ASYMPTOTIC NORMALITY OF GRAPH STATISTICS

Krzysztof NOWICKI

Department of Mathematical Statistics, University of Lund, Box 118, S 221 00 Lund, Sweden

Received 27 February 1986; revised manuscript received 11 February 1988
Recommended by O. Frank

Abstract: Various types of graph statistics for Bernoulli graphs are represented as numerators of incomplete U-statistics. Asymptotic normality of these statistics is proved for Bernoulli graphs in which the edge probability is constant. In addition it is shown that subgraph counts asymptotically are linear functions of the number of edges in the graph.

AMS Subject Classification: Primary 05C99; Secondary 62E99, 92A20.

Key words: Random graphs; subgraph counts; induced subgraph counts; U-statistics; incomplete U-statistics; Markov graphs.

1. Introduction

The statistical analysis of graph data is often based on subgraph counts. In some cases, subgraph counts are sufficient statistics, as, for example, in the case of certain Markov graphs, see Frank and Strauss (1986). However, for other models and for empirical graph data, subgraph counts are merely used to describe structural properties. Kendall and Babington Smith (1940) and Moran (1947) considered measures of inconsistency and agreement between judges in paired comparisons. Several authors have proposed and investigated indices for measuring transitivity of contacts and friendship choices, e.g. Holland and Leinhardt (1971, 1975), Harary and Kommel (1979), Frank (1980) and Frank and Harary (1982).

In random graphs, the subgraph counts are random variables with complicated mutual dependencies. Even for very simple graph models, studies of their asymptotic distributions have been restricted to special classes of subgraphs. Moon (1968) studied the distribution of the numbers of 3- and 4-cycles in a random tournament and proved their asymptotic normality. For the most common random graph model, so-called Bernoulli graphs (a graph in which edges occur independently with a common probability p), subgraph counts were investigated for so-called balanced subgraphs. The classical reference is two papers by Erdős and Rényi (1959, 1960). Further references can be found in the text books by Bollobás (1985) and Palmer (1985) and in the survey articles by Karoński (1982, 1984).
In statistical analysis, Bernoulli graphs can be useful as representations of pure chance phenomena; they can thus be used to test if empirical graph data are generated from some more complicated model than a Bernoulli graph (Frank, 1980). Furthermore, Bernoulli graphs have been used to describe networks of irregularly appearing contacts between vertices; see Frank (1979). Barlow and Proschan (1975) considered reliability problems by using Bernoulli graphs. Holland and Leinhardt (1975, 1979) and Wasserman (1977) considered moments of dyad and triad counts (subgraph counts of order 2 and 3) for simple random graph models, while Frank (1979) gave the first two moments of subgraph counts of arbitrary order for general random graph models.

The purpose of this paper is to study the asymptotic behaviour of subgraph counts for Bernoulli graphs in which the edge probability $p$ is constant. It is proved that subgraph counts are asymptotically normally distributed. In addition, we show that subgraph counts asymptotically are linear functions of the number of edges in the graph. This allows us to obtain simultaneous distributions for multiple counts directly from distributions of single counts.

In Section 2 we will give basic concepts and notations. Section 3 presents some results from the theory of U-statistics. In Section 4 we will give the asymptotic distribution of a single subgraph count with certain extensions in Section 5. Section 6 is devoted to investigations of the simultaneous distribution of subgraph counts, as well as their application to testing a Bernoulli graph versus a Markov graph. Finally, in Section 7 we study induced subgraph count statistics, both for a single count and for simultaneous counts.

2. Basic concepts and notations

Let $N = \{1, \ldots, n\}$ be a set of elements called vertices, and let $E$ be a subset of the collection of $n' = \binom{n}{2}$ unordered pairs of vertices. The elements of $E$ are called edges. A graph $G$ consists of an ordered pair $(N, E)$ of sets of vertices and edges. The number of vertices is called the order of $G$ while the number of edges is called the size of $G$. The size of $G$ can be at most $n'$ and a graph on $n$ vertices with size $n'$ is called a complete graph and is denoted $K_n$.

A subgraph $H = (V, \xi)$ of $G$ is itself a graph whose vertex set $V$ is a subset of $N$, and edge set $\xi$ is a subset of $E$. Moreover, if every pair of vertices $u$ and $v$ of $H$ is connected by an edge in $H$ whenever it is connected by an edge in $G$, then $H$ is an induced subgraph of $G$.

Two graphs $G$ and $H$ are isomorphic if there is a one-to-one function $\Psi$ from the vertex set of $G$ onto the vertex set of $H$, such that for any two vertices $t$ and $u$ of $G$ we have an edge between $t$ and $u$ in $G$ if and only if we have an edge between $\Psi(t)$ and $\Psi(u)$ in $H$. We let $t(G, H)$ denote the number of subgraphs of $G$ which are isomorphic to $H$. In particular, let $G = K_n$ and let $H$ be a graph of order $v$, $v < n$. Then, we have that $t(K_n, H) = \binom{n}{v} t(K_v, H)$. 
3. Asymptotic normality of an incomplete U-statistic

Hoeffding (1948) considered a sequence of independent identically distributed random variables (i.i.d. r.v.) \( X_1, \ldots, X_n \) and a symmetric function \( g(x_1, \ldots, x_k) \), which we call a kernel. He estimated the parameter \( \theta = E(g(X_1, \ldots, X_k)) \) by

\[
U_0 = \frac{1}{\binom{n}{k}} \sum_{i=1}^{\binom{n}{k}} g_i,
\]

where the \( g_i \)'s are obtained by inserting the \( \binom{n}{k} \) possible ordered subsequences \( X_{j_1}, \ldots, X_{j_k}, 1 \leq j_1 < \cdots < j_k \leq n \), in \( g \), the numbering of the \( g_i \) being arbitrary. The r.v. \( U_0 \) is called a complete U-statistic. Now, set

\[
\sigma^2 = V(E(g(X_1, \ldots, X_k) | X_1))
\]

and suppose that \( \sigma^2 > 0 \). Using Hajek’s projection method it can be proved that \( \sqrt{n}(U_0 - \theta) \) is asymptotically normally distributed with mean 0 and variance \( k^2 \sigma^2 \).

In order to calculate \( U_0 \) we have to evaluate the sum of \( \binom{n}{k} \) terms. To reduce these calculations Blom (1976) introduced an incomplete U-statistic

\[
U = \frac{1}{m} \sum_{i=1}^{m} g_i/m, \quad m < \binom{n}{k}.
\]

Denote by \( S \) the set of the \( m \) subsequences \( X_{j_1}, \ldots, X_{j_k} \) in the \( g_i \)'s. Due to the strong dependence between many of the \( g_i \)'s in \( U_0 \) we can often choose \( S \) so that \( U \) is asymptotically equivalent to \( U_0 \). This is shown by the following theorem, which was essentially given by Blom (1976).

The \( m \) subsequences \( X_{j_1}, \ldots, X_{j_k} \) appearing in \( S \), form \( m^2 \) pairs. Let \( f_c, c = 1, \ldots, k \), be the number of pairs with \( c \) \( X \)-values in common, and set \( p_c = f_c/m^2 \). Then we have

**Theorem 1.** Suppose that for \( m/n \to \infty \),

\[
p_1 = k^2/n + o(1/n)
\]

and

\[
\sum_{i=2}^{k} p_i = o(1/n).
\]

Then (A) \( \sqrt{n}(U - \theta) \) is asymptotically normally distributed with mean 0 and variance \( k^2 \sigma^2 \).

(B) \( nE(U - U^*)^2 \to 0 \) as \( n \to \infty \), where

\[
U^* = \frac{k}{n} \sum_{i=1}^{n} h(X_i) - (k - 1)\theta
\]

and

\[
h(x_i) = E(g(X_{i_1}, \ldots, X_{i_k}) | X_{i_1} = x_i), \quad 1 \leq i_1 < \cdots < i_k \leq n.
\]
Proof. For the proof of part (A) we refer the reader to Blom (1976). To prove part (B) we write
\[ E(U - U^*)^2 = E(U - U_0)^2 + E(U_0 - U^*)^2 + 2E\{(U - U_0)(U_0 - U^*)\}. \]
Furthermore, as proved by Hoeffding (1948) we have that
\[ nE(U_0 - U^*)^2 \to 0, \quad \text{as } n \to \infty, \]
and, as proved by Blom (1976), under the assumption of our theorem
\[ nE(U - U_0)^2 \to 0, \quad \text{as } n \to \infty. \]
Finally, by invoking the Minkowski inequality we obtain (B). □

We shall now give the two other conditions that ensure the asymptotic normality of \( U \). Let \( a_t, t = 1, \ldots, n, \) denote the number of \( g_i \)'s in \( U \) which contain \( X_t \), and \( a_{iu} \) the number of \( g_i \)'s which contain both \( X_t \) and \( X_u \). Then we have:

**Theorem 2.** If
\[ \sum_{t=1}^{n} \frac{a_t^2}{m^2} = \frac{k^2}{n} + o\left(\frac{1}{n}\right) \]  
and
\[ \sum_{1 \leq i < u \leq n} \frac{a_{iu}^2}{m^2} = o\left(\frac{1}{n}\right), \]
then the conditions (3.3a) and (3.3b) of Theorem 1 are satisfied.

Proof. Consider the \( g_i \)'s appearing in \( U \). There are \( a_t^2 \) pairs \((g_i, g_j)\) which have \( X_t \) in common and \( a_{iu}^2 \) pairs which have both \( X_t \) and \( X_u \) in common. We seek the number \( f_1 + \cdots + f_k \) of pairs with at least one \( X \)-value in common. Clearly, we have the double inequality
\[ \sum_{t=1}^{n} a_t^2 - \sum_{1 \leq i < u \leq n} a_{iu}^2 \leq f_1 + \cdots + f_k \leq \sum_{t=1}^{n} a_t^2. \]
Similarly it is found that
\[ f_2 + \cdots + f_k \leq \sum_{1 \leq i < u \leq n} a_{iu}^2, \]
Dividing these inequalities by \( m^2 \) and using the conditions of the theorem we have
\[ p_1 + \cdots + p_k = \frac{k^2}{n} + o\left(\frac{1}{n}\right), \quad p_2 + \cdots + p_k = o\left(\frac{1}{n}\right). \]
Hence the conditions (3.3a) and (3.3b) are satisfied. □

In the sequel we shall be particularly interested in the numerators
\[ T_0 = \sum_{i=1}^{\binom{n}{2}} g_i \quad \text{and} \quad T = \sum_{i=1}^{m} g_i \]
of \( U_0 \) and \( U \). We call these quantities complete and incomplete \( T \)-statistics, respectively. All asymptotic results for \( U_0 \) and \( U \) hold, of course, also for \( T_0 \) and \( T \) after multiplication by a suitable factor.

4. Subgraph counts

Consider a Bernoulli graph \( G_{n,p} \) on a finite set \( \{1, \ldots, n\} \) of vertices, i.e. a graph where edges occur independently with the same probability \( p \), \( 0 < p < 1 \). Such a graph can be described by the sequence \( X_{iu} \), \( 1 \leq i < u \leq n \), of \( n^t = \binom{n}{2} \) independent r.v., where

\[
X_{iu} = \begin{cases} 
1 & \text{if an edge occurs between vertices } i \text{ and } u, \\
0 & \text{otherwise}.
\end{cases}
\]

We want to count the number of subgraphs of \( G_{n,p} \) with prescribed properties. For this purpose we may use the theory of \( U \)-statistics. We begin with a simple problem leading to a complete \( T \)-statistic: count the number of all subgraphs with \( s \) edges. In what follows, we consider only subgraphs without isolated vertices, i.e. subgraphs defined by specifying their set of edges. Subgraph counts for arbitrary graphs are then obtained by multiplying these subgraph counts by a suitable factor.

For this purpose introduce the symmetric function

\[
g(x_1, \ldots, x_s) = x_1 x_2 \cdots x_s,
\]

and the corresponding complete \( T \)-statistic

\[
T_0 = \sum_{i=1}^{\binom{s}{2}} g_i.
\]

Here the \( g_i \)'s are obtained by inserting in \( g \) all possible \( s \)-subsequences taken from the sequence \( \{X_{iu}\} \). Clearly, \( g_i = 1 \) if all \( X_{iu} \) corresponding to the \( i \)-th subsequence are equal to 1. Hence \( T_0 \) counts the number of all subgraphs with \( s \) edges.

It follows from Hoeffding (1948) that \( T_0 \) is asymptotically normally distributed. In this case we have

\[
\theta = E(g) = p^s.
\]

Moreover, using (3.1) we obtain that

\[
\sigma^2 = V(E(g \mid X_{iu})) = V(p^{s-1} X_{iu}) = p^{2s-1}(1-p).
\]

Hence we have proved the following theorem:

**Theorem 3.** For \( T_0 \) defined as before,

\[
\sqrt{n^t(T_0/(\binom{n}{s}) - p^s)}
\]

is asymptotically normally distributed with mean 0 and variance \( s^2 p^{2s-1}(1-p) \).
We now consider a fixed subgraph $H$ of order $v$ and size $s$ and seek the number $T_H = t(G_{n,p}, H)$ of subgraphs of $G_{n,p}$ isomorphic to $H$.

Introduce an incomplete $T$-statistic

$$T_H = \sum_{i=1}^{t(K_n, H)} g_i$$

with the same kernel $g$ as in (4.1) and with $t(K_n, H)$ being the number of subgraphs isomorphic to $H$ in $K_n$. By suitable choice of a set $S_H$ consisting of $t(K_n, H)$ $s$-subsequences from $\{X_{tu}\}$ and inserting these subsequences in $g$, $T_H$ counts the number of subgraphs isomorphic to $H$.

For example, take $s = 4$, $n = 4$ and let $H$ be a 4-cycle, i.e. an undirected graph on four vertices in which each vertex is connected with two other vertices. Set

$$T_H = g(X_{12}, X_{13}, X_{24}, X_{34}) + g(X_{12}, X_{14}, X_{23}, X_{34}) + g(X_{13}, X_{14}, X_{23}, X_{24}).$$

Clearly, $T_H$ counts the number of 4-cycles in $G_{4,p}$.

For any given $H$ the incomplete $T$-statistic in (4.3) is asymptotically normally distributed. To prove this we shall check that the conditions of Theorem 2 are satisfied for $n = n'$.

Let $a_{tu}$ be the frequency of the single variable $X_{tu}$ in $S_H$, also let $a_{tu, wz}$ be the frequency of the pair $(X_{tu}, X_{wz})$; and $a_{tu, uw}$, the frequency of the pair $(X_{tu}, X_{uw})$. Due to the invariance of the structure of $S_H$ when vertices are relabeled, the frequencies $a_{tu}, a_{tu, wz}, a_{tu, uw}$ do not depend on the indices. Let us therefore call them $c_1, c_2, c_3$, respectively. It is immediately seen that $c_1 = s t(K_n, H)/n'$; taking $m = t(K_n, H)$, $a_i = a_{tu}$ in condition (3.4a) we see that this condition is fulfilled.

We now turn to condition (3.4b). Its left side takes the form

$$\frac{3(c_1^2 + 3(c_2^2))}{(t(K_n, H))^2},$$

where we have used the fact that there are $3(c_1)$ pairs of type $(X_{tu}, X_{wz})$ and $3(c_2)$ pairs of type $(X_{tu}, X_{uw})$. In order to find $c_2$ and $c_3$ denote the order of $H$ by $v$ and note that a pair $(X_{tu}, X_{wz})$ can be completed to a subsequence of $S_H$ in $a_1(n-4)$ ways, where $a_1$ depends on the form of $H$ but not on $n$. Moreover, a pair $(X_{tu}, X_{uw})$ can be completed to a subsequence of $S_H$ in $a_2(n-3)$ ways, even here $a_1$ depends on the form of $S_H$ but not on $n$ and $n'$. Hence $c_2$ is of order $n^{v-4}$, and $c_3$ of order $n^{v-3}$. Noting that $t(K_n, H)$ is of order $n^v$, we see from (4.4) that the left side in condition (4.3b) is of order $n^{-3}$. But $n' = (\frac{n}{4})$. Thus the left side of (3.4b) converges faster than $1/n'$ and hence (3.4b) is satisfied for $n = n'$.

We now observe that we still have $\theta$ and $\sigma^2$ as in (4.2ab). Thus, from Theorem 1 we obtain:

**Theorem 4.** Let $T_H$ count subgraphs isomorphic to a given graph $H$ of size $s$ and order $v$ in $G_{n,p}$. Then

$$\sqrt{n'}(T_H/t(K_n, H) - p^v)$$
is asymptotically normally distributed with mean 0 and variance $s^2 p^{2s-1}(1-p)$.

Let us now investigate the asymptotic distributions for some special subgraph count statistics. Frank and Strauss (1986) showed that the number of triangles and the number of $k$-stars for $k = 1, \ldots, n-1$ are sufficient statistics for any homogeneous Markov graph of order $n$. The distributions of these statistics in the case of a Bernoulli graph can be obtained from Theorem 4, namely:

**Corollary 5.** If $T$ is the number of triangles, we have that
\[
\sqrt{n} (T/(\binom{n}{3}) - p^3)
\]
is asymptotically normally distributed with mean 0 and variance $9(p^5 - p^6)$.

**Corollary 6.** If $S_k$ is the number of $k$-stars, where a $k$-star has order $k+1$ and size $k$, we have that
\[
\sqrt{n} (S_k/((\binom{n}{k+1})(k+1)) - p^k)
\]
is asymptotically normally distributed with mean 0 and variance $k^2(p^{2k-1} - p^{2k})$.

5. Extensions

The ideas developed in Section 4 can be generalized to counting the number of all subgraphs isomorphic to at least one of a given set of graphs $H_1, \ldots, H_L$, each of size $s$. As before, we consider only graphs without isolated vertices. For this purpose introduce an incomplete $T$-statistic
\[
T = \sum_{i=1}^{m} g_i
\]
with the same kernel $g(x_1, \ldots, x_s) = x_1 x_2 \cdots x_s$ as before. Let $S$ be the union of $S_{H_1}, \ldots, S_{H_L}$ and let $m$ denote the number of subsequences in $S$. Clearly, $T$ counts the number of subgraphs isomorphic to at least one of $H_1, \ldots, H_L$. Now, checking the conditions of Theorem 2 for $S$ we prove that $T$ is asymptotically normally distributed.

First, $a_{iu}$ is the sum of $a_{iu}^1, \ldots, a_{iu}^L$ which are frequencies of $X_{iu}$ in $S_{H_1}, \ldots, S_{H_L}$, respectively. Hence
\[
a_{iu}^2 = (t(K_n, H_1) + \cdots + t(K_n, H_L))^2 s^2/n^2
\]
and condition (3.4a) in Theorem 2 is fulfilled.

Second, the left side of condition (3.4b) takes the form
\[
(3\binom{s}{2} c_2^2 + 3\binom{s}{3} c_3^2)/m^2
\]
where now the orders of magnitude of $c_2, c_3$ and $m$ are determined by the highest
order among $H_1, \ldots, H_L$. Hence condition (3.4b) holds true because it is true for each $S_{H_i}$.

Thus we have proved that $T$ in (5.1) is asymptotically normally distributed and the following theorem can be formulated:

**Theorem 7.** Let $T$ count the number of all subgraphs isomorphic to at least one of a given set of graphs $H_1, \ldots, H_L$, each of size $s$ in $G_{n,p}$. Then

$$\sqrt{n'} (T/m - p^s),$$

where

$$m = t(K_n, H_1) + \cdots + t(K_n, H_L),$$

is asymptotically normally distributed with mean 0 and variance $s^2 p^{2s-1} (1 - p)$.

In particular, let us count all subgraphs in $G_{n,p}$ of order $v$ and size $s$. Then we obtain the following corollary to Theorem 7:

**Corollary 8.** Let $T$ count all subgraphs in $G_{n,p}$ of order $v$ and size $s$, where $s > (v-1)/2$. Then

$$\sqrt{n'} (T/m - p^s),$$

where

$$m = \binom{v}{s} \binom{s}{s},$$

is asymptotically normally distributed with mean 0 and variance $s^2 p^{2s-1} (1 - p)$.

6. **Asymptotic linear dependency between subgraph counts**

For the purpose of investigating the asymptotic distributions of multiple subgraph counts and linear combinations of subgraph counts the following theorem is useful:

**Theorem 9.** Let $T_H$ count subgraphs in $G_{n,p}$ isomorphic to a given graph $H$ of order $v$ and size $s$. Then, it holds that

$$E\{\sqrt{n'} (T_H/t(K_n, H) - p^s) - \sqrt{n'} (sp^{s-1} R/n' - sp^s)\}^2 \to 0, \quad \text{as } n \to \infty,$$

where $R$ is the number of edges in $G_{n,p}$.

**Proof.** By Theorem 1(B) we have that $\sqrt{n'} (T_H/t(K_n, H) - p^s)$ is asymptotically equal (in mean square) to

$$\sqrt{n'} \left( \frac{s}{n'} \sum_{1 \leq i < j \leq n} h(X_{ij}) - sp^s \right). \quad (6.1)$$

Furthermore, we have that

$$h(X_{ij}) = E(g \mid X_{ij}) = p^{s-1} X_{ij}.$$
Thus, (6.1) is equal to
\[ \sqrt{n'} \left( \frac{s}{n'} \sum_{1 \leq i < j \leq n} p^{s-1} X_{ij} - sp^s \right), \]
and by substituting \( R = \sum_{1 \leq i < j \leq n} X_{ij} \) we obtain the result of the theorem. ☐

Consider now \( T = (T_1, \ldots, T_r) \) to be a vector of \( r \) subgraph counts. Each \( T_j \) is assumed to count subgraphs isomorphic to a given graph \( H_j \) of order \( v_j \) and size \( s_j \), respectively. The following result is obtained from Theorem 9:

**Corollary 10.** The random vector \((T_1, \ldots, T_r)\) where
\[ \tilde{T}_j = \sqrt{n'} \left( T_j / (t(K_{n}, H_j) - p s) \right), \quad j = 1, \ldots, r, \]
is asymptotically normally distributed with mean vector 0 and covariance matrix \( C = (C_{ij})_{i,j=1}^r \) of rank 1. Here
\[ C_{ij} = C_i C_j, \quad \text{where} \quad C_i = s_j p^{s_j} \sqrt{1 - p/p}, \quad i = 1, \ldots, r. \]

**Proof.** By Theorem 4 each component of the considered vector is asymptotically normally distributed. Thus, it remains to prove that a linear combination \( \sum_{j=1}^r \alpha_j \tilde{T}_j \) is asymptotically normally distributed. First, we introduce
\[ X_j = \tilde{T}_j - \sqrt{n'} \left( s_j p^{s_j-1} R/n' - s_j p s_j \right), \quad j = 1, \ldots, r, \]
i.e. the distance between \( \tilde{T}_j \) and its projection on \( \{X_{ij}, 1 \leq i < j \leq n\} \). We have that \( EX_j = 0 \) and, from Theorem 9, that \( V(X_j) = E(X_j^2) \to 0 \) as \( n \to \infty \). Second, we have that \( E(\sum_{j=1}^r \alpha_j X_j)^2 \to 0 \) as \( n \to \infty \). Namely,
\[ E\left( \sum_{j=1}^r \alpha_j X_j \right)^2 = \sum_{j=1}^r \alpha_j^2 E(X_j^2) + 2 \sum_{1 \leq k < m \leq n} \alpha_k \alpha_m E(X_k X_m) \]
\[ \leq \sum_{j=1}^r \alpha_j^2 E(X_j^2) + 2 \sum_{1 \leq k < m \leq n} |\alpha_k \alpha_m| \{E(X_k^2) E(X_m^2)\}^{1/2} \]
\[ \to 0 \quad \text{as} \quad n \to \infty. \]
Thus, we have that
\[ E\left( \sum_{j=1}^r \alpha_j \tilde{T}_j - \sqrt{n'} \beta (R/n' - p) \right)^2 \to 0 \quad \text{as} \quad n \to \infty \]
where
\[ \beta = \sum_{j=1}^r \alpha_j s_j p^{s_j-1}. \]
Finally, since \( R \in \text{Bin}(n', p) \) we have that \( \sqrt{n'} (R/n' - p) \) and consequently \( \sum_{j=1}^r \alpha_j \tilde{T}_j \) are asymptotically normally distributed.
To calculate the covariance matrix we write
\[ C_{ij} = C(a_i + b_i R, a_j + b_j R) - b_i b_j \text{Var}(R). \]
Hence, since \( b_i = s_i p_i^{s_i - 1} / \sqrt{n} \), \( b_j = s_j p_j^{s_j - 1} / \sqrt{n} \) and \( \text{Var}(R) = n' p(1 - p) \), we obtain the covariance matrix. Finally, since each component of the vector is a multiple of \( R \) we obtain that the rank is equal to 1.

Another application of Theorem 9 is in the investigation of linear combinations of subgraph counts. We will consider the example of testing a Bernoulli graph against a Markov graph. Here, we devote ourselves to a Markov graph model used by Frank and Strauss (1986) with probability function
\[ P_M(G) = c^{-1} \exp(\varrho r + \sigma s + \tau t), \]
where \( r, s, t \) are equal to the number of edges, 2-stars and triangles in graph \( G \), respectively. This model captures both transitivity and clustering in graph data and can be seen as a natural generalisation of a Bernoulli graph model. In fact if \( \sigma = \tau = 0 \), \( P_M(G) \) reduces to a probability function of a Bernoulli graph \( G_{n,p} \) which will be denoted by
\[ P_B(G) = p^r(1 - p)^{n - r} \quad \text{where} \quad p = \varrho / (1 + e^\varrho). \]
Consider now the hypothesis
\[ H_0: \quad P_B(G) = c^{-1} \exp(\varrho r) \]
against
\[ H_1: \quad P_M(G) = c_1^{-1} \exp(\varrho_1 r + \sigma_1 s + \tau_1 t) \]
where \( \varrho, \varrho_1, \sigma_1, \) and \( \tau_1 \) are specified values of the parameters. By using the Neyman–Pearson lemma we obtain that the critical region of the likelihood test is of the type
\[ (\varrho - \varrho_1) r + \sigma_1 s + \tau_1 t > k. \]
Now, let \( R, S \) and \( T \) be the number of edges, 2-stars and triangles in \( G_{n,p} \), respectively. Then, from Theorem 9 we obtain that
\[ S = 2p(n - 2)R - 3p^2(\xi) + \varepsilon_s \quad \text{and} \quad T = p^2(n - 2)R - 2p^3(\xi) + \varepsilon_t \]
where r.v. \( \varepsilon_s \) and \( \varepsilon_t \) are such that
\[ E(\varepsilon_s) = E(\varepsilon_t) = 0 \quad \text{and} \quad \sigma^2(\varepsilon_s) = \sigma^2(\varepsilon_t) = O(n^2). \]
Consequently, under \( H_0 \) the critical region is approximately
\[ ((\varrho - \varrho_1) + 2\sigma_1 p(n - 2) + \tau_1 p^2(n - 2))R > k + 3\sigma_1 p^2(\xi) + 2\tau_1 p^3(\xi), \]
from which we finally obtain that the test is of the following type:
\[ R > k_1 \quad \text{if} \quad (\varrho - \varrho_1) + 2\sigma_1 p(n - 2) + \tau_1 p^2(n - 2) > 0, \]
\[ R < k_2 \quad \text{if} \quad (\varrho - \varrho_1) + 2\sigma_1 p(n - 2) + \tau_1 p^2(n - 2) < 0. \]
7. Induced subgraph counts

Consider, as before, $G_{n,p}$ to be a Bernoulli graph. We want to count the number $T_{H}^{\text{ind}}$ of induced subgraphs of $G_{n,p}$ isomorphic to a given graph $H$ of order $v$ and size $s$. We assume that $H$ does not have isolated vertices; for graph $H$ with isolated vertices we count instead induced subgraphs of $G_{n,1-p}$ isomorphic to the complement of $H$. Since an induced subgraph of $G_{n,p}$ having $i_1, \ldots, i_v$, say, as vertices, is either isomorphic to $H$ or nonisomorphic to $H$ we can represent $T_{H}^{\text{ind}}$ as

$$ T_{H}^{\text{ind}} = \sum_{1 \leq i_1 < \ldots < i_v \leq n} g(X_{i_1}, \ldots, X_{i_v}) $$

where

$$ g(X_{i_1}, \ldots, X_{i_v}) = \begin{cases} 1 & \text{if the subgraph of } G_{n,p} \text{ induced by} \\ i_1, \ldots, i_v \text{ is isomorphic to } H, \\ 0 & \text{otherwise.} \end{cases} $$

For instance, take $H$ to be a cycle of size 4 and order 4. Then,

$$ g(X_{ij}, X_{jk}, X_{jk}, X_{jkl}, X_{kl}) = X_{ij}X_{jk}(1 - X_{ik})(1 - X_{jl}) + X_{ij}X_{ik}X_{jl}(1 - X_{ij})(1 - X_{jk}) + X_{jk}X_{ij}X_{jl}(1 - X_{ik})(1 - X_{kl}). $$

It is easy to verify that $g$ is no longer a symmetric function and consequently $T_{H}^{\text{ind}}$ is not a numerator of an incomplete U-statistic. However, $T_{H}^{\text{ind}}$ can be written as a linear combination of subgraph counts for all subgraphs $H_i$ such that $H \subseteq H_i \subseteq K_v$. Namely,

$$ T_{H}^{\text{ind}} = \sum_{H \subseteq H_i \subseteq K_v} (-1)^{s_i - s} t(H_i, H) T_{H_i} $$

where $s_i$ is the size of $H_i$ and $T_{H_i}$ is the number of subgraphs of $G_{n,p}$ isomorphic to $H_i$. Thus we arrive at the following theorem:

**Theorem 11.** Let $T_{H}^{\text{ind}}$ count induced subgraphs in $G_{n,p}$ isomorphic to a given graph $H$ of order $v$ and size $s$. Then

$$ \sqrt{n} \left( \frac{T_{H}^{\text{ind}}}{\binom{n}{v}} - \theta \right), \quad \text{where} \quad \theta = t(K_v, H)p^v(1-p)^{v-s}, $$

is asymptotically normally distributed with mean 0 and variance $\theta^2(s - \binom{s}{2})p^2/(p(1-p))$.

**Proof.** Asymptotical normality of $T_{H}^{\text{ind}}$ follows from Corollary 10 since $T_{H}^{\text{ind}}$ is a linear combination of subgraph counts of the same order. Moreover, the expectation of $T_{H}^{\text{ind}}$ can be easily obtained from representation (7.1) since each function $g(X_{i_1}, \ldots, X_{i_v})$ is a sum of $t(K_v, H)$ random indicators, each with the expectation $p^v(1-p)^{v-s}$.

In order to obtain the variance we use (7.2) and write

$$ \sqrt{n} \left( \frac{T_{H}^{\text{ind}}}{\binom{n}{v}} - \theta \right) = \sqrt{n} \left( \sum_{H \subseteq H_i \subseteq K_v} (-1)^{s_i - s} t(H_i, H) T_{H_i}/\binom{n}{v} - \theta \right). $$
Thus we have to calculate
\[ n' \text{Var} \left( \sum_{H \subseteq H \subseteq K_v} (-1)^{s_i-s} t(H_i, H) T_{H_i}/(\binom{n}{s_i}) \right) \]
which, by arguments as in the proof of Corollary 10, is asymptotically equal to
\[ n' \text{Var} \left( \sum_{H \subseteq H \subseteq K_v} (-1)^{s_i-s} t(H_i, H) s_{ij} p^{s_j-1} t(K_{u_j}, H_i) R/n' \right) \]
and since \( \text{Var}(R) = n'pq \) to
\[ pq \left( \sum_{H \subseteq H \subseteq K_v} (-1)^{s_i-s} t(H_i, H) s_{ij} p^{s_j-1} t(K_{u_j}, H_i) \right)^2. \] (7.3)
Finally, this expression can be simplified by calculating the expectation \( \theta \) using (7.2). Namely, we obtain that
\[ \theta = \sum_{H \subseteq H \subseteq K_v} (-1)^{s_i-s} t(H_i, H) t(K_{u_j}, H_i) p^{s_j} \]
and consequently
\[ \frac{d\theta}{dp} = \sum_{H \subseteq H \subseteq K_v} (-1)^{s_i-s} t(H_i, H) t(K_{u_j}, H_i) s_{ij} p^{s_j-1}. \]
Hence, expression (7.3) can be written \( pq (d\theta/dp)^2 \) and consequently variance is obtained by evaluating the expression
\[ pq \left( \frac{d(t(K_{u_j}, H)p^{s_j}(1-p)^{u_j-s_j})}{dp} \right)^2. \]

Remark 12. Observe that when \( p = s/(\binom{n}{2}) \) the asymptotic variance in Theorem 11 is equal to 0 and we obtain only the convergence in probability to 0. Maehara (1987) proved, independently, Theorem 11 when \( p = 0.5 \).

Let us now consider \( T = (T_1, \ldots, T_r) \) to be a vector of induced subgraph count statistics, in which each component \( T_j \) counts induced subgraphs \( H_j \) of order \( u_j \) and size \( s_j \), respectively. Then we can give the following theorem:

Theorem 13. We have that
\[ (\sqrt{n'}(T_1/(\binom{n}{u_1}) - \theta_1), \ldots, \sqrt{n'}(T_r/(\binom{n}{u_r}) - \theta_r) \]
is asymptotically normally distributed with mean vector 0 and covariance matrix \( C = (C_{ij})_{i,j=1}^r \) of rank 1. Here
\[ \theta_i = t(K_{u_j}, H_i)p^{s_j}(1-p)^{u_j-s_j} \]
and
\[ C_{ij} = C_i C_j, \quad C_i = \theta_i (s_i - \binom{u_j}{2} p) / \sqrt{(1-p)p}, \quad i, j = 1, \ldots, r. \]
Proof. The asymptotic normality can be proved by a method similar to that used to prove Corollary 10. In order to establish covariance matrix \( C \) we use a method similar to that used to prove Theorem 11. It allows us to conclude that

\[
C_{ij} = \text{Cov}\left\{ \sum_{H_k \subset H_i \subset K_i} (-1)^{x-i} t(H_k, H_i) s_k p^{x-1} t(K_{\alpha}, H_k) R / \sqrt{n'}, \right. \\
\left. \sum_{H_j \subset H_m \subset K_j} (-1)^{y-m} t(H_m, H_j) s_m p^{y-1} t(K_{\alpha}, H_m) R / \sqrt{n'} \right\}
\]

and after calculation we obtain that

\[
C_{ij} = pq (d \theta_i / dp)(d \theta_j / dp).
\]

Finally, the evaluation of this expression gives the covariance matrix. \( \square \)

Consider now \( T = (T_0, T_1, T_2, T_3) \) to be a vector of triad counts for a Bernoulli graph where the components \( T_r \) give the number of triads with \( r \) edges where \( r = 0, \ldots, 3 \), i.e. \( T_r \) is of order 3 and size \( r \). Frank (1979) obtained the first- and second-degree moments of triad counts. We can use Theorem 13 in order to generalize his results. The following corollary gives the asymptotic distribution of \( T \):

**Corollary 14.** Let \( T = (T_0, \ldots, T_3) \) be a vector of triad counts in \( G_{n,p} \). Then

\[
\sqrt{n'} \left( T / \binom{3}{2} \right) \theta,
\]

where

\[
\theta = (\theta_0, \ldots, \theta_3), \quad \theta_i = \binom{3}{i} p^i (1 - p)^{3-i} \text{ for } i = 0, \ldots, 3,
\]

is asymptotically normally distributed with mean vector 0 and covariance matrix \( C = (C_{ij})_{i,j=0} \) where

\[
C_{ij} = C_i C_j
\]

and

\[
C_i = \binom{3}{i} p^i (1 - p)^{3-i} (3p - i) / \sqrt{(1 - p)p}, \quad i, j = 0, \ldots, 3.
\]

**Acknowledgement**

I am indebted to Professor Gunnar Blom and Professor Ove Frank for offering so much of their time for discussion, valuable suggestions and comments. I also want to thank Professor Svante Jansson who gave me valuable comments and Dozent Daniel Thorburn who further extended my understanding of the asymptotic behaviour of subgraph counts.
References