Discrete Mathematics 72 (1988) 299–310
North-Holland

SUBGRAPH COUNTS IN RANDOM GRAPHS USING INCOMPLETE *U*-STATISTICS METHODS

Krzysztof NOWICKI
Department of Mathematical Statistics, Lund University, S-22100 Lund, Sweden
and
John C. WIERMAN
Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218, U.S.A.

Received 26 August 1986
Revised 20 September 1987

The random graph $K_{n,p}$ is constructed on n labelled vertices by inserting each of the $\binom{n}{2}$ possible edges independently with probability p, $0>p<1$. For a fixed graph G, the threshold function for existence of a subgraph of $K_{n,p}$ isomorphic to G has been determined by Erdös and Rényi [8] and Bollobás [3]. Bollobás [3] and Karoński [14] have established asymptotic Poisson and normal convergence for the number of subgraphs of $K_{n,p}$ isomorphic to G for sequences of $p(n)\to 0$ which are slightly greater than the threshold function. We use techniques from asymptotic theory in statistics, designed to study sums of dependent random variables known as U-statistics. We note that a subgraph count has the form of an incomplete U-statistic, and prove asymptotic normality of subgraph counts for a wide range of values of p, including any constant p and sequences of $p(n)$ tending to 0 or 1 sufficiently slowly.

1. Introduction

Erdös and Rényi [8] introduced the random graph model $K_{n,p}$, in which a graph is constructed on n labelled vertices with each of the $\binom{n}{2}$ possible edges present with probability p, $0<p<1$, independently. Without loss of generality, we will label the vertices $\{1, 2, \ldots, n\}$. We may view the random graph $K_{n,p}$ as determined by a set $\{X(i,j)\}$, $1\leqslant i<j\leqslant n$, of independent Bernoulli random variables with $p=\{X(i,j)=1)=1-P(X(i,j)=0)$ for all $1\leqslant j<j\leqslant n$, where $X(i,j)=1$ indicates that an edge is present between i and j, and $X(i,j)=0$ indicates the absence of an edge between i and j. For convenience, we let $q=1-p$.

One often views a random graph as a structure which evolves as edges are added successively, or as p increases from 0 to 1. Many important properties of graphs appear suddenly in this evolution as the probability p crosses a threshold, on opposite sides of which $K_{n,p}$ has the property with probability 0 or 1 asymptotically as $n\to\infty$. For most properties of interest, the threshold is a function of n which tends to 0 as $n\to\infty$. For a given function $p(n)$, a property is said to hold for *almost all graphs* if it holds for a set of random graphs $K_{n,p}$ with

probability tending to one as $n \to \infty$. A more complete discussion of the extensive literature on random graphs is available in the recent introduction by Palmer [17] and research monograph by Bollobás [4].

For a fixed graph G, we consider the *subgraph count*, $S_n(G)$, a random variable defined as the number of subgraphs of $K_{n,p}$ which are isomorphic to G. Introduce the indicator function notations

$$I(A \subseteq K_{n,p}) = \begin{cases} 1 & \text{if } A \text{ is a subgraph of } K_{n,p} \\ 0 & \text{otherwise} \end{cases}$$

$$= \prod_{e \in E(A)} X(e)$$

where $E(A)$ denotes the edge set of graph A.

$$I(A \sim G) = \begin{cases} 1 & \text{if } A \text{ is isomorphic to } G \\ 0 & \text{otherwise} \end{cases}$$

We may express the subgraph count as

$$S_n(G) = \sum_{A \subseteq K_n} I(A \sim G) I(A \subseteq K_{n,p})$$

$$= \sum_{A \subseteq K_n} I(A \sim G) \prod_{e \in E(A)} X(e)$$

where K_n denotes the complete graph on the vertex set $\{1, 2, \ldots, n\}$. Note that we identify a graph with its set of edges, so the approach applies directly only to graphs G with no isolated vertices. However, the results can be easily extended to graphs which have isolated vertices.

If G is a graph with k edges, we need only consider subgraphs A with k edges, so

$$S_n(G) = \sum_{\substack{A \subseteq K_n \\ |E(A)| = k}} I(A \sim G) \prod_{e \in E(A)} X(e)$$

which has the form

$$\sum_{\substack{e_1, e_2, \ldots, e_k \text{ distinct} \\ e_i \in E(K_n)}} w(e_1, e_2, \ldots, e_k) h(X(e_1), X(e_2), \ldots, X(e_k)) \qquad (*)$$

in which $w(e_1, \ldots, e_k)$ is a nonrandom indicator function.

The primary goal of this paper is to establish that the subgraph count random variable $S_n(G)$ has an asymptotic normal distribution for all graphs G with no isolated vertices, and for a broad range of probability functions $p(n)$. A secondary purpose is to introduce statistical methods for the treatment of random variables of the form (*), which are called weighted U-statistics, which may be relevant to other random graph problems. Previous results on the threshold for existence of a subgraph of $K_{n,p}$ isomorphic to G, and on limiting Poisson and normal distributions for subgraph counts for sequences $p(n)$ converging to 0 but slightly above the existence threshold, are discussed in Section 2. The principal

result, stated in Section 3, establishes asymptotic normality of subgraph counts when p is constant, and when the sequence $p(n)$ converges to 0 or 1 sufficiently slowly. The result fills a substantial gap in the literature on asymptotic subgraph count distributions. Examples illustrating the application of our result to stars, trees, cycles, and complete graphs are also presented in Section 3. Section 4 discusses statistical tools for treating sums of dependent random variables, from the theory of U-statistics, and adapts them for treatment of subgraph counts. For computation of variances of the relevant statistics, counts of intersections of isomorphic subgraphs are considered in Section 5.

2. Previous subgraph count results

The determination of the threshold for the existence of a given graph G as a subgraph of $K_{n,p}$ was the focus of Theorem 1 of Erdös and Rényi [8]. For this problem, Erdös and Rényi introduced the concept of a balanced graph. Define the *degree* of a graph G by

$$d(G) = |E(G)|/|V(G)|.$$

A graph G is *balanced* if $d(G) \geq d(H)$ for every subgraph H of G. Let $A \supset B$ denote that A contains a subgraph isomorphic to B. Erdös and Rényi proved that if G is balanced

$$\lim_{n\to\infty} P[K_{n,p} \supset G] = \begin{cases} 0 & \text{if } p(n)n^{1/d(G)} \to 0 \text{ as } n \to \infty \\ 1 & \text{if } p(n)n^{1/d(G)} \to \infty \text{ as } n \to \infty \end{cases}$$

which identifies the threshold function for existence of a balanced graph G as $n \to \infty$.

Bollobás [3] generalized this result to arbitrary graphs G. Define $m(G)$ as the maximal degree of any subgraph of G. Note that $m(G) = d(G)$ if and only if G is balanced. Then for any graph G,

$$\lim_{n\to\infty} P[K_{n,p} \supset G] = \begin{cases} 0 & \text{if } p(n)n^{1/m(G)} \to 0 \text{ as } n \to \infty \\ 1 & \text{if } p(n)n^{1/m(G)} \to \infty \text{ as } n \to \infty \end{cases}$$

i.e. the existence threshold is $n^{-1/m(G)}$. The Bollobás proof uses a rather intricate method called grading, but a short elementary proof has been supplied by Rucinski and Vince [19] using the second moment method.

A graph G is *strictly balanced* if $d(G) > d(H)$ for every proper subgraph H of G. This concept plays a crucial role in obtaining asymptotic distributions for subgraph counts near the threshold function for existence. Independently, Bollobás [3] and Karoński and Rucinski [15] proved the following: Let G be a strictly balanced graph with k edges, l vertices, $m = m(G) = k/l$, and an automorphism group of order a. Let $p(n)n^{1/m} \to c$ as $n \to \infty$, for some $0 < c < \infty$. Then $S_n(G)$ has an asymptotic Poisson distribution with mean c^k/a.

Research of Rucinski and Vince [19] shows that the factorial moment

convergence method for establishing Poisson convergence does not apply if G is not strictly balanced. While their result does not prove that Poisson convergence is impossible, it strongly suggests that Poisson convergence of a subgraph count holds if and only if the graph is strictly balanced.

For a small range of sequences $p(n)$ above the existence threshold, Karoński and Rucinski [15] established asymptotic normality for subgraph counts. If G is a strictly balanced graph, and $p(n)n^{1/m(G)} \to \infty$, but for any $\delta > 0$, $p(n)n^{1/m(G)-\delta} \to 0$, then $S_n(G)$ is asymptotically normally distributed.

Novicki [16] treats multiple subgraph count statistics, obtaining multivariate normal limiting distributions, and also treats induced subgraph count statistics, for constant values of p.

Janson [13] applies the method of semi-invariants to derive limiting normal distributions for induced subgraph counts to deal with graphs for which the usual normalization is not valid for certain constant values of p.

3. Statement of results

Theorem 3.1. *Let G be a graph with no isolated vertices. Suppose that G has k edges and l vertices, and that*

$$np^{k-1} \to +\infty$$
$$n^2(1-p) \to +\infty.$$

Then

$$\frac{S_n(G) - E[S_n(G)]}{\binom{n-2}{l-2}\frac{2k}{a}(l-2)!\sqrt{\binom{n}{2}p(1-p)}}$$

has an asymptotic Normal (0, 1) *distribution.*

Remark 1. Let G' be a graph with $l+m$ vertices, m of which are isolated vertices, and let G be the graph obtained by deleting the isolated vertices from G'. For an isomorphic copy of G on a set of l vertices, there are $\binom{n-l}{m}$ sets of vertices that may be added to obtain a copy of G'. Thus, $S_n(G') = \binom{n-l}{m}S_n(G)$, and the asymptotic distribution for $S_n(G')$ is easily obtained from Theorem 3.1.

Remark 2. The subgraph count $S_n(G)$ is a sum of dependent random variables, for which the exact calculation of the variance may be long and tedious. Our method approximates $S_n(G)$ by a sum of independent random variables, called the projection of $S_n(G)$, for which the variance calculation is elementary, providing the denominator of the normalized random variable in the conclusion of Theorem 3.1.

Remark 3. The upper end of the range of normal convergence does not depend on the graph G, as the lower end does, and is the best possible bound: If $n^2(1-p)\to c$ for some $0<c<\infty$, then $\binom{n}{2}(1-p)\to \frac{1}{2}c$, which implies that the number of edges of K_n which are absent in $K_{n,p}$ has a limiting Poisson distribution with mean $\frac{1}{2}c$. As a consequence, the number of subgraphs of K_n isomorphic to G which are not subgraphs of $K_{n,p}$, properly standardized, has an asymptotic Poisson distribution. If $n^2(1-p)\to 0$, then almost all graphs are complete, so the subgraph count is a deterministic function of n with probability tending to one.

Remark 4. For a star on d vertices, Theorem 3.1 provides normal convergence when $p = w_n n^{-1/(d-1)}$ where $w_n \to +\infty$. Since the star is strictly balanced with average degree $(d-1)/d$, the threshold function for existence of d-stars is $n^{-d/(d-1)}$, and normal convergence results of Karonski and Rucinski [15] apply when $p_n = w_n n^{-d/(d-1)}$ where $w_n \to \infty$ but $w_n = o(n^{\delta})$ for any $\delta > 0$. The range between $n^{-1/(d-1)}$ and $n^{-d/(d-1)}$ is not covered by either result.

For a cycle of length d, Theorem 3.1 applies when $p_n = w_n n^{-1/(d-1)}$ where $w_n \to \infty$, and previous results apply for $p_n = w_n n^{-1}$ where $w_n \to \infty$ but $w_n = o(n^{\delta})$ for any $\delta > 0$.

For the complete graph on d vertices, Theorem 3.1 gives

$$p_n = w_n n^{-1/[\binom{d}{2}-1]} = w_n n^{-2/[(d-2)(d+1)]},$$

where $w_n \to \infty$, as the lower bound on the range of normal convergence. However, the previous results give normal convergence for sequences near the threshold $n^{-2/(d-1)}$, again leaving a range of values where the asymptotic behavior is unknown. Rucinski [18] recently proved that normal convergence holds in this range.

4. *U*-statistic methodology

Consider a sequence $X_1, X_2, X_3, \ldots$ of independent identically distributed random variables. If h is a symmetric function of k variables, the *U-statistic* with kernel h based on n observations is defined by

$$U_n = \binom{n}{k}^{-1} \sum_{\{i_1,\ldots,i_k\}\in C(n,k)} h(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$$

where $C(n, k)$ denotes the set of subsets of size k from $\{1, 2, \ldots, n\}$.

U-statistics were introduced by Hoeffding [11] as a generalization of the sample mean, and provide a class of unbiased estimators of distributional parameters in statistical estimation theory. Many common statistics, such as the sample mean, sample variance, and Wilcoxon test statistics, are *U*-statistics. The strong law of large numbers for *U*-statistics was established by Hoeffding [12] and Berk [1] using martingale methods. The central limit theorem was established in the original

paper of Hoeffding, and the rate of convergence to normality has been investigated by Grams and Serfling [9], Bickel [2], Chan and Wierman [7], Callaert and Jansen [6], and Helmers and Van Zwet [10]. U-statistics are useful as approximations to other classes of statistics, such as linear combinations of order statistics.

If each term $h(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ is weighted by a factor $w(i_1, i_2, \ldots, i_k)$, we have the more general form of a *weighted U-statistic*

$$W_n = \sum_C w(i_1, i_2, \ldots, i_k) h(X_{i_1}, X_{i_2}, \ldots, X_{i_k}).$$

If the weights $w(i_1, i_2, \ldots, i_k)$ take only 0 or 1 as values, the statistic W_n represents an 'incomplete' or 'reduced' U-statistic sum. Incomplete U-statistics are designed to be computationally simpler than the full sum, based on the reasoning that it should be possible to use fewer terms without much loss of information. Incomplete U-statistics have been investigated by Brown and Kildea [5], who proved asymptotic normality under certain balance conditions on the weights. Shapiro and Hubert [20] investigated asymptotic normality for weighted U-statistics.

The asymptotic behavior of a weighted U-statistic may often be determined using Hajek's projection method, approximating the weighted U-statistic by a sum of independent random variables. Define W_n^*, the projection of W_n, by

$$W_n^* = \sum_{i=1}^{n} E[W_n \mid X_i] - (n-1)E[W_n].$$

The original weighted U-statistic may be investigated by writing

$$W_n - E[W_n] = [W_n^* - E[W_n]] + [W_n - W_n^*],$$

and considering each term on the right side separately.

The projection W_n^* has the same mean value as W_n, so

$$W_n^* - E[W_n] = \sum_{i=1}^{n} [E[W_n \mid X_i] - E[W_n]]$$

is a sum of independent, identically distributed, mean zero random variables. Under appropriate conditions, its asymptotic distribution may be derived by a standard central limit theorem.

We will rely on a central limit theorem formulated for a double array of random variables $\{X_{nj}\}$, in which for each $n \geq 1$, there are k_n random variables $\{X_{nj}, 1 \leq j \leq k_n\}$, where it is assumed that $k_n \to \infty$ as $n \to \infty$. Denote the distribution function of X_{nj} by F_{nj}, and let

$$\mu_{nj} = E[X_{nj}]$$

$$\mu_n = E\left[\sum_{i=1}^{k_n} X_{nj}\right] = \sum_{j=1}^{k_n} \mu_{nj}$$

and

$$\sigma_n^2 = \mathrm{Var}\Big[\sum_{j=1}^{k_n} X_{nj}\Big].$$

Theorem 4.1. *Let* $\{X_{nj}: 1 \leq j \leq k_n\}$ *be a double array in which the random variables in each row are independent. If the Lindeberg condition*

$$\frac{1}{\sigma_n^2}\sum_{j=1}^{k_n}\int_{|t-\mu_{nj}|>\varepsilon\sigma_n}(t-\mu_{nj})^2\,dF_{nj}(t)\to 0 \text{ as } n\to\infty$$

is satisfied for each $\varepsilon > 0$, *then*

$$\frac{\sum_{i=1}^{k_n}(X_{nj}-\mu_n)}{\sigma_n}$$

has an asymptotic standard normal distribution.

Note that the independence is assumed only within rows, but random variables in different rows may be arbtrarily strongly dependent.

To complete the analysis, one wishes to show that the error in the approximation, $W_n - W_n^*$, is negligible, so the asymptotic behavior of W_n is the same as that of W_n^*. This is often accomplished by the use of moment inequalities, such as Chebyshev's Inequality. If $\mathrm{Var}(W_n - W_n^*) = o(\mathrm{Var}(W_n^*))$, then by Chebyshev's inequality

$$P\Big(\frac{|W_n - W_n^*|}{\mathrm{Var}(W_n^*)} \geqslant \varepsilon\Big) \leqslant \frac{\mathrm{Var}(W_n - W_n^*)}{\varepsilon^2\,\mathrm{Var}(W_n^*)} \to 0$$

for every $\varepsilon > 0$, so $(W_n - W_n^*)/\sqrt{\mathrm{Var}(W_n^*)}$ converges to 0 in probability, and the error in the approximation by the projection is negligible. Then the asymptotic normal distribution is obtained by writing

$$\frac{W_n - E[W_n]}{\sqrt{\mathrm{Var}(W_n^*)}} = \frac{W_n^* - E[W_n]}{\sqrt{\mathrm{Var}(W_n^*)}} + \frac{W_n - W_n^*}{\sqrt{\mathrm{Var}(W_n^*)}} \xrightarrow{\mathcal{D}} N(0,1),$$

by applying then central limit theorem to the projection term.

To aid in computing the variance of $W_n - W_n^*$, we now show that it may also be written in the form of a weighted U-statistic. Without loss of generality, we assume that $E[h(X_1, \ldots, X_k)] = 0$. The projection W_n^* is given by

$$\sum_{\{i_1,\ldots,i_k\}\in C(n,k)} w(i_1, \ldots, i_k)[g(X_{i_1}) + \cdots + g(X_{i_k})]$$

where

$$g(x) = E[h(X_1, \ldots, X_{k-1}, X_k)\,|X_1 = x].$$

Thus, we have the weighted U-statistic representation.

$$W_n - W_n^* = \sum_{C(n,k)} w(i_1, i_2, \ldots, i_k)\psi(X_{i_1}, X_{i_2}, \ldots, X_{i_k}),$$

where

$$\psi(X_1, X_2, \ldots, X_k) = h(X_1, X_2, \ldots, X_k) - \sum_{i=1}^{k} g(X_i).$$

The variance of $W_n - W_n^*$ is a weighted sum of expected values of products of the form

$$\psi(X_{i_1}, \ldots, X_{i_k})\psi(X_{j_1}, \ldots, X_{j_k}).$$

Note that if $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_k\}$ are disjoint, the expected value of the product is zero, by independence of the sets of random variables. In addition, if there is only one index in common, the expected value is still zero. To see this, compute

$$\begin{aligned} &E[\psi(X_{i_1}, \ldots, X_{i_k})\psi(X_{i_1}, X_{j_2}, \ldots, X_{j_k})] \\ &\quad = E[E[\psi(X_{i_1}, \ldots, X_{i_k})\psi(X_{i_1}, X_{j_2}, \ldots, X_{j_k}) \mid X_{i_1}, X_{j_2}, \ldots, X_{j_k}]] \\ &\quad = E[\psi(X_{i_1}, X_{j_2}, \ldots, X_{j_k})E[\psi(X_{i_1}, X_{i_2}, \ldots, X_{i_k}) \mid X_{i_1}]] \end{aligned}$$

and use

$$\begin{aligned} &E[\psi(X_{i_1}, X_{i_2}, \ldots, X_{i_k}) \mid X_{i_1}] \\ &\quad = E[h(X_{i_1}, \ldots, X_{i_k}) - g(X_{i_1}) - \cdots - g(X_{i_k}) \mid X_{i_1}] \\ &\quad = E[h(X_{i_1}, \ldots, X_{i_k}) \mid X_{i_1}] - g(X_{i_1}) \\ &\quad = 0. \end{aligned}$$

Thus, only terms with two or more indices in common make a contribution to the variance of $W_n - W_n^*$.

By the Cauchy–Schwartz inequality, the contribution of any non-zero term is no greater than $E[\psi(X_1, \ldots, X_k)^2]$, which is bounded by $(k+1)^2 E[h(X_1, \ldots, X_k)]^2$.

5. Application to subgraph counts

Let G be a graph with k edges and l vertices, none of which are isolated, and consider the subgraph count

$$S_n \equiv S_n(G) = \sum_{\substack{A \subseteq K_n \\ |E(A)|=k}} I(A \sim G) \prod_{e \in E(A)} X(e).$$

By subtracting the mean, we obtain

$$S_n(G) - E[S_n(G)] = \sum_{\substack{A \subseteq K_n \\ |E(A)|=k}} I(A \sim G)\left\{ \prod_{e \in E(A)} X(e) - p^k \right\}$$

which is a weighted U-statistic with kernel

$$h(X_1, \ldots, X_k) = \prod_{i=1}^{k} X_i - p^k,$$

which has mean zero. The corresponding conditional expectation is

$$g(X_1) = E\left[\prod_{i=1}^{k} X_i \mid X_1\right] - p^k$$
$$= p^{k-1}X_1 - p^k,$$

so the projection S_n^* has the form

$$S_n^* = \sum_{e \in K_n} a_n(p^{k-1}X_e - p^k)$$

where a_n denotes the number of subgraphs of K_n which contain a fixed edge and are isomorphic to G.

To compute a_n, note that there are $\binom{n-2}{l-2}$ different sets of l vertices containing the endpoints of the fixed edge. Thus, $a_n = \binom{n-2}{l-2} b_l$ where b_l denotes the number of subgraphs isomorphic to G on a set of l vertices [which without loss of generality we take to be $\{1, 2, \ldots, l\}$] with an edge between a fixed pair of vertices.

Let $C(n, G)$ denote the set of labelled subgraphs of K_n which are isomorphic to G. To find b_l, we count the number of edges in $C(l, G)$ by two different procedures. First, there are $\binom{l}{2}$ pairs of vertices, each having an edge in b_l subgraphs, for a total of $\binom{l}{2} b_l$ edges. Second, since there are $l!$ orderings of the vertices, there are $l!/a$ different subgraphs isomorphic to G, where a denotes the order of the automorphism group of G. Since each of these subgraphs has k edges, there is a total of $k(l!)/a$. Therefore,

$$b_l = \frac{k(l!)}{a\binom{l}{2}} = \frac{2k}{a}(l-2)!$$

so

$$S_n^* = \sum_{e \in K_n} \binom{n-2}{l-2} \frac{2k}{a}(l-2)!\,(p^{k-1}X_e - p^k).$$

Since the summands in S_n^* are independent and identically distributed,

$$\begin{aligned} \operatorname{Var}(S_n^*) &= \binom{n}{2} \operatorname{Var}\left(\binom{n-2}{l-2} \frac{2k}{a}(l-2)!\, p^{k-1}X_e\right) \\ &= \binom{n}{2}\binom{n-2}{l-2}^2 \frac{4k^2}{a^2}[(l-2)!]^2 p^{2k-2}p(1-p) \qquad (5.1) \\ &= \mathcal{O}(n^{2l-2}p^{2k-1}(1-p)) \quad \text{as } n \to \infty. \end{aligned}$$

Each of the summands in S^* is bounded by

$$B_n = \binom{n-2}{l-2} \frac{2k}{a}(l-2)! \max\{p^{k-1} - p^k, p^k\}.$$

If $B_n = o\sqrt{\operatorname{Var}(S_n^*)}$, then the Lindeberg condition in Theorem 4.1 is satisfied. If

p is a constant independent of n, then $B_n = \mathcal{O}(n^{l-2})$ and $\mathrm{Var}(S_n^*) = \mathcal{O}(n^{2l-2})$, so the Lindeberg condition is satisfied. If $p(n) \to 0$, $B_n = o\sqrt{\mathrm{Var}(S_n^*)}$ if and only if

$$\frac{n^{l-2}p^{k-1}\sqrt{1-p}}{n^{l-1}p^{k-1/2}\sqrt{1-p}} = \frac{1}{np^{\frac{1}{2}}} \to 0$$

i.e.

$$pn^2 \to \infty.$$

If $p(n) \to 1$, $B_n = o\sqrt{\mathrm{Var}(S_n^*)}$ if and only if

$$\frac{n^{l-2}p^k}{n^{l-1}p^{k-\frac{1}{2}}\sqrt{1-p}} \approx \frac{1}{n\sqrt{1-p}} \to 0$$

i.e. $(1-p)n^2 \to \infty$. When both of these convergence conditions hold, by the Lindeberg–Feller central limit theorem [Theorem 4.1], we have an asymptotic standard normal distribution for

$$\frac{S_n^* - E[S_n]}{\binom{n-2}{l-2}\frac{2k}{a}(l-2)!\sqrt{\binom{n}{2}p(1-p)}}.$$

To complete the proof of Theorem 3.1, we next show that $\mathrm{Var}(S_n - S_n^*) = o(\mathrm{Var}(S_n^*))$. This, by Chebyshev's inequality, implies that the error in the approximation of S_n by S_n^* is negligible. By the discussion in Section 4, $S_n - S_n^*$ may be represented as

$$\sum_{\{i, \ldots, i_k\} \subseteq \{1, \ldots, n\}} I(A \sim G)\varphi(X_{i_1}, \ldots, X_{i_k})$$

where

$$\psi(X_1, \ldots, X_k) = \left\{\prod_{i=1}^{k} X_i - p^k\right\} - \left\{\sum_{i=1}^{k} (p^{k-1}X_i - p^k)\right\}.$$

The terms in $S_n - S_n^*$ are uncorrelated unless there are two or more indices in common, which is equivalent to the corresponding subgraphs having two or more common edges.

For $d = 2, \ldots, k$, define

$$f_d = \left|\{(A, B) : |E(A) \cap E(B)| = d,\ A, B \in C(n, G)\}\right|.$$

i.e. f_d is the number of pairs of subgraphs isomorphic to G with exactly d common edges.

To compute f_d, decompose the set according to the number of common vertices, defining

$$f_d(i) = \left|\{(A, B) : |E(A) \cap E(B)| = d,\ |V(A) \cap V(B)| = i;\ A, B \in C(n, G)\}\right|$$

for $i = 3, 4, \ldots, l$ (since if A and B share two or more common edges, they have at least 3 common vertices). Then,

$$f_d = \sum_{i=3}^{l} f_d(i).$$

For each $i = 3, 4, \ldots, l$, we choose the i common vertices and $l - i$ additional vertices in each of A and B, so

$$f_d(i) = \binom{n}{i, l-1, l-i} e_d(i),$$

where $e_d(i)$ denotes the number of pairs (A, B) which can be obtained on two fixed sets of vertices $V_1 = V(A)$ and $V_2 = V(B)$ such that $|V_1 \cap V_2| = i$ and $|E(A) \cap E(B)| = d$. Since $e_d(i)$, $i = 3, 4, \ldots, k$, is a sequence of constants independent of n, we find that

$$\begin{aligned} f_d &= \sum_{i=3}^{k} \binom{n}{i, l-1, l-i} e_d(i) \\ &= \sum_{i=3}^{k} \binom{n}{2l-i} \binom{2l-1}{i} \binom{2l-2i}{l-1} e_d(i) \\ &= \mathcal{O}(n^{2l-3}). \end{aligned}$$

Therefore

$$\sum_{d=2}^{k} f_d = \mathcal{O}(n^{2l-3}).$$

Since

$$\begin{aligned} \operatorname{Var}(S_n - S_n^*) &\leq \left(\sum_{d=2}^{k} f_d \right) E[\psi^2(X_1, \ldots, X_k)] \\ &\leq \mathcal{O}(n^{2l-3}(k+1)^2 E[h^2(X_1, \ldots, X_k)]) \\ &= \mathcal{O}(n^{2l-3} p^k (1-p^k)), \end{aligned}$$

we have

$$\begin{aligned} \frac{\operatorname{Var}(S_n - S_n^*)}{\operatorname{Var}(S_n^*)} &= \mathcal{O}\left(\frac{n^{2l-3} p^k (1-p)^k}{n^{2l-2} p^{k-1} (1-p)} \right) \\ &= \mathcal{O}\left(\frac{1-p^k}{1-p} \frac{1}{np^{k-1}} \right) \to 0 \end{aligned}$$

if $np^{k-1} \to \infty$.

Thus, under the hypotheses of Theorem 3.1, the projection is asymptotically normal and the error in the approximation is negligible, so the conclusion follows by the general theory discussed in Section 4.

Acknowledgements

Dr. Wierman's research is supported in part by the U.S. National Science foundation under grant DMS-8303238. The authors became acquainted on a research exchange visit by Dr. Wierman to Poland sponsored by the National Academy of Science of the USA and the Polish Academy of Science, and learned

of their independent results presented here. Dr. Wierman thanks Michal Karoński for introducing him to the subgraph count problem.

References

[1] R.H. Berk, Limiting behavior of posterior distributions when the model is correct, Ann. Math. Statist. 37 (1966) 51–58.
[2] P.J. Bickel, Edgeworth expansions in nonparametric statistics, Ann. Statist. 2 (1974) 1–20.
[3] B. Bollobás, Threshold functions for small subgraphs, Math. Proc. Camb. Phil. Soc. 90 (1981) 197–206.
[4] B. Bollobás, Random graphs (Academic Press, 1985).
[5] B.M. Brown and D.G. Kildea, Reduced U-statistics and the Hodges–Lehmann estimator. Ann. Statist. 6 (1978) 828–835.
[6] H. Callaert and P. Janssen, The Berry–Essen theorem for U-statistics. Ann. Statist. 6 (1978) 417–421.
[7] Y.K. Chan and J.C. Wierman, On the Berry–Essen theorem for U-statistics, Ann. Prob. 5 (1977) 136–139.
[8] P. Erdös and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Sci. 5 (1960) 17–61.
[9] W.F. Grams and R.J. Serfling, Convergence rates for U-statistics and related statistics, Ann. Statist. 1 (1973) 153–160.
[10] R. Helmers and W.R. VanZwet, The Berry–Essen bound for U-statistics, Statistical Decision Theorem and Related Topics III, Vol. 1, S.S. Gupta and J.O. Berger (eds.) (1982) 497–512.
[11] W. Hoeffding, A class of statistics with asymptotically normal distribution, Ann. Math. Stat. 19 (1948) 293–325.
[12] W. Hoeffding, The strong law of large number for U-statistics, Univ. of North Carolina Institute of Statistics Mimeo Series, No. 302 (1961).
[13] S. Janson, Normal convergence by higher semi-invariants with applications to sums of dependent random variables and random graphs, Uppsala University Department of Mathematics Technical Report (1985) 12.
[14] M. Karoński, Balanced subgraphs of large random graphs (Adam Mickiewicz University Press, 1984).
[15] M. Karoński and A. Rucinski, On the number of strictly balanced subgraphs of a random graph, Graph theory. Lagow 1981, Lecture Notes in Mathematics Volume 1018 (Springer-Verlag, 1983) 79–83.
[16] K. Nowicki, Asymptotic normality of graph statistics, University of Lund Department of Mathematical Statistics Technical Report (1985) 7.
[17] E. Palmer, Graphical Evolution: an introduction to the theory of random graphs. (John Wiley & Sons, 1985).
[18] A. Rucinski, When small subgraphs of a random graph are normally distributed; Probability theory and related field, in press.
[19] A. Rucinski and A. Vince, Balanced graphs and the problem of subgraphs of random graphs, Congressus Numerantium 49 (1985) 181–190.
[20] C.P. Shapiro and L. Hubert, Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. Ann. Statist. 7 (1979) 788–794.