ON COMBINATORIAL HOEFFDING DECOMPOSITION AND ASYMPTOTIC NORMALITY OF SUBGRAPH COUNT STATISTICS

M. Bloznelis

Vilnius University and Institute of Mathematics and Informatics
Naugarduko 24, Vilnius, Lithuania; E-MAIL: mblozn@ieva.maf.vu.lt

Abstract. Given k and n, consider a graph with k vertices and n "blue" edges. We assume that the set of "blue" edges $\{X_1, \dots, X_n\}$ is uniformly distributed among n-subsets of $N = \genfrac{}{}{0pt}{}{k}{2}$ pairs of vertices. Given a graph G, the number $\mathcal{N}_G$ of blue copies of G is a U-statistic based on random sample $X_1, \dots, X_n$. We show how the combinatorial Hoeffding decomposition of the random variable $\mathcal{N}_G$ can be applied to establish the asymptotic normality of $\mathcal{N}_G$ as $k, n \to \infty$. Several examples are considered.

1. Introduction. Given a complete graph based on k vertices let $\mathcal{X} = \{x_1, \dots, x_N\}$ denote the set of edges. Let $\mathbb{X} = \{X_1, \dots, X_n\} \subset \mathcal{X}$ be a random n-subset uniformly distributed over the class of n-subsets of $\mathcal{X}$. Here $n < N$. We paint edges $X_1, \dots, X_n$ blue. The graph based on k vertices and (random) blue edges is denoted by $G(k, n)$. Given a graph G let $\mathcal{N}_G$ denote the number of copies of G in $G(k, n)$. We are interested when the random variable $(\mathcal{N}_G - \mathbf{E}\, N_G)/\sigma(N_G)$ is asymptotically standard normal as $k, n \to \infty$. Here $\sigma^2(\mathcal{N}_G)$ denotes the variance of $\mathcal{N}_G$.

Another random graph model assumes that edges become blue independently with probability $t \in (0, 1)$. Let $\nu_1, \dots, \nu_N$ be independent Bernoulli random variables with success probability t, i.e., $\mathbf{P}\{\nu_i = 1\} = 1 - \mathbf{P}\{\nu_i = 0\} = t$ for every i. We paint the edge x_i blue if $\nu_i = 1$. The graph based on k vertices and (random) blue edges is denoted by $G'(k, t)$ and called Bernoulli random graph. Given a graph G let $\mathcal{N}'_G = N'_G(k, t)$ denote the number of copies of G in $G'(k, t)$. Note that the

1991 *Mathematics Subject Classification.* Primary 60C05; secondary 60F05.

Key words and phrases. Hoeffding decomposition, combinatorial U-statistics, subgraph count, central limit theorem.

conditional distribution of $\mathcal{N}'_G$ given the event $\nu_1 + \cdots + \nu_N = n$ coincides with the distribution of $\mathcal{N}_G$. Therefore, the problems of the asymptotic normality of distributions of $\mathcal{N}_G$ and $\mathcal{N}'_G$ are closely related.

The asymptotic normality for $\mathcal{N}'_G$ as $k \to \infty$ was studied by several authors using different methods (method of moments, Stein's method, projection method and martingale limit theorems). An overview of the results and methods is given in the book of Janson, Luczak and Rucinski (2000). Let us mention that the first complete description of the conditions that are necessary and sufficient for the asymptotic normality of $\mathcal{N}'_G$ was given by Rucinski (1988).

The asymptotic normality of $\mathcal{N}_G$ is shown in Janson (1990). He consider the random graph process $\{G(k,t), t \in [0,1]\}$, where for every t the random graph $G(k,t)$ is defined as above, but with $\nu_i = \nu_i(t) = \mathbb{I}_{\{u_i \leq t\}}$. Here $u_1, u_2, \ldots$ denote independent random variables uniformly distributed in $[0,1]$. For every $k = 1, 2, \ldots,$ the collection of random variables $\{N'_G(k,t), t \in [0,1]\}$ can be viewed as a random process with sample paths in the Skorokhod space $D[0,1]$. Using a martingale convergence theorem Janson (1990) proved a functional limit theorem for the sequence of random processes $\{\mathcal{N}'_G(k,\cdot)\}$ and then derived the asymptotic normality of $N'_G(k,t_n)$, where the random times $t_n = \min\{t : \nu_1(t) + \cdots + \nu_N(t) = n\}$. Since the distributions of $N'_G(k,t_n)$ and $\mathcal{N}_G$ coincide, this implies the asymptotic normality of $\mathcal{N}_G$.

The present paper proposes another approach to the asymptotic normality of $\mathcal{N}_G$.

Using this approach we show the asymptotic normality for the simplest subgraph count statistic: the number $T = \mathcal{N}_{P_2}$ of $2-$stars ($G = P_2$). Denote $n^* = \min\{n, N-n\}$.

Proposition. *Assume that $\sigma^2(T) \to \infty$ as $k, n^* \to \infty$. Then the distribution of $(T - \mathbf{E}\,T)/\sigma(T)$ is asymptotically standard normal.*

By $\sigma^2(T)$ we denote the variance of the random variable T. The proof combines projection's method and Stein's method. By means of Hoeffding's decomposition, the random variable $\mathcal{N}_G$ is expanded into a sum of mutually uncorrelated U statistics of random variables $X_1, \ldots, X_n$. The decomposition enables us to write the characteristic function $f(t)$ of (properly standardized) random variable $\mathcal{N}_G$ in Erdős-Rényi form, see Erdős and Rényi (1959). Finally, we show that

$$f'(t) + tf(t) \to 0 \tag{1}$$

as $k, n \to \infty$. This implies the asymptotic normality, see Stein (1970) and Tikhomirov (1976, 2001).

The remaining part of the paper is organized as follows. In Section 2 we construct Hoeffding's decomposition for three examples of subgraph count statistics: the number of $2-$stars ($G = P_2$), the number of triangles ($G = K_3$) and the number of $4-$cycles ($G = C_4$). In section 3 we write Erdős-Rényi representation for characteristic function f and prove Proposition 1.

2. Decomposition. Let $(X_1, \dots, X_N)$ denote the random permutation of the sequence $x_1, \dots, x_N$. A real function $t(x_{i_1}, \dots, x_{i_n})$ defined on $n-$ subsets $\{x_{i_1}, \dots, x_{i_n}\} \subset \mathcal{X}$ defines the random variable $T = t(X_1, \dots, X_n)$. Hoeffding's decomposition expands the random variable T in the series of $U-$ statistics

$$T = \mathbf{E}\, T + U_1 + \dots + U_{n^*}, \qquad n^* = \min\{n, N-n\}. \tag{2}$$

Here $U_r = \sum_{1 \le i_1 < \dots < i_r \le n} g_r(X_{i_1}, \dots, X_{i_r})$. The function g_r is defined on $r-$ subsets of $\mathcal{X}$ and satisfies $\mathbf{E}\, g_r(X_{i_1}, \dots, X_{i_r}) = 0$. Furthermore, for every $s < r$ and every $1 \le i_1 < \dots < i_r \le N$ and $1 \le j_1 < \dots < j_s \le N$ we have almost surely

$$\mathbf{E}\left(g_r(X_{i_1}, \dots, X_{i_r}) \middle| X_{j_1}, \dots, X_{j_s}\right) = 0. \tag{3}$$

The kernels $g_r(x_{i_1}, \dots, x_{i_r})$ are linear combinations of conditional expectations $\mathbf{E}\,(T | X_1 = y_1, \dots, X_j = y_j)$ for $1 \le j \le r$ and $\{y_1, \dots, y_j\} \subset \{x_{i_1}, \dots, x_{i_r}\}$. Basic facts about the decomposition and formulas defining g_r can be found in Bloznelis and Götze (2001) and Bloznelis (2003), see also Zhao and Chen (1990). Note that (3) implies that random variables U_r are mutually uncorrelated. This yields the variance decomposition

$$\sigma^2(T) = \sigma^2(U_1) + \dots + \sigma^2(U_{n^*}).$$

Furthermore using (3) it is easy to show, see Bloznelis and Götze (2001), that

$$\sigma^2(U_r) = \frac{\genfrac{}{}{0pt}{}{n}{r}\genfrac{}{}{0pt}{}{N-n}{r}}{\genfrac{}{}{0pt}{}{N-r}{r}}\,\sigma_r^2 \sim \frac{N^r}{r!}\,(pq)^r \sigma_r^2, \tag{4}$$

as $k, n^* \to \infty$. Here

$$p = \frac{n}{N}, \qquad q = \frac{N-n}{N}, \qquad \sigma_r^2 = \mathbf{E}\, g_r^2(X_{i_1}, \dots, X_{i_r}).$$

2.1. Let T denote the number of blue copies of $2-$star P_2. Given two edges $x, y \in \mathcal{X}$ let L_{xy} be the indicator of the event that x and y are incident. We have $T = \sum_{1 \le i < j \le n} L_{X_1 X_2}$. Hoeffding's decomposition

$$T = \mathbf{E}\, T + U_1 + U_2, \qquad U_1 \equiv 0, \qquad U_2 = \sum_{1 \le i < j \le n} \bigl(L_{X_i X_j} - p_L\bigr). \tag{5}$$

Here $p_L := \mathbf{E}\, L_{X_i X_j} = 2(k-2)/(N-1)$ and $\sigma_2^2 = p_L(1 - p_L)$.
The variance

$$\sigma^2(T) = \sigma^2(U_2) = \frac{\genfrac{}{}{0pt}{}{n}{2}\genfrac{}{}{0pt}{}{N-n}{2}}{\genfrac{}{}{0pt}{}{N-2}{2}}\,\sigma_2^2 \sim \frac{(pq)^2}{2}\,k^3 \qquad \text{as} \quad k, n^* \to \infty. \tag{6}$$

2.2. Let T denote the number of blue copies of the triangle K_3. Given three edges $x, y, z \in \mathcal{X}$, let $a(x, y, z)$ be the indicator of the event that x, y and z make up a triangle. We have $T = \sum_{1 \le i_1 < i_2 < i_3 \le n} a(X_{i_1}, X_{i_2}, X_{i_3})$. Hoeffding's decomposition

$$
\begin{aligned}
(7) \qquad & T = \mathbf{E}\, T + U_1 + U_2 + U_3, \\
& g_1 \equiv 0, \qquad g_2(x, y) = \frac{n-2}{N-4}\,\bigl(L_{xy} - p_L\bigr), \\
& g_3(x, y, z) = a(x, y, z) - p_a - \frac{1}{N-1}\,\bigl(L_{xy} + L_{xz} + L_{yz} - 3p_L\bigr).
\end{aligned}
$$

Here $p_a := \mathbf{E}\, a(X_1, X_2, X_3) = \frac{\genfrac{}{}{0pt}{}{k}{3}}{\genfrac{}{}{0pt}{}{N}{3}}$. Using (3), (4) we obtain as $k, n^* \to \infty$

$$
(8) \qquad \sigma^2(U_2) \sim \frac{p^4 q^2}{2}\, k^3, \quad \sigma^2(U_3) \sim \frac{(pq)^3}{6}\, k^3, \quad \sigma^2(T) \sim \frac{(pq)^3}{6}\, (3\,\frac{p}{q} + 1) k^3.
$$

2.3. Let T denote the number of blue copies of the cycle C_4. Given edges $x, y, z, w \in \mathcal{X}$, let $b(x, y, z)$ be the indicator of the event that x, y and z make up a path and let $d(x, y, z, w)$ be the indicator of the event that x, y, z and w make up a cycle. We have $T = \sum_{1 \le i_1 < i_2 < i_3 < i_4 \le n} d(X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4})$. Hoeffding's decomposition

$$
\begin{aligned}
& T = \mathbf{E}\, T + U_1 + U_2 + U_3 + U_4, \\
& g_1 \equiv 0, \qquad g_2(x, y) = \binom{n-2}{2}\, \frac{N-2}{N-4}\, \frac{N-3}{N-5}\, Q_{\{x,y\}}, \\
& g_3(x, y, z) = (n-3)\, \frac{N-3}{N-6}\, Q_{\{x,y,z\}}, \\
& g_4(x, y, z, w) = d(x, y, z, w) - p_d - \frac{N-3}{N-6} \sum_{A \subset \{x,y,z,w\}, |A|=3} Q_A \\
& \qquad\qquad - \frac{N-2}{N-4}\, \frac{N-3}{N-5} \sum_{A \subset \{x,y,z,w\}, |A|=2} Q_A.
\end{aligned}
$$

Here we denote

$$
\begin{aligned}
& Q_{\{xy\}} = \frac{k-5}{\genfrac{}{}{0pt}{}{N-2}{2}}\, \bigl(L_{xy} - p_L\bigr), \\
& Q_{\{xyz\}} = \frac{1}{N-3}\, \bigl(b(x, y, z) - p_b\bigr) - \frac{N-2}{N-4} \sum_{A \subset \{x,y,z\}, |A|=2} Q_A.
\end{aligned}
$$

Furthermore,

$$
p_d := \mathbf{E}\, d(X_1, X_2, X_3, X_4) = 3\, \frac{\genfrac{}{}{0pt}{}{k}{4}}{\genfrac{}{}{0pt}{}{N}{4}}, \qquad p_b = \mathbf{E}\, b(X_1, X_2, X_3) = 12\, \frac{\genfrac{}{}{0pt}{}{k}{4}}{\genfrac{}{}{0pt}{}{N}{3}}.
$$

Using (3), (4) we obtain as $k, n^* \to \infty$

$$\sigma^2(U_2) \sim \frac{p^6 q^2}{2}\, k^5, \qquad \sigma^2(U_3) \sim \frac{p^5 q^3}{2}\, k^4, \qquad \sigma^2(U_4) \sim \frac{p^4 q^4}{8}\, k^4. \tag{9}$$

2.3. Examples suggest that the linear part of the decomposition of $\mathcal{N}_G$ vanishes, i.e. $U_1 \equiv 0$ almost surely. This can be easily shown for arbitrary G. Furthermore, different $U-$ statistics contributing to the Hoeffding's decomposition (2) may have the same stochastic order even in the case where the parameter $p = p(k)$ is bounded away from 0 and 1, i.e., for some $\varepsilon > 0$,

$$\varepsilon < p(k) < 1 - \varepsilon \qquad \text{as} \qquad k, n^* \to \infty. \tag{10}$$

Thus, we have $\sigma^2(U_j) = \Theta(k^3)$ for $j = 2, 3$ in (8) and $\sigma^2(U_j) = \Theta(k^4)$ for $j = 3, 4$ in (9). Similarly, Hoeffding's decomposition of the number $\mathcal{N}_{K_4}$ of blue copies of the complete graph K_4, is the sum $\mathbf{E}\,\mathcal{N}_{K_4} + U_2 + \cdots + U_6$, where $\sigma^2(U_i) = \Theta(k^5)$ for $i = 2, 3$, and $\sigma^2(U_i) \sim \Theta(k^4)$ for $i = 4, 5, 6$ provided that (10) holds. We do not present formulas of various parts of the decomposition of $\mathcal{N}_{K_4}$ here because the notation becomes awkward.

One may expect that, for a large class of graphs G, the leading $U-$ statistics of the decomposition ($U-$statistics having largest variances) of $\mathcal{N}_G$ correspond to the components P_2 and K_3, i.e., they are U_2 and (in the case where $K_3 \subset G$) U_3.

3. Asymptotic normality. Recall that $\nu_1, \dots, \nu_N$ denotes a sequence of independent Bernoulli random variables with success probability $p = \mathbf{P}\{\nu_i = 1\} = 1 - \mathbf{P}\{\nu_i = 0\}$. We shall assume that this sequence and the random permutation $(X_1, \dots, X_N)$ are independent.

3.1. Let us write the characteristic function of the distribution of T in the Erdős-Rényi form using the decomposition (2). Denote

$$U_k^\star = \sum_{1 \le i_1 < \cdots < i_k \le N} g_k(X_{i_1}, \dots, X_{i_k})\nu_{i_1} \dots \nu_{i_k}.$$

Replacing the factors ν_i by $w_i = (\nu_i - p)$ in the formula of $U_k^\star$ we obtain the random variable U_k^*. Note that (3) implies $U_k^\star = U_k^*$ almost surely. Denote $S = w_1 + \cdots + w_N$.

The distribution of $T - \mathbf{E}\,T$ coincides with the conditional distribution of the sum

$$U_1^\star + \cdots + U_{n^*}^\star = U_1^* + \cdots + U_{n^*}^* =: T^*$$

given the event $\{S = 0\}$. Therefore, the characteristic function

$$\mathbf{E}\,\exp\{it(T - \mathbf{E}\,T)\} = \frac{1}{2\pi \mathbf{P}\{S = 0\}} \int_{-\pi}^{\pi} \mathbf{E}\,\exp\{itT^* + isS\} ds. \tag{11}$$

This way of representing the characteristic function of a linear statistic, like U_1, was used by Erdős and Rényi (1959).

3.2. Here we prove Proposition 1. Denote $f(t) = \mathbf{E}\,\exp\{it(T - \mathbf{E}\,T)/\sigma(T)\}$. In view of (6) it suffices to show that the relation $p^2q^2k^3 \to \infty$ implies for every $t \in R$ that $f'(t) + tf(t) \to 0$ as $k, n^* \to \infty$.

It follows from (5), (11) that

$$f(t) = \frac{1}{\lambda}\int_{-\pi}^{\pi} \mathbf{E}\,e^{iJ}ds, \qquad J = tH + sS, \qquad \lambda = 2\pi\mathbf{P}\{S = 0\},$$
$$H = \sum_{1\le i<j\le N} h_{ij}, \qquad h_{ij} = g_{ij}w_iw_j, \quad g_{ij} = \frac{1}{\sigma(T)}\,(L_{X_iX_j} - p_L).$$

Furthermore, by symmetry, we obtain

$$f'(t) = \frac{i}{\lambda}\int_{-\pi}^{\pi} \mathbf{E}\,He^{iJ}ds = i\,\frac{\binom{N}{2}}{\lambda}\int_{-\pi}^{\pi} \mathbf{E}\,h_{12}e^{iJ}ds. \tag{13}$$

Split $S = S_* + S_0$ and $H = h_{12} + H_* + H_0$ where $S_* = w_1 + w_2$ and

$$S_0 = \sum_{3\le j\le N} w_j, \qquad H_* = \sum_{3\le j\le N}(h_{1j} + h_{2j}), \qquad H_0 = \sum_{3\le i<j\le N} h_{ij}.$$

Expanding e^{iJ} in powers of ith_{12}, itH_* and isS_* we show that

$$\frac{\binom{N}{2}}{\lambda}\left|\int_{-\pi}^{\pi} \mathbf{E}\,h_{12}(e^{iJ} - ith_{12}e^{iJ_0})ds\right| \to 0, \tag{14}$$

where $J_0 = tH_0 + sS_0$. Therefore, we replace $\mathbf{E}\,h_{12}e^{iJ}$ by $itp^2q^2\mathbf{E}\,g_{12}^2e^{iJ_0}$ in the right integral (13).

Similarly, expanding e^{iJ} in powers of ith_{12}, itH_* and isS_* we show that

$$\frac{\binom{N}{2}}{\lambda}\left|p^2q^2\int_{-\pi}^{\pi} \mathbf{E}\,g_{12}^2(e^{iJ_0} - e^{iJ})ds\right| \to 0. \tag{15}$$

Using (14) and (15) we replace $\mathbf{E}\,h_{12}e^{iJ}$ by $itp^2q^2\mathbf{E}\,g_{12}^2e^{iJ}$ in the right integral (13). Furthermore, by symmetry, we can replace $\binom{N}{2}\mathbf{E}\,g_{12}^2e^{iJ}$ by $\mathbf{E}\,a^2e^{iJ} = a^2\mathbf{E}\,e^{iJ}$, since the number $a^2 := \sum_{1\le i<j\le N} g_{ij}^2$ is non-random. Finally, invoking (4) we obtain

$$a^2p^2q^2 = \frac{\mathbf{E}\,(L_{X_1X_2} - p_L)^2}{\sigma^2(T)}\binom{N}{2}p^2q^2 = 1 + O(1/Npq).$$

Since $O(1/Npq\lambda) = O(1/\sqrt{Npq})$ we can replace $p^2q^2a^2\mathbf{E}\,e^{iJ}$ by $\mathbf{E}\,e^{iJ}$. We have shown that

$$i\,\frac{\binom{N}{2}}{\lambda}\int_{-\pi}^{\pi}\mathbf{E}\,h_{12}e^{iJ}ds = -\,\frac{t}{\lambda}\int_{-\pi}^{\pi}\mathbf{E}\,e^{iJ}ds + o(1)$$

thus completing the proof.

The proof of (14) and (15) is rather technical and laborious. We do not present it here and refer to an extended version of the paper (Bloznelis 2004). Let us mention that in the proof of (14) and (15) we apply techniques developed in Bloznelis and Gözte (2002) for the analysis of the accuracy of the normal approximation of $U-$ statistics based on samples drawn without replacement, see also Bentkus, Götze and van Zwet (1997), Helmers and van Zwet (1982).

3.3. Note that the orthogonal decomposition (projection method) was used by Janson and Nowicki (1991) to prove limit theorems for subgraph count statistics of Bernoulli random graphs. The present paper can be considered as an attempt to extend these techniques to subgraph count statistics in the random graph model $G(k, n)$. In contrast to Bernoulli graph case the subgraph count statistics studied here have decompositions with vanishing linear part. Therefore, known results on the central limit theorem for asymptotically linear statistics based on samples drawn without replacement (see e.g., Bloznelis and Götze (2002), Zhao and Chen (1990)) are not applicable. We show that Erdős-Rényi (1959) representation (11) combined with Stein method (1) can be used to establish the asymptotic normality.

References

Bentkus, V., Götze, F. and van Zwet, W. R., *An Edgeworth expansion for symmetric statistics,*, Ann. Statist. **25** (1997), 851–896.

Bloznelis, M., *Some results on the orthogonal decomposition and asymptotic normality of subgraph count statistics*, Preprint 2004-10 Faculty of Mathematics and Informatics, Vilnius university (2004).

Bloznelis, M., *Orthogonal decomposition of symmetric functions defined on random permutations*, Combinatorics, Probability and Computing, Accepted for publication (2003).

Bloznelis, M. and Götze, F., *Orthogonal decomposition of finite population statistic and its applications to distributional asymptotics*, Ann. Statist. **29** (2001), 899–917.

Bloznelis, M. and Götze, F., *An Edgeworth expansion for symmetric finite population statistics*, Ann. Probab. **30** (2002), 1238–1265.

Erdős, P. and Rényi, A., *On the central limit theorem for samples from a finite population*, Publ. Math. Inst. Hungar. Acad. Sci. **4** (1959), 49–61.

Helmers, R., and van Zwet, W. R., *The Berry–Esseen bound for U-statistics.* **1** (1982), Statistical Decision Theory and Related Topics,III. Vol. 1. (S.S. Gupta and J.O. Berger, eds.), Academic Press, New York, 497–512.

Janson, S. (1990), *A functional limit theorem for random graphs with applications to subgraph count statistics*, Random. Struct. Algorithms **1**, 15–37.

Janson, S., Luczak, T., and Rucinski, A., *Random graphs* (2000), Wiley-Interscience, New York.

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

Ruciński, A. (1988), *When are small subgraphs of a random graph normally distributed?*, Probab. Theory Related Fields **78**, 1–10.

Stein, Ch., *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables* **2** (1970), Proc Sixth Berkeley Symp. Math. Stat. Probab., 583–602.

Tikhomirov, A.N., *On the rate of convergence in the central limit theorem for weak dependent variables*, Vestn. Leningr. Univ. (Mat. Mekh. Astron.) **7** (1976), 158-159.

Tikhomirov, A.N., *On the central limit theorem.*, Vestn. Syktyvkar. Univ., Ser. 1, Mat. Mekh. Inform. **4** (2001), 51–76.

Zhao, L. C. and Chen, X. R., *Normal approximation for finite-population U-statistics*, Acta Math. Appl. Sinica **6** (1990), 263–272.