^2 t^2}{2\big(N_F^* + \varepsilon t/3\big)}\right\} \le \exp\left\{-\frac{\varepsilon^2 t^2}{2(t^{3/2} + \varepsilon t/3)}\right\}$$

for $N_F^* \le t^{3/2}$.

Now we assume that $v_F \ge 3$. In this case the proof of (32) is much more involved. In the proof we use often the fact [14, Lemma 3.5] that

$$\mathbb{E}^*\big(\Delta_F^*\big)^2 \asymp \frac{\big(N_F^*\big)^2}{\Phi_F(X, Q)}(1 - Q). \tag{33}$$

We also use the simple relation $N_F^* \asymp a_F \Psi_F(X, Q)$.

To prove $P_2' = o(t^{-\alpha})$, given (X, Q) with $0 < Q < 1$ (cases 0 and 1 are trivial), we apply Janson's inequality [14, Theorem 2.14] to $p_\varepsilon^* := \mathbb{P}^*\big\{\Delta_F^* < -\varepsilon N_F^*\big\}$. In what follows, we assume that the random graph $\mathrm{G}(X, Q)$ and complete graph $\mathcal{K}_X$ are both defined on the same vertex set of size X, and that $X \ge 1$. Let

$$\bar{\delta} := \mathbb{E}^*\big(N_F^2\big) - \delta, \qquad \delta := \sum_{F' \subset \mathcal{K}_X} \sum_{\substack{F'' \subset \mathcal{K}_X \\ \mathcal{E}_{F'} \cap \mathcal{E}_{F''} = \emptyset}} \mathbb{E}^*(\mathbf{1}_{F'}\mathbf{1}_{F''}).$$

Here, the sum runs over ordered pairs (F', F'') of subgraphs of $\mathcal{K}_X$ such that F' and F'' are copies of F and their edge sets $\mathcal{E}_{F'}$ and $\mathcal{E}_{F''}$ are disjoint. Furthermore, $\mathbf{1}_{F'}$ stands for the indicator of the event that F' is present in $G(X, Q)$. Janson's inequality implies

$$\mathbb{P}^*\big\{\Delta_F^* < -\eta N_F^*\big\} \le \mathrm{e}^{-\left(\eta N_F^*\right)^2/\bar{\delta}} \qquad \text{for all} \quad \eta \in (0, 1). \tag{34}$$

Next, we bound $\bar{\delta}$ from above. The (variance) identity $\mathbb{E}^*(N_F^2) - \big(N_F^*\big)^2 = \mathbb{E}^*\big(\Delta_F^*\big)^2$ implies that

$$\bar{\delta} = \mathbb{E}^*\big(\Delta_F^*\big)^2 + \big(N_F^*\big)^2 - \delta. \tag{35}$$

Furthermore, using the observation that $V_{F'} \cap V_{F''} = \emptyset$ implies $\mathcal{E}_{F'} \cap \mathcal{E}_{F''} = \emptyset$, and that the latter relation implies $\mathbb{E}^*(\mathbf{1}_{F'}\mathbf{1}_{F''}) = (\mathbb{E}^*\mathbf{1}_{F'})(\mathbb{E}^*\mathbf{1}_{F''}) = Q^{2e_F}$, we bound δ from below:

$$\delta \ge \sum_{F' \subset \mathcal{K}_X} \sum_{\substack{F'' \subset \mathcal{K}_X \\ V_{F'} \cap V_{F''} = \emptyset}} \mathbb{E}^*(\mathbf{1}_{F'}\mathbf{1}_{F''}) = a_F^2 \binom{X}{k}\binom{X-k}{k} Q^{2e_F} = \frac{(X-k)_k}{(X)_k}\big(N_F^*\big)^2.$$

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

Then, we lower bound the fraction

$$\frac{(X-k)_k}{(X)_k} \geq \left(1-\frac{k}{X-k}\right)^k \geq 1-\frac{k^2}{X-k} \qquad \text{for } X \geq 2k,$$

and obtain that $\delta \geq \big(N_F^*\big)^2\big(1-k^2(X-k)^{-1}\big)$. Invoking this bound in (35) we obtain $\bar{\delta} \leq \mathbb{E}^*\big(\Delta_F^*\big)^2 + \big(N_F^*\big)^2 k^2 (X-k)^{-1}$. Hence, the ratio in the exponent of (34) satisfies

$$\frac{\big(N_F^*\big)^2}{\bar{\delta}} \geq \frac{\big(N_F^*\big)^2}{2\max\left\{\mathbb{E}^*\big(\Delta_F^*\big)^2, \big(N_F^*\big)^2 k^2 (X-k)^{-1}\right\}} = \frac{1}{2}\min\left\{\frac{\big(N_F^*\big)^2}{\mathbb{E}^*\big(\Delta_F^*\big)^2}, \frac{X-k}{k^2}\right\}. \tag{36}$$

We will show below that there exists $c_k > 0$ (independent of t) such that $N_F^* > t$ implies

$$\frac{\big(N_F^*\big)^2}{\mathbb{E}^*\big(\Delta_F^*\big)^2} > c_k t^{2/k}. \tag{37}$$

We also note that $N_F^* > t$ implies $X > (t/a_F)^{1/k}$, using $a_F\binom{X}{k} \geq a_F\binom{X}{k}Q^{e_F} = N_F^*$). Therefore, on the event $N_F^* > t$ the right-hand side of (36) is at least

$$\frac{1}{2}\min\left\{c_k t^{2/k}, \ \frac{(t/a_F)^{1/k}-k}{k^2}\right\}, \tag{38}$$

and this quantity scales as $t^{1/k}$ as $t \to +\infty$. Finally, from (34), (36), and (38) we obtain that, on the event $N_F^* > t$, $p_\varepsilon^* \leq \mathrm{e}^{-\varepsilon^2\Theta(t^{1/k})} = o(t^{-\alpha})$ as $t \to +\infty$. We conclude that $P_2' = o(t^{-\alpha})$. It remains to show (37). We observe that the inequalities $N_F^* \leq a_F \Psi_F(X, Q)$ and $N_F^* > t$ imply $\Psi_F(X, Q) > t/a_F > 1$, where the last inequality holds for $t > a_F$. Then, (27) implies $\Phi_F(X, Q) \geq (\Psi_F(X, Q))^{2/k}$, and (33) implies

$$\frac{\big(N_F^*\big)^2}{\mathbb{E}^*\big(\Delta_F^*\big)^2} \asymp \frac{\Phi_F(X, Q)}{1-Q} \geq \Phi_F(X, Q) \geq \Psi_F^{2/k}(X, Q) \geq (t/a_F)^{2/k}.$$

To prove $P_3 = o(t^{-\alpha})$ we apply exponential inequalities for upper tails of subgraph counts in Bernoulli random graphs [15]. For the reader's convenience, we state the result from [15] that we will use. Let Δ_F be the maximum degree of F. Let

$$M_F(n,p) = \begin{cases} 1 & \text{if } p < n^{-1/m_F}, \\ \min_{H\subset F}\big(\Psi_H(n,p)\big)^{1/\alpha_H^*} & \text{if } n^{-1/m_F} \leq p \leq n^{-1/\Delta_F}, \\ n^2 p^{\Delta_F} & \text{if } p \geq n^{-1/\Delta_F}. \end{cases}$$

Here, α_H^* is the fractional independence number of a graph H [15]. We do not define the fractional independence number here as we only use the upper bound $\alpha_H^* \leq v_H - 1$ that holds for any H with $e_H > 0$ [15, (A.1)]. Let ξ_F be the number of copies of F in $G(n, p)$. By [15, Theorems 1.2 and 1.5], for any $\eta > 0$ there exists $c_{\eta,F} > 0$ such that, uniformly in p and $n \geq k$ (recall that $k = v_F$ is the number of vertices of F),

$$\mathbb{P}\big\{\xi_F \geq (1+\eta)\mathbb{E}\xi_F\big\} \leq \mathrm{e}^{-c_{\eta,F} M_F(n,p)}. \tag{39}$$

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

We will apply (39) to the number N_F of copies of F in $G(X, Q)$ conditionally given X, Q; see (43).

We write, for short, $s = \varepsilon t$ and estimate $P_3 \le \mathbb{P}\{\Delta_F^* > s\}$. Let $\eta > 0$. We split

$$\mathbb{P}\{\Delta_F^* > s\} = \mathbb{P}\{\Delta_F^* > \eta N_F^*,\ \Delta_F^* > s\} + \mathbb{P}\{\Delta_F^* \le \eta N_F^*,\ \Delta_F^* > s\} =: P_{31} + P_{32}$$

and estimate the probabilities P_{31} and P_{32} separately. The second probability,

$$P_{32} \le \mathbb{P}\{N_F^* > s/\eta\} = \eta^\alpha (a + o(1)) s^{-\alpha}, \tag{40}$$

can be made negligibly small by choosing η arbitrarily small.

Now we upper bound the remaining probability P_{31}. Introduce the events

$$\mathcal{A}_1 = \{Q \le X^{-1/m_F}\}, \quad \mathcal{A}_{21} = \{X^{-1/m_F} < Q < X^{-1/\Delta_F}\}, \quad \mathcal{A}_{22} = \{Q \ge X^{-1/\Delta_F}\},$$

and put $\mathcal{A}_2 = \mathcal{A}_{21} \cup \mathcal{A}_{22}$ (note that $\Delta_F \ge 2m_F = 2e_F/v_F$). We split

$$P_{31} = \tilde{P}_1 + \tilde{P}_2, \qquad \tilde{P}_i := \mathbb{P}\{\Delta_F^* > \eta N_F^*,\ \Delta_F^* > s,\ \mathcal{A}_i\},$$

and estimate $\tilde{P}_1$ and $\tilde{P}_2$ separately.

We first consider $\tilde{P}_1$. The inequality $Q \le X^{-1/m_F}$ implies $\Psi_F(X, Q) \le 1$. Consequently, (27) implies $\Phi_F(X, Q) \ge \Psi_F(X, Q)$. The latter inequality, together with (33), imply $\mathbb{E}^*(\Delta_F^*)^2 \le c_k \Psi_F(X, Q) \le c_k$ for some $c_k > 0$. Hence, on the event $\mathcal{A}_1$ we have $\mathbb{E}^*(\Delta_F^*)^2 \le c_k$. Finally, by Markov's inequality,

$$\tilde{P}_1 \le \mathbb{P}\{\Delta_F^* > s,\ \mathcal{A}_1\} = \mathbb{E}\big(\mathbf{1}_{\mathcal{A}_1} \mathbb{E}^* \mathbf{1}_{\{\Delta_F^* > s\}}\big) \le \mathbb{E}\big(\mathbf{1}_{\mathcal{A}_1} \mathbb{E}^* (\Delta_F^*)^2 s^{-2}\big) \le c_k s^{-2}. \tag{41}$$

Second, we consider $\tilde{P}_2$. The inequality $X^{-1/m_F} < Q$ implies $\Psi_F(X, Q) > 1$. For balanced F this yields $\Psi_H(X, Q) > 1$ for every $H \subset F$ with $e_H > 0$. Then, by using $\alpha_H^* \le v_H - 1$ we obtain

$$\min_{H \subset F:\, e_H > 0} (\Psi_H(X, Q))^{1/\alpha_H^*} \ge \min_{H \subset F:\, e_H > 0} (\Psi_H(X, Q))^{1/v_H} = (\Psi_F(X, Q))^{1/v_F}.$$

In the last step we used the fact that F is balanced once again. Hence, on the event $\mathcal{A}_{21}$ we have (recall that $v_F = k$)

$$M_F(X, Q) \ge (\Psi_F(X, Q))^{1/k}. \tag{42}$$

We observe that (42) holds on the event $\mathcal{A}_{22}$ as well. Indeed, the inequality $Q \ge X^{-1/\Delta_F}$ yields $M_F(X, Q) \ge X^2 Q^{\Delta_F} \ge X$. Now the inequality $X^{v_F} \ge \Psi_F(X, Q)$ implies (42).

From (39) and (42) we obtain the exponential bound

$$\mathbb{P}^*\{\Delta_F^* > \eta N_F^*\} \le \mathrm{e}^{-c_{\eta,F} M_F(X,Q)} \le \mathrm{e}^{-c_{\eta,F} (\Psi_F(X,Q))^{1/k}}. \tag{43}$$

Let us bound $\tilde{P}_2$ from above. We fix a (large) number $B > 0$ and introduce the events $\mathcal{B}_1 = \{\Psi_F(X_1, Q_1) > B \ln^k s\}$ and $\mathcal{B}_2 = \{\Psi_F(X_1, Q_1) \le B \ln^k s\}$. We then split $\tilde{P}_2 = \tilde{P}_{21} + \tilde{P}_{22}$, $\tilde{P}_{2i} = \mathbb{P}\{\Delta_F^* > \eta N_F^*,\ \Delta_F^* > s,\ \mathcal{A}_2,\ \mathcal{B}_i\}$, and bound $\tilde{P}_{21}$ from above, using (43):

$$\begin{aligned} \tilde{P}_{21} &\le \mathbb{P}\{\Delta_F^* > \eta N_F^*,\ \mathcal{A}_2,\ \mathcal{B}_1\} = \mathbb{E}\left(\mathbf{1}_{\mathcal{B}_1} \mathbf{1}_{\mathcal{A}_2} \mathbb{P}^*\{\Delta_F^* > \eta N_F^*\}\right) \\ &\le \mathbb{E}\Big(\mathbf{1}_{\mathcal{B}_1} \exp\left\{-c_{\eta,F} (\Psi_H(X_1, Q_1))^{1/k}\right\}\Big) \le \mathrm{e}^{-c_{\eta,F} B^{1/k} \ln s}. \end{aligned} \tag{44}$$

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press

It remains to upper bound $\tilde{P}_{22}$. The inequality $\Psi_F(X, Q) > 1$, which holds on the event $\mathcal{A}_2$, implies (see (27)) $\Phi_F(X, Q) \geq (\Psi_F(X, Q))^{2/k}$. Furthermore, (33) implies $\mathbb{E}^*\big(\Delta_F^*\big)^2 \leq c_F(\Psi_F(X, Q))^{2-(2/k)}(1-Q)$, where $c_F > 0$ depends only on F. Note that on the event $\mathcal{B}_2$ the right-hand side is upper bounded by $c_F(B \ln^k s)^{2-(2/k)}$. Hence, by Markov's inequality,

$$\mathbb{P}^*(\Delta_F^* > s) \leq s^{-2}\mathbb{E}^*\big(\Delta_F^*\big)^2 \leq c_F B^{2-(2/k)} s^{-2} \ln^{2k-2} s.$$

Finally, we obtain

$$\tilde{P}_{22} \leq \mathbb{P}\big\{\Delta_F^* > s, \mathcal{A}_2, \mathcal{B}_2\big\} = \mathbb{E}\big(\mathbf{1}_{\mathcal{A}_2}\mathbf{1}_{\mathcal{B}_2}\mathbb{P}^*\big\{\Delta_F^* > s\big\}\big) \leq c_F B^{2-(2/k)} s^{-2} \ln^{2k-2} s. \tag{45}$$

We complete the proof by showing that, for any $0 < \varepsilon < 1$, the probability P_3, which depends on ε, satisfies $P_3 = o(t^{-\alpha})$ as $t \to +\infty$. Recall that $s = \varepsilon t$. We have, for any $\eta > 0$,

$$\begin{aligned}\limsup_{t\to+\infty} t^\alpha P_3 &\leq \limsup_{t\to+\infty} t^\alpha \mathbb{P}\big\{\Delta_F^* > \varepsilon t\big\} = \varepsilon^{-\alpha} \limsup_{s\to+\infty} s^\alpha \mathbb{P}\big\{\Delta_F^* > s\big\} \\ &\leq \varepsilon^{-\alpha} \limsup_{s\to+\infty} s^\alpha \big(\tilde{P}_1 + \tilde{P}_{21} + \tilde{P}_{22} + P_{32}\big) \leq (\eta/\varepsilon)^\alpha a.\end{aligned}$$

Hence, $\limsup_{t\to+\infty} t^\alpha P_3 = 0$. The last inequality above follows from (40), (41), (44), and (45). Indeed, given $\eta > 0$, we choose $B = B(\eta)$ (in (44) and (45)) large enough that $c_{\eta,F} B^{1/k} > 2$. Then $\tilde{P}_{21} \leq s^{-2}$ and $\limsup_s s^\alpha \tilde{P}_{21} = 0$. We also mention the obvious relations $\limsup_s s^\alpha \tilde{P}_1 = 0$ and $\limsup_s s^\alpha \tilde{P}_{22} = 0$. □

Funding information

JK was supported by the Magnus Ehrnrooth Foundation and Academy of Finland grant 346311 – Finnish Centre of Excellence in Randomness and Structures.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] AMBROISE, C. AND MATIAS, C. (2012). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *J. R. Statist. Soc. B* **74**, 3–35.
[2] BENSON, A. R., GLEICH, D. AND LESKOVEC, J. (2016). Higher-order organization of complex networks. *Science* **353**, 163–166.
[3] BHATTACHARYA, B. B., CHATTERJEE, A. AND JANSON, S. (2021). Fluctuations of subgraph counts in graphon-based random graphs. Preprint, arXiv:2104.07259.
[4] BLOZNELIS, M. AND JAWORSKI, J. (2018). The asymptotic normality of the global clustering coefficient in sparse random intersection graphs. In *Algorithms and Models for the Web Graph – 15th International Workshop, WAW 2018* (Lect. Notes Comp. Sci. 10836), eds A. Bonato. P. Prałat and A. Raigorodskii. Springer, New York, pp. 16–29.
[5] BLOZNELIS, M. AND KURAUSKAS, V. (2016). Clustering coefficient of random intersection graphs with infinite degree variance. *Internet Math.* doi: 10.24166/im.02.2017.
[6] BLOZNELIS, M. AND LESKELÄ, L. (2023). Clustering and percolation on superpositions of Bernoulli random graphs. *Random Structures Algorithms* **63**, 283–342.
[7] BLOZNELIS, M., KARJALAINEN, J. AND LESKELÄ, L. (2022). Assortativity and bidegree distributions on Bernoulli random graph superpositions. *Prob. Eng. Inf. Sci.* **36**, 1188–1213.
[8] EIKMEIER, N., RAMANI, A. S. AND GLEICH, D. (2018). The HyperKron graph model for higher-order features. In *Proc. 2018 IEEE International Conference on Data Mining*. IEEE, Piscataway, NJ, pp. 941–946.

[9] GNEDENKO, B. V. AND KOLMOGOROV, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Cambridge.
[10] GODEHARDT, E. AND JAWORSKI, J. (2001). Two models of random intersection graphs and their applications. *Electron. Notes Discrete Math.* **10**, 129–132.
[11] GRÖHN, T., KARJALAINEN, J. AND LESKELÄ, L. (2019). Clique and cycle frequencies in a sparse random graph model with overlapping communities. Preprint, arXiv:1911.12827.
[12] HLADKÝ, J., PELEKIS, CH. AND ŠILEIKIS, M. (2021). A limit theorem for small cliques in inhomogeneous random graphs. *J. Graph Theory* **97**, 578–599.
[13] HONEY, C. J., KÖTTER, R., BREAKSPEAR, M. AND SPORNS, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Nat. Acad. Sci. USA* **104**, 10240–10245.
[14] JANSON, S., ŁUCZAK, T. AND RUCIŃSKI, A. (2000). *Random Graphs*. John Wiley, New York.
[15] JANSON, S., OLESZKIEWICZ, K. AND RUCIŃSKI, A. (2004). Upper tails for subgraph counts in random graphs. *Israel. J. Math.* **142**, 61–92.
[16] KARJALAINEN, J., VAN LEEUWAARDEN, J. S. H. AND LESKELÄ, L. (2018). Parameter estimators of random intersection graphs with thinned communities. In *Algorithms and Models for the Web Graph – 15th International Workshop, WAW 2018* (Lect. Notes Comp. Sci. **10836**), eds A. Bonato, P. Prałat and A. Raigorodskii. Springer, New York, pp. 44–58.
[17] KAUR, G. AND RÖLLIN, A. (2021). Higher-order fluctuations in dense random graph models. *Electron. J. Prob.* **26**, 1–36.
[18] KURAUSKAS, V. (2022). On local weak limit and subgraph counts for sparse random graphs. *J. Appl. Prob.* **59**, 755–776.
[19] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHLOVSKII, D. AND ALON, U. (2002): Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827.
[20] OSPINA-FORERO, L., DEANE, C. M. AND REINERT G. (2019). Assessment of model fit via network comparison methods based on subgraph counts. *J. Complex Networks* **7**, 226–253.
[21] PRIVAULT, N. AND SERAFIN, G. (2020). Normal approximation for sums of weighted U-statistics: Application to Kolmogorov bounds in random subgraph counting. *Bernoulli* **26**, 587–615.
[22] RÖLLIN, A. (2022). Kolmogorov bounds for the normal approximation of the number of triangles in the Erdös–Rényi random graph. *Prob. Eng. Inf. Sci.* **36**, 747–773.
[23] RUCIŃSKI, A. (1988). When are small subgraphs of a random graph normally distributed? *Prob. Theory Relat. Fields* **78**, 1–10.
[24] SHEN-ORR, S. S., MILO, R., MANGAN, S. AND ALON, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics* **31**, 64–68.
[25] UGANDER, J., BACKSTROM, L. AND KLEINBERG, J. (2013). Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *Proc. 22nd Int. Conf. World Wide Web*, 1307–1318.
[26] VAN DER HOFSTAD, R., KOMJÁTHY, J. AND VADON, V. (2021). Random intersection graphs with communities. *Adv. Appl. Prob.* **53**, 1061–1089.
[27] VAN LEEUWAARDEN, J. S. AND STEGEHUIS, C. (2021). Robust subgraph counting with distribution-free random graph analysis. *Phys. Rev. E* **104**, 044313.
[28] YANG, J. AND LESKOVEC, J. (2012). Community affiliation graph model for overlapping network community detection. In *Proc. 2012 IEEE 12th Int. Conf. Data Mining*. IEEE, Piscataway, NJ, pp. 1170–1175.
[29] YANG, J. AND LESKOVEC, J. (2014). Structure and overlaps of ground-truth communities in networks. *ACM Trans. Intell. Syst. Technol.* **5**, 1–35.
[30] ZHANG, Z. S. (2022). Berry–Esseen bounds for generalized *U*-statistics. *Electron. J. Prob.* **27**, 1–36.

https://doi.org/10.1017/jpr.2023.48 Published online by Cambridge University Press