Note on normal approximation for number of triangles in heterogeneous Erdős-Rényi graph

Logachov, A.V., Mogulskii, A.A. and Yambartsev, A.A.*

Abstract

We obtain an estimate of the convergence rate in the central limit theorem for the number of triangles in an heterogeneous Erdős-Rényi graphs. Our approach is reminiscent of Hoeffding decomposition (a common technique in the theory of U – statistics). We demonstrate that the centered and normalized number of triangles asymptotically behaves as well as the normalized sum of centered independent random variables, as the number of vertices of the graph increases. The proposed method is characterized by its simplicity and probabilistic intuition.

1 Introduction

The counting number of triangles and subgraphs in the random Erdős-Rényi graph $G(n,p)$ starts from the original works of Erdős and Rényi, and it is not a new research area. For the history of the central limit theorem (CLT) for the number of triangles (and subgraphs in general), we refer the reader to [1], where we found a good history introduction. This note contributes to this matter in the following points.

- We extend CLT to the case where the probability of an edge between two vertices can depend on vertices.

When the probability $p_{u,v}$ to connect two vertices u and v can depend on vertices, we call inhomogeneous Erdős-Rény random graph, $G(n,(p_{u,v}))$. We impose the strong condition (2.1) requiring a non-zero gap $\rho > 0$ between 0 and 1. However, even with this restrictive assumption, a large class of real-world networks can be effectively modeled by $G(n,(p_{u,v}))$. For example, a *stochastic block model* has such characteristics, [2]. At a first glance, it appears, that all inhomogeneous random graph models mentioned in [3] can be accommodated within our condition. It is worth noting that a term *inhomogeneous random graph* utilized in [3] differs somewhat from our usage. Therefore, we will refrain from referring to the random graph $G(n,(p_{u,v}))$ as an *inhomogeneous random graph*, opting instead for the term *heterogeneous Erdős-Rényi random graph* or simply $G(n,(p_{u,v}))$.

- We reduce the problem to the classic problem of the sum of independent random variables. It makes the proof easy and more "probabilistic".

*Logachev A.V. and Yambartsev A.A. thank FAPESP for the finantial support via the grant 2022/01030-0. Yambartsev A.A. thanks also FAPESP grant 2017/10555-0.

We show that in the simple decomposition for the sum of triangles, (2.3), the main contribution to the count of triangles is provided by the sum of independent random variables (the term $\eta_{3,n}$ in the decomposition (2.3)). This fact allows us to use well-developed machinery to deal with sums of independent random variables. In all earlier works, the results were obtained using combinatorics, the method of moments, and a detailed analysis of the characteristic function. All this complicates the proof, although it gives more accurate results [4].

Our proof is both simple and purely probabilistic, making it highly accessible to students who have completed a standard course in probability theory. It is notably concise, serving as an illustrative example of standard techniques for centering random variables and the application of Chebyshev and Berry-Essen inequalities.

It's worth noting that decomposition (2.3) can be viewed as a Hoeffding-type decomposition commonly used in U-statistics. Additionally, many graph statistics can be considered as incomplete U-statistics. This is another, statistics, way to establish the normal approximation. The validity of the Central Limit Theorem for these cases are well-established facts applied for homogeneous Erdős-Rényi (see [1, 5, 8]). Perhaps one of the earliest proofs of normal approximation for U-statistics can be attributed to the work of Wassily Hoeffding, as referenced in [18].

We establish convergence based on the Barry-Essen bound outlined in [6, p. 115, Theorem 6], which provides a convergence rate in the Kolmogorov distance of order $n^{-1/3}$. We will show (see section 3) that our approach can, unfortunately, improve this rate only to the form $n^{-\alpha}$, $\alpha < \frac{1}{2}$. While our paper does not primarily focus on the rate of convergence, it is worth noting that recent research has yielded more profound Barry-Essen-type bounds for the rate of convergence. For instance, references such as [15, 16, 17] offer improved rates of order n^{-1} applied for homogeneous Erdős-Rényi random graph. We believe extending these results to the case of heterogeneous random graphs presents a promising avenue for future research.

Consequently, we have established moderate deviation within a somewhat restricted area (see (2.13)). Recently, the precise asymptotic behavior of moderate deviation for the number of subgraphs in the homogeneous Erdős-Rényi random graph, $G(n, p)$, has been derived in [9].

It's also worth noting the significance of papers such as [10], [11], which explore the principle of large deviations for the number of subgraphs in the homogeneous Erdős-Rényi random graph, and [12]–[14], which encompass the Hoeffding-type inequalities for the number of subgraphs in random graphs of diverse types.

- Finally, we are sure that our approach can readily apply to the number of copies of some fixed arbitrary subgraph G in considered heterogeneous random graphs.

It is easy to see that the decomposition (2.3) works for any subgraph. Moreover, rough combinatorics shows that the sum of independent random variables provides the main asymptotic. We comment it in Remark 2.6.

In the next section we formulate and prove the main results. Throughout the paper, we consider all random elements on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and $\mathbf{E}, \mathbf{D}$ are expectation and variance with respect to the probability measure $\mathbf{P}$.

2 Main result

We define the *heterogeneous* Erdős-Rényi random graph with n vertices, $n \in \mathbb{N}$ by the following. Denote $[n]$, the set of vertices, $[n] = \{1, \dots, n\}$. Consider the family of independent random variables X_{ij}, $1 \le i < j \le n$ with Bernoulli distribution with success probability $p_{ij,n}$ (i.e. $\mathbf{P}(X_{ij} = 1) = p_{ij,n}$, $\mathbf{P}(X_{ij} = 0) = 1 - p_{ij,n}$). We assume that if $X_{ij} = 1$, then the vertices i and j are connected by an edge, and there is no edge, otherwise. In other words, X_{ij} is an indicator of an edge between vertex i and j. Note that $p_{ij,n}$ also depends on n, the total number of vertices. For this random graph we adopt the notation $G(n, (p_{ij,n}))$.

In this way, a heterogeneous Erdős-Rényi graph, $G(n, (p_{ij,n}))$, differs from a homogeneous Erdős-Rényi graph, $G(n, p)$, by the distributions of X_{ij}: in homogeneous case the all $p_{ij,n} \equiv p$ for all $1 \le i < j \le n$). Sometimes the following notations $X_{(ij),n}, p_{(ij),n}$ for random variables $X_{ij,n}$ and probabilities will be useful. We denote (ij) the pair of vertices i and j without obeying the order, for example, $X_{(ij),n}$ where $i < j$ and $X_{(ji),n}$ are the same variable $X_{(ij),n} \equiv X_{(ji),n}$, and for probabilities $p_{(ij),n} \equiv p_{(ji),n}$.

Throughout the paper, we require a (small) gap from 0 and 1 for probabilities $p_{ij,n}$: there exist $p_{\min}$ and $p_{\max}$ such that

$$0 < p_{\min} \le p_{ij,n} \le p_{\max} < 1, \tag{2.1}$$

for all $1 \le i < j \le n$, $n \in \mathbb{N}$. While this assumption may appear restrictive, it's worth noting that the most intriguing phenomena often emerge when connection probabilities decrease with n. But in practical terms, when dealing with finite networks, this assumption may not be as stringent. Nevertheless, we will also describe and discuss cases where the probabilities $p_{ij,n}$ or $1 - p_{ij,n}$ tend to zero, see Remark 2.7 and Remark 2.8.

Let T_n be the number of triangles in our heterogeneous Erdős-Rényi graph. Then

$$T_n := \sum_{1 \le i < j < k \le n} X_{ij} X_{jk} X_{ik}, \quad \mathbf{E}T_n = \sum_{1 \le i < j < k \le n} p_{ij,n} p_{jk,n} p_{ik,n}.$$

Let

$$Z(n) := \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij,n}(1 - p_{ij,n}) Q_{ij,n}^2,$$

where

$$Q_{ij,n} := \sum_{r=1}^{i-1} p_{ri,n} p_{rj,n} + \sum_{r=i+1}^{j-1} p_{ir,n} p_{rj,n} + \sum_{r=j+1}^{n} p_{ir,n} p_{jr,n}. \tag{2.2}$$

Here we obviously assume that $\sum\limits_{r=1}^{0} = \sum\limits_{r=i+1}^{i} = \sum\limits_{r=n+1}^{n} = 0$.

We are interested in the rate of convergence in CLT for the sequence

$$\eta_n := \frac{T_n - \mathbf{E}T_n}{\sqrt{Z(n)}}.$$

Let's formulate and prove the main result of the paper.

Theorem 2.1. *Let the condition (2.1) holds. Then for all $n \geq 3$*

$$\sup_{x\in\mathbb{R}} |\mathbf{P}(\eta_n < x) - \Phi(x)| \leq \frac{1}{n^{1/3}} \left(\frac{3^{7/2}A}{2\rho^7} + \frac{18}{\rho^6} + \frac{4}{\sqrt{2\pi}} \right),$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$, A is the Berry-Esseen constant, and $\rho := \min(p_{\min}, 1 - p_{\max})$.

Proof. Let write η_n in the following way.

$$\begin{aligned}
\eta_n =& \frac{\sum\limits_{1\leq i<j<k\leq n} (X_{ij} \pm p_{ij,n})(X_{jk} \pm p_{jk,n})(X_{ik} \pm p_{ik,n}) - \sum\limits_{1\leq i<j<k\leq n} p_{ij,n}p_{jk,n}p_{ik,n}}{\sqrt{Z(n)}} \\
=& \frac{\sum\limits_{1\leq i<j<k\leq n} (X_{ij} - p_{ij,n})(X_{jk} - p_{jk,n})(X_{ik} - p_{ik,n})}{\sqrt{Z(n)}} \\
&+ \frac{\sum\limits_{1\leq i<j\leq n} p_{ij,n} \sum\limits_{k\in[n]\setminus\{i,j\}} (X_{(ik)} - p_{(ik),n})(X_{(jk)} - p_{(jk),n})}{\sqrt{Z(n)}} \\
&+ \frac{\sum\limits_{1\leq i<j\leq n} (X_{ij} - p_{ij,n}) \sum\limits_{k\in[n]\setminus\{i,j\}} p_{(ik),n}p_{(jk),n}}{\sqrt{Z(n)}} =: \eta_{1,n} + \eta_{2,n} + \eta_{3,n}.
\end{aligned} \tag{2.3}$$

Note that in (ij) notation $Q_{ij,n} = \sum\limits_{k\in[n]\setminus\{i,j\}} p_{(ik),n}p_{(jk),n}$, and thus

$$\eta_{3,n} = \frac{\sum\limits_{i=1}^{n-1} \sum\limits_{j=i+1}^{n} (X_{ij} - p_{ij,n})Q_{ij,n}}{\sqrt{Z(n)}}.$$

Moreover, observe that $Z(n)$ is exactly the variance of the numerator of $\eta_{3,n}$, and the following inequality holds true for all $n \geq 3$

$$\begin{aligned}
\frac{n^4}{2} \geq Z(n) = \mathbf{D} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (X_{ij} - p_{ij,n})Q_{ij,n} &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij,n}(1 - p_{ij,n})Q_{ij,n}^2 \\
&\geq \frac{n(n-1)(n-2)^2\rho^6}{2} = \frac{n^4(1-1/n)(1-2/n)^2\rho^6}{2} \geq \frac{n^4\rho^6}{27}.
\end{aligned} \tag{2.4}$$

Using Chebyshev inequality, uncorrelatedness of terms in the sum, and using (2.4), we obtain that the "contribution" of $\eta_{1,n}$ and $\eta_{2,n}$ into η_n is negligible. Indeed,

$$\begin{aligned}
\mathbf{P}\left(|\eta_{1,n}| > \frac{1}{n^{1/3}} \right) &\leq \frac{n^{2/3}\mathbf{E}\left(\sum\limits_{1\leq i<j<k\leq n} (X_{ij} - p_{ij,n})(X_{jk} - p_{jk,n})(X_{ik} - p_{ik,n}) \right)^2}{Z(n)} \\
&= \frac{n^{2/3} \sum\limits_{1\leq i<j<k\leq n} p_{ij,n}(1 - p_{ij,n})p_{jk,n}(1 - p_{jk,n})p_{ik,n}(1 - p_{ik,n})}{Z(n)} \leq \frac{27n^{2/3}\binom{n}{3}}{n^4\rho^6} \leq \frac{9}{2\rho^6 n^{1/3}}.
\end{aligned} \tag{2.5}$$

In the same way, we will obtain the obtain bound for the $\eta_{2,n}$. But to make the calculation easier instead of $\eta_{2,n}$ we apply Chebyshev inequality to the following three terms which compose $\eta_{2,n}$:

$$\eta_{2,n} = \frac{\sum\limits_{1\le i<j<k\le n}(X_{ij}-p_{ij,n})(X_{jk}-p_{jk,n})p_{ik,n}}{\sqrt{Z(n)}} + \frac{\sum\limits_{1\le i<j<k\le n}(X_{ij}-p_{ij,n})p_{jk,n}(X_{ik}-p_{ik,n})}{\sqrt{Z(n)}} + \frac{\sum\limits_{1\le i<j<k\le n}p_{ij,n}(X_{jk}-p_{jk,n})(X_{ik}-p_{ik,n})}{\sqrt{Z(n)}} =: \eta_{2,n}^{(1)} + \eta_{2,n}^{(2)} + \eta_{2,n}^{(3)}.$$

Finally, for $r = 1, 2, 3$ we obtain

$$\mathbf{P}\left(|\eta_{2,n}^{(r)}| > \frac{1}{n^{1/3}}\right) \le \frac{27n^{2/3}\binom{n}{3}}{n^4\rho^6} \le \frac{9}{2\rho^6 n^{1/3}}. \tag{2.6}$$

For the rate of convergence in CLT for $\eta_{3,n}$ we use the following statement

Theorem 2.2. *[6, p. 115, Theorem 6] Let $Y_1, \dots, Y_m$ independent random variables with $\mathbf{E}Y_j = 0$, $\mathbf{E}|Y_j|^{2+\delta} < \infty$ for some $\delta \in (0, 1]$, $1 \le j \le m$. Then*

$$\sup_{x\in\mathbb{R}}\left|\mathbf{P}\left(\frac{\sum\limits_{j=1}^{m}Y_j}{\sqrt{B_m}} < x\right) - \Phi(x)\right| \le \frac{A}{B_m^{1+\frac{\delta}{2}}}\sum_{j=1}^{m}\mathbf{E}|Y_j|^{2+\delta},$$

where $B_m := \sum\limits_{j=1}^{m}\mathbf{E}Y_j^2$.

Using formulas (2.4) and inequality $Q_{ij,n} \le n$, we obtain for $\delta \in (0, 1]$

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbf{E}|(X_{ij}-p_{ij,n})Q_{ij,n}|^{2+\delta} \le \binom{n}{2}n^{2+\delta} \le \frac{n^{4+\delta}}{2}, \tag{2.7}$$

$$(Z(n))^{1+\delta/2} \ge \frac{n^{4+2\delta}\rho^{6+3\delta}}{3^{3+3\delta/2}}. \tag{2.8}$$

From (2.7), (2.8), and from Theorem 2.2, for $\delta = \frac{1}{3}$, $m = \binom{n}{2}$ it is follows that

$$\sup_{x\in\mathbb{R}}|\mathbf{P}(\eta_{3,n} < x) - \Phi(x)| \le A\frac{3^{7/2}n^{13/3}}{2n^{14/3}\rho^7} = A\frac{3^{7/2}}{2\rho^7 n^{1/3}}. \tag{2.9}$$

We finish the proof by constructing the upper and lower bounds for $\mathbf{P}(\eta_n < x) - \Phi(x)$. It is easy to see that for all $\delta > 0$, $x \in \mathbb{R}$

$$0 < \Phi(x+\delta) - \Phi(x) \le \frac{\delta}{\sqrt{2\pi}}. \tag{2.10}$$

Let denote $\tilde{\eta}_n := |\eta_{1,n}| + |\eta^{(1)}_{2,n}| + |\eta^{(2)}_{2,n}| + |\eta^{(3)}_{2,n}|$. Using inequalities (2.5), (2.6), (2.9), (2.10), we obtain for all $x \in \mathbb{R}$

$$\begin{aligned}\mathbf{P}(\eta_n < x) - \Phi(x) &\le \mathbf{P}\left(\eta_n < x, \tilde{\eta}_n \le \frac{4}{n^{1/3}}\right) + \mathbf{P}\left(\tilde{\eta}_n > \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\le \mathbf{P}\left(\eta_{3,n} < x + \frac{4}{n^{1/3}}\right) + \frac{18}{2\rho^6 n^{1/3}} \pm \Phi\left(x + \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\le A\frac{3^{7/2}}{2\rho^7 n^{1/3}} + \frac{18}{\rho^6 n^{1/3}} + \frac{4}{\sqrt{2\pi} n^{1/3}}.\end{aligned} \tag{2.11}$$

Applying (2.5), (2.6), (2.9), (2.10), and the inequality $\mathbf{P}(A \cap B) \ge \mathbf{P}(A) - \mathbf{P}(\overline{B})$, for all $x \in \mathbb{R}$ we obtain

$$\begin{aligned}\mathbf{P}(\eta_n < x) - \Phi(x) &\ge \mathbf{P}\left(\eta_n < x, \tilde{\eta}_n \le \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\ge \mathbf{P}\left(\eta_{3,n} < x - \frac{4}{n^{1/3}}\right) - \mathbf{P}\left(\tilde{\eta}_n > \frac{4}{n^{1/3}}\right) \pm \Phi\left(x - \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\ge \mathbf{P}\left(\eta_{3,n} < x - \frac{4}{n^{1/3}}\right) - \Phi\left(x - \frac{4}{n^{1/3}}\right) - \frac{36}{2\rho^6 n^{1/3}} - \frac{4}{\sqrt{2\pi} n^{1/3}} \\ &\ge -A\frac{3^{7/2}}{2\rho^7 n^{1/3}} - \frac{18}{\rho^6 n^{1/3}} - \frac{4}{\sqrt{2\pi} n^{1/3}}.\end{aligned} \tag{2.12}$$

From (2.11), (2.12) it follows that

$$\sup_{x\in\mathbb{R}} |\mathbf{P}(\eta_n < x) - \Phi(x)| \le \frac{1}{n^{1/3}}\left(A\frac{3^{7/2}}{2\rho^7} + \frac{18}{\rho^6} + \frac{4}{\sqrt{2\pi}}\right).$$

□

Corollary 2.3. *The proof of Theorem 2.1 implies that $\mathbf{D}T_n \sim Z(n)$ as n goes to infinity. Thus, CLT works with normalization $\sqrt{\mathbf{D}T_n}$.*

Consider the following family of numeric sequences

$$\mathcal{X} := \left\{x = x(n) : \lim_{n\to\infty} x(n) = \infty, \quad \lim_{n\to\infty} \frac{x(n)}{\sqrt{\ln n}} = 0\right\}. \tag{2.13}$$

Let

$$\xi_n := \frac{\eta_n}{x}, \quad x \in \mathcal{X}.$$

Corollary 2.4. *Theorem 2.1 implies the following moderate deviation principle for the sequence ξ_n. Let $x \in \mathcal{X}$ then, for any Borelian set $B \subset \mathbb{R}$*

$$\limsup_{n\to\infty} \frac{1}{x^2} \ln \mathbf{P}\left(\xi_n \in B\right) \le -\inf_{\alpha\in[B]} \frac{\alpha^2}{2};$$

$$\liminf_{n\to\infty} \frac{1}{x^2} \ln \mathbf{P}\left(\xi_n \in B\right) \ge -\inf_{\alpha\in(B)} \frac{\alpha^2}{2},$$

where $[B]$, (B) are the closure and interior of B correspondingly.

Proof. Theorem 2.1 implies that for any $\alpha \in \mathbb{R}$

$$\frac{1}{\sqrt{2\pi}}\int_{x(\alpha-\varepsilon)}^{x(\alpha+\varepsilon)} e^{-\frac{t^2}{2}}dt + \frac{2}{n^{1/3}}\left(\frac{3^{7/2}A}{2\rho^7}+\frac{18}{\rho^6}+\frac{4}{\sqrt{2\pi}}\right) \geq \mathbf{P}\left(\xi_n \in (\alpha)_\varepsilon\right)$$
$$\geq \frac{1}{\sqrt{2\pi}}\int_{x(\alpha-\varepsilon)}^{x(\alpha+\varepsilon)} e^{-\frac{t^2}{2}}dt - \frac{2}{n^{1/3}}\left(\frac{3^{7/2}A}{2\rho^7}+\frac{18}{\rho^6}+\frac{4}{\sqrt{2\pi}}\right).$$

Thus, denoting $C := 2\left(\frac{3^{7/2}A}{2\rho^7}+\frac{18}{\rho^6}+\frac{4}{\sqrt{2\pi}}\right)$ we have

$$\frac{2\varepsilon}{\sqrt{2\pi}}e^{-\frac{x^2(\max(|\alpha-\varepsilon|,|\alpha+\varepsilon|))^2}{2}} + \frac{C}{n^{1/3}} \geq \mathbf{P}\left(\xi_n \in (\alpha)_\varepsilon\right) \geq \frac{2\varepsilon}{\sqrt{2\pi}}e^{-\frac{x^2(\min(|\alpha-\varepsilon|,|\alpha+\varepsilon|))^2}{2}} - \frac{C}{n^{1/3}}. \tag{2.14}$$

Using the condition (2.13) and inequality (2.14), we obtain

$$\lim_{\varepsilon\to 0}\liminf_{n\to\infty}\frac{1}{x^2}\ln \mathbf{P}\left(\xi_n \in (\alpha)_\varepsilon\right) \geq \lim_{\varepsilon\to 0}\liminf_{n\to\infty}\frac{1}{x^2}\ln\left(\frac{\varepsilon}{\sqrt{2\pi}}e^{-\frac{x^2(\max(|\alpha-\varepsilon|,|\alpha+\varepsilon|))^2}{2}}\right) = -\frac{\alpha^2}{2},$$
$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{x^2}\ln \mathbf{P}\left(\xi_n \in (\alpha)_\varepsilon\right) \leq \lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{x^2}\ln\left(\frac{3\varepsilon}{\sqrt{2\pi}}e^{-\frac{x^2(\min(|\alpha-\varepsilon|,|\alpha+\varepsilon|))^2}{2}}\right) = -\frac{\alpha^2}{2}.$$

It means the moderate large deviation principle for the sequence ξ_n.

In order to prove the exponential tightness we need to show that for any $M > 0$ there exists $N_M < \infty$ such that

$$\limsup_{n\to\infty}\frac{1}{x^2(n)}\ln \mathbf{P}\left(|\xi_n| \geq N_M\right) \leq -M.$$

Theorem 2.1 and condition (2.13) imply that for any $N_M > 0$ and for sufficiently large n the following inequality holds

$$\mathbf{P}\left(|\xi_n| \geq N_M\right) \leq \frac{2}{\sqrt{2\pi}}\int_{xN_M}^{\infty} e^{-\frac{t^2}{2}}dt + \frac{C}{n^{1/3}} \leq \frac{3}{\sqrt{2\pi}}e^{-\frac{x^2(n)N_M^2}{2}}.$$

Thus, choosing $N_M := \sqrt{2M}$, we have

$$\limsup_{n\to\infty}\frac{1}{x^2}\ln \mathbf{P}\left(|\xi_n| \geq N_M\right) \leq \limsup_{n\to\infty}\frac{1}{x^2}\ln\left(\frac{3}{\sqrt{2\pi}}e^{-\frac{x^2N_M^2}{2}}\right) = -M.$$

It finishes the proof of exponential tightness for ξ_n. The local large deviation principle and exponential compactness imply the large deviation principle for ξ_n (see, for example, [7, Lemma 4.1.23]). □

Remark 2.5. *Note that the method does not allow us to improve the bound of the rate of convergence $O(1/n^{1/3})$. Indeed, if we want to improve the bound (2.5), then we need to consider*

$$\mathbf{P}\left(|\eta_{1,n}| > \frac{1}{n^\alpha}\right),$$

for $\alpha < 1/3$, then for any fixed x

$$\Phi\left(x+\frac{1}{n^\alpha}\right) - \Phi\left(x\right) = O\left(\frac{1}{n^\alpha}\right) \gg \frac{1}{n^{1/3}},$$

which will make the bound (2.10) worse.

Remark 2.6. *The method allows (with simple modifications) to prove similar statements for the number of any fixed subgraphs in heterogeneous Erdős-Rényi graph.*

Indeed, we saw that only the sum of independent variables $\eta_{3,n}$ contributed to the η_n. The same holds for any subgraph under the condition (2.1). Suppose a subgraph contains k vertices. Then the decomposition (2.3) will contain again the sums of random variables $\eta_{1,n} + \ldots + \eta_{k-1,n}$ each of them is represented by a sum of uncorrelated products of $(X_{ij} - p_{ij})$'s and the last sum $\eta_{k,n}$ is the sum of independent variables. The variance of all uncorrelated sums will be negligible with respect to the variance of the last sum of independent variables:

$$\lim_{n\to\infty} \frac{\mathbf{D}\eta_{1,n} + \ldots + \mathbf{D}\eta_{k-1,n}}{\mathbf{D}\eta_{k,n}} = 0.$$

Let us consider a weaker than (2.1) condition (2.15) in which the numbers $p_{\min} = p_{\min,n} > 0$, $p_{\max} = p_{\max,n} < 1$ can depend on the parameter $n \in \mathbb{N}$:

$$0 < p_{\min,n} \le p_{ij,n} \le p_{\max,n} < 1, \tag{2.15}$$

for all $1 \le i < j \le n$, $n \in \mathbb{N}$.

Remark 2.7. *Note that the condition (2.1) can be replaced by the weaker condition (2.15). Theorem 2.1 will hold if we require additional condition*

$$\lim_{n\to\infty} \frac{1}{n^{1/3}\rho^7} = 0,$$

which come from the "worst" bound (2.9). In this case the CLT holds for η_n with the rate of convergence $O\left(\frac{1}{n^{1/3}\rho^7}\right)$, as n goes to infinity.

Let us consider the following conditions:

$$\lim_{n\to\infty} \frac{\sum\limits_{1\le i<j<k\le n} \max(p_{ij,n}, 1-p_{ij,n})^2 \max(p_{jk,n}, 1-p_{jk,n})^2 \max(p_{ik,n}, 1-p_{ik,n})^2}{\sum\limits_{i=1}^{n-1}\sum\limits_{j=i+1}^{n} p_{ij,n}(1-p_{ij,n})Q_{ij,n}^2} = 0, \tag{2.16}$$

$$\lim_{n\to\infty} \frac{1}{Z(n)} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{E}\left(Q_{ij,n}^2 (X_{ij,n} - p_{ij,n})^2 \mathbf{I}\big(|Q_{ij,n}(X_{ij,n} - p_{ij,n})| > \varepsilon\sqrt{Z(n)}\big)\right) = 0, \tag{2.17}$$

here $\mathbf{I}(\cdot)$ is an indicator of the set, $Q_{ij,n}$ defined in (2.2).

Remark 2.8. *If we are interested in the CLT only (without estimating of the rate of convergence), then conditions (2.16), (2.17) are sufficient.*

Indeed, we used Chebyshev inequality to bound variability of the terms $\eta_{1,n}$ and $\eta_{2,n}$. To maintain the contribution of these terms negligible we need the condition (2.16). And, we can impose, for example, the Lindeberg's condition (2.17) for the CLT to hold for a sum of independent random variables (see, for example, [6]).

Writing conditions (2.16) and (2.17) in the homogeneous case, when $p_n = p_{ij,n}$ for all $i, j \in [n]$, we obtain the following conditions

$$np_n^2 \to \infty, \quad n^2(1-p_n) \to \infty.$$

Observe that these conditions coincide with that was obtained in [8]. Note that this condition is stronger than necessary and sufficient conditions obtained in [4]:

$$np_n \to \infty, \quad n^2(1-p_n) \to \infty.$$

3 An improvement of the convergence rate estimate of the Kolmogorov distance

We showed that the contribution of bounds like (2.5) is of order $n^{-1/3}$. Note that we used the second moments in these bounds. The only way to improve the rate is by utilizing the moments greater than two in inequality (2.5). Let us show that within this approach, for any $\alpha \in \left[\frac{1}{3}, \frac{1}{2}\right)$ the following bounds hold: there exists a positive constant C_α such that

$$\sup_{x\in\mathbb{R}} |\mathbf{P}(\eta_n < x) - \Phi(x)| < \frac{C_\alpha}{n^\alpha}. \tag{3.1}$$

When we expand the brackets in the expression

$$\left(\sum_{1\le i<j<k\le n} (X_{ij} - p_{ij,n})(X_{jk} - p_{jk,n})(X_{ik} - p_{ik,n}) \right)^{2r},$$

each term will be a product of the form

$$\Pi_l = \prod_{v_l=1}^{2r} (X_{i_{v_l}j_{v_l}} - p_{i_{v_l}j_{v_l},n})(X_{j_{v_l}k_{v_l}} - p_{j_{v_l}k_{v_l},n})(X_{i_{v_l}k_{v_l}} - p_{i_{v_l}k_{v_l},n}), \quad 1 \le l \le \binom{n}{3}^{2r},$$

where l denote a number of an ordered set of $2r$ triangles, and for each its factor

$$(X_{i_{v_l}j_{v_l}} - p_{i_{v_l}j_{v_l},n})(X_{j_{v_l}k_{v_l}} - p_{j_{v_l}k_{v_l},n})(X_{i_{v_l}k_{v_l}} - p_{i_{v_l}k_{v_l},n}), \tag{3.2}$$

we established a one-to-one correspondence with the triangle $(i_{v_l}, j_{v_l}, k_{v_l})$, and index v_l stands for the ordinal number of the triangle $(i_{v_l}, j_{v_l}, k_{v_l})$ in the ordered set of $2r$ triangles in term Π_l.

Note that $\mathbf{E}\Pi_l \neq 0$ if and only if each factor in (3.2) appears at least twice in the product Π_l. In simpler terms, every side of a triangle shares at least two triangles included in Π_l. It means that term Π_l with a non-zero expectation contains no more than $3r$ distinct vertices. Indeed, when each vertex appears at least twice among $2r$ triangles, the total number of vertices cannot exceed $3r$. This observation provides an upper bound for the number of arrangements of $2r$ triangles that yield a non-zero expectation for Π_l:

choosing $3r$ vertices gives us

$$\binom{n}{3r} < n^{3r};$$

the upper bound for the number of an ordered set of $2r$ triangles with chosen vertices is

$$\binom{3r}{3}^{2r} < (3r)^{6r};$$

Ultimately, the number of terms Π_l with a non-zero expectation does not exceed

$$n^{3r}(3r)^{6r}. \tag{3.3}$$

Denote $\overline{f}(r) := (3r)^{6r}$. Finally, utilizing the bound $|\mathbf{E}\Pi_l| \leq 1$ we obtain

$$\mathbf{E}\left(\sum_{1\leq i<j<k\leq n}(X_{ij}-p_{ij,n})(X_{jk}-p_{jk,n})(X_{ik}-p_{ik,n})\right)^{2r} \leq n^{3r}\overline{f}(r). \tag{3.4}$$

To ensure that the expectation aligns with the same order as the upper bound in (3.4), we will establish the lower bound for the expectation. Note that $2r$ triangles can cover each of the 3r sides and vertices exactly twice only if they are r pairs of coincident triangles. Let U and $|U|$ is the set and number of such configurations, respectively. It is easy to see, that

$$|U| = \binom{n}{3r}\frac{\binom{3r}{3}\binom{3r-3}{3}\dots\binom{6}{3}\binom{3}{3}}{r!}\binom{2r}{2}\binom{2r-2}{2}\dots\binom{4}{2}\binom{2}{2} = \frac{n(n-1)...(n-3r+1)(2r)!}{12^r r!}. \tag{3.5}$$

If $\Pi_l \notin U$ and $\mathbf{E}\Pi_l \neq 0$ then Π_l is constructed by at most $3r-3$ vertices. The number of such Π_l has the following upper bound

$$|\{\Pi_l : \Pi_l \notin U, \mathbf{E}\Pi_l \neq 0\}| \leq n^{3r-3}(3r-3)^{6r} \leq n^{3r-3}\overline{f}(r). \tag{3.6}$$

Using (3.5) and (3.6), we can conclude that there exists $\underline{f}(r,\rho) > 0$ such that for sufficiently large n

$$n^{3r}\underline{f}(r,\rho) \leq \mathbf{E}\left(\sum_{1\leq i<j<k\leq n}(X_{ij}-p_{ij,n})(X_{jk}-p_{jk,n})(X_{ik}-p_{ik,n})\right)^{2r} \leq n^{3r}\overline{f}(r). \tag{3.7}$$

The last means that for a given r and for sufficiently large n there exists constant $C_r > 0$ such that

$$\mathbf{E}\left(\sum_{1\leq i<j<k\leq n}(X_{ij}-p_{ij,n})(X_{jk}-p_{jk,n})(X_{ik}-p_{ik,n})\right)^{2r} \sim C_r n^{3r},$$

as $n \to \infty$.

Utilizing Markov inequality for $2r$-th moment and relation (3.7) we obtain

$$\mathbf{P}\left(|\eta_{1,n}| > \frac{1}{n^\alpha}\right) \leq \frac{\overline{f}(r)n^{3r}n^{2\alpha r}}{Z^r(n)} \leq \frac{27^r\overline{f}(r)n^{2\alpha r}}{n^r\rho^{6r}}. \tag{3.8}$$

In a similar way, one can obtain an estimate of the same order for $|\eta_{2,n}|$.

We can now determine the optimal α by solving the equation

$$\frac{n^{2\alpha r}}{n^r} = \frac{1}{n^\alpha}, \quad \alpha = \frac{r}{2r+1}. \tag{3.9}$$

From (3.8), (3.9) it follows that for any $\alpha \in \left[\frac{1}{3}, \frac{1}{2}\right)$ and choosing $1 \le r < \infty$ we can obtain upper bound (3.1). Thus, we have proved the following theorem.

Theorem 3.1. *Let the condition (2.1) holds. Then for any $\alpha \in \left[\frac{1}{3}, \frac{1}{2}\right)$ there exists a constant $C_\alpha > 0$ such that inequality (3.1) is met for all $n \ge 3$.*

References

[1] Gilmer, J., Kopparty, S. (2016). A local central limit theorem for triangles in a random graph. Random Structures & Algorithms, 48(4), 732-750.

[2] Lee, C., Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. Applied Network Science, 4(1), 1-50.

[3] Bollobás, B., Janson, S., Riordan, O. (2007). The phase transition in inhomogeneous random graphs. Random Structures & Algorithms, 31(1), 3-122.

[4] Ruciński, A. (1988). When are small subgraphs of a random graph normally distributed? Probability Theory and Related Fields, 78(1), 1-10.

[5] Sah, A., Sawhney, M. (2022). Local limit theorems for subgraph counts. Journal of the London Mathematical Society, 105(2), 950-1011.

[6] Petrov, V.V. Sum of independent random variables, Springer, 1975.

[7] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, New York, Springer, 1998.

[8] Nowicki, K., Wierman, J. C. (1988). Subgraph counts in random graphs using incomplete U-statistics methods. Discrete Mathematics, 72(1-3), 299-310.

[9] Goldschmidt, C., Griffiths, S., Scott, A. (2020). Moderate deviations of subgraph counts in the Erdős-Rényi random graphs $G(n, m)$ and $G(n, p)$. Transactions of the American Mathematical Society, 373(8), 5517-5585.

[10] Chatterjee, S., Varadhan, S.R.S. (2011). The large deviation principle for the Erdős–Rényi random graph. European J. Combin., 32(7), 1000-1017.

[11] Logachov, A.V., Mogulskii, A.A. (2020). Exponential Chebyshev Inequalities for Random Graphons and Their Applications. Sib Math J 61(4), 697-714.

[12] Bystrov, A.A., Volodko, N.V., (2022). Exponential Inequalities for the Distribution Tails of the Number of Cycles in the Erdős-Rényi Random Graphs. Siberian Advances in Mathematics, 2022, 32(2), 87-93.

[13] Bystrov, A.A., Volodko, N.V., (2023). Exponential inequalities for the number of subgraphs in the Erdős-Rényi random graph. Statistics and Probability Letters, 2023, 195, 109763.

[14] Bystrov, A.A., Volodko, N.V., (2023). Exponential Inequalities for the Tail Probabilities of the Number of Cycles in Generalized Random Graphs. Siberian Advances in Mathematics, 2023, 33(3), 181-189.

[15] Eichelsbacher, P., and Rednoß, B. (2023). Kolmogorov bounds for decomposable random variables and subgraph counting by the Stein–Tikhomirov method. Bernoulli, 29(3), 1821-1848.

[16] Privault, N., and Serafin, G. (2020). Normal approximation for sums of weighted U-statistics–application to Kolmogorov bounds in random subgraph counting. Bernoulli, 26(1), 2020, 587-615.

[17] Zhang, Z. S. (2022). Berry–Esseen bounds for generalized U-statistics. Electronic Journal of Probability, 27, 1-36.

[18] Hoeffding, W. (1992). A class of statistics with asymptotically normal distribution. Breakthroughs in Statistics: Foundations and Basic Theory, 308-334.