S©MR

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ
Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru
ISSN 1813-3304

Vol. 21, No. 2, pp. 914–926 (2024)
https://doi.org/10.33048/semi.2024.21.060

УДК 519.21
MSC 05C80, 60F05, 60F10

NOTE ON NORMAL APPROXIMATION FOR NUMBER OF TRIANGLES IN HETEROGENEOUS ERDŐS-RÉNYI GRAPH

A.V. LOGACHOV, **A.A. MOGULSKII**, **A.A. YAMBARTSEV**

Communicated by N.S. Arkashov

Abstract: We obtain a bound for the convergence rate in the central limit theorem for the number of triangles in a heterogeneous Erdős-Rényi graphs. Our approach is reminiscent of Hoeffding decomposition (a common technique in the theory of U-statistics). We show that the centered and normalized number of triangles asymptotically behaves as the normalized sum of centered independent random variables when the number of vertices increases. The proposed method is simple and intuitive.

Keywords: Erdős-Rényi random graphs, central limit theorem, large deviations principle.

Logachov, A.V., Mogulskii, A.A., Yambartsev, A.A. Note on normal approximation for number of triangles in heterogeneous Erdos-Renyi graph.

Logachov A.V. thanks Mathematical Center in Akademgorodok under the agreement N. 075-15-2022-281 with the Ministry of Science and Higher Education of the Russian Federation; Mogulskii A.A. is supported by the Ministry of Science and Higher Education of the Russian Federation FWNF-2022-0010; Logachov A.V., Yambartsev A.A. thanks FAPESP grant 2022/01030-0; Yambartsev A.A. thanks FAPESP grant 2017/10555-0.
Received March, 7, 2024, Published November, 1, 2024.

1 Introduction

The study of counting the number of triangles 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], which provides a good historical 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.

A natural generalization of the "classical" random graph $G(n,p)$ can be obtained by replacing the probability p by a symmetric $n \times n$ matrix (p_{ij}) with $0 \le p_{ij} \le 1$. We write $G(n,(p_{ij}))$ for the random graph with the set of vertices $[n] := \{1,\dots,n\}$, where vertices i and j are connected with probability p_{ij}. We will refer to the graph $G(n,(p_{ij}))$ as a heterogeneous Erdős-Rényi random graph. Although the term *inhomogeneous random graph* was used in [3], it differs somewhat from our usage. Therefore, to avoid confusion, we will refrain from using the term *inhomogeneous random graph* for $G(n,(p_{ij}))$ and instead use *heterogeneous Erdős-Rényi random graph* or simply $G(n,(p_{ij}))$.

In this paper, we impose a strong condition on connection probabilities, requiring a non-zero gap $\varepsilon > 0$ between 0 and 1: $\varepsilon \le p_{ij} \le 1-\varepsilon$ for any $i,j \in [n]$, see (1). However, even with this restrictive assumption, a large class of real-world networks can be effectively modeled by $G(n,(p_{ij}))$. For example, a *stochastic block model* has such characteristics [2]. Moreover, at first glance, all inhomogeneous random graph models mentioned in [3] can be accommodated within our condition.

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

We show that in the simple decomposition for the number of triangles, (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 (3)). This fact enables us to utilize well-established techniques for handling sums of independent random variables. Previous works relied on combinatorics, the method of moments, and detailed analyses of characteristic functions to obtain results. While these approaches provide more precise results [4], they also complicate the proofs.

Our proof is simple and purely probabilistic, making it highly accessible to students who have completed a standard course in probability theory. It is notably concise and provides a clear example of standard techniques for centering random variables and applying Chebyshev's and Berry-Esseen's inequalities.

It's worth noting that decomposition (3) can be interpreted as a Hoeffding-type decomposition commonly used in U-statistics. Additionally, many graph statistics can be viewed as incomplete U-statistics. This provides an alternative "statistical"approach to establishing the normal approximation. The validity of the Central Limit Theorem for these cases is a well-established fact, applied to homogeneous Erdős-Rényi graphs (see [1, 5, 8]). 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 using the Berry-Esseen bound outlined in [6, p. 115, Theorem 6]. This bound provides a convergence rate of order $n^{-1/3}$ in the Kolmogorov distance. In Section 3, we demonstrate that our approach only improves this rate to the order $n^{-\alpha}$, for $\alpha < \frac{1}{2}$. While our paper does not primarily focus on the rate of convergence, it is worth noting that recent research has produced more profound Berry-Esseen-type bounds for the rate of convergence. References such as [15]–[17] offer improved rates of order n^{-1} applied to homogeneous Erdős-Rényi random graphs. We believe that extending these results to the case of heterogeneous random graphs presents a promising avenue for future research.

As a result, we have established moderate deviation within a somewhat restricted area (see (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 various types.

- Finally, we are confident that our approach can readily be applied to counting the number of copies of an arbitrary fixed subgraph G in the considered heterogeneous random graphs.

It is easy to see that the decomposition (3) works for any subgraph. Moreover, basic combinatorial analysis reveals that the sum of independent random variables provides the main asymptotic. We discuss this further in Remark 2.

In the next section, we formulate and prove the main result. Throughout the paper, we consider all random elements on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and $\mathbf{E}, \mathbf{D}$ denote 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, where $n \in \mathbb{N}$, as follows. Denote $[n]$ as the set of vertices, $[n] = 1, \dots, n$. Consider the family of independent random variables X_{ij}, where $1 \le i < j \le n$, with a Bernoulli distribution having a success probability $p_{ij,n}$, i.e., $\mathbf{P}(X_{ij} = 1) = p_{ij,n}$ and $\mathbf{P}(X_{ij} = 0) = 1 - p_{ij,n}$. 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} indicates the presence of an edge between vertices i and j. Note that $p_{ij,n}$ also depends on the total number of vertices n. 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 the homogeneous case, all $p_{ij,n}$ are equal to p for all $1 \le i < j \le n$. Sometimes, the following notations $X_{(ij)}$ and $p_{(ij),n}$ for random variables $X_{ij,n}$ and probabilities will be useful. We denote (ij) as the pair of vertices i and j without obeying the order. For example, $X_{(ij)}$, where $i<j$, and $X_{(ji)}$ are the same variable, $X_{(ij)} \equiv X_{(ji)}$, 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

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

for all $1 \le i < j \le n$, $n \in \mathbb{N}$. This assumption may seem very restrictive. Indeed, the most intriguing phenomena often emerge when connection probabilities decrease with n. However, in practice, this assumption may not be as stringent when dealing with finite networks. Nevertheless, in Remark 3 and Remark 4, we discuss cases when the probabilities $p_{ij,n}$ or $1-p_{ij,n}$ tend to zero.

Let T_n be the number of triangles in $G(n,(p_{ij,n}))$. Then

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

Denote

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

where

$$Q_{ij,n} := \sum_{r=1}^{i-1} p_{ri,n}p_{rj,n} + \sum_{r=i+1}^{j-1} p_{ir,n}p_{rj,n} + \sum_{r=j+1}^{n} p_{ir,n}p_{jr,n}. \tag{2}$$

In (2) we obviously assume that $\sum_{r=1}^{0} = \sum_{r=i+1}^{i} = \sum_{r=n+1}^{n} = 0$.

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 us now formulate and prove the main result of the paper.

Theorem 1. *Suppose condition (1) holds. Then for all $n \ge 3$,*

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

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

Proof. Let rewrite η_n in the following way

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

Note that in (ij)-notations we can rewrite (2) as $Q_{ij,n} = \sum\limits_{k\in[n]\setminus\{i,j\}} p_{(ik),n}p_{(jk),n}$, 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 $Z(n)$ exactly represents the variance of the numerator of $\eta_{3,n}$, and the following inequalities hold for all $n \ge 3$:

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

By applying Chebyshev's inequality and considering the uncorrelatedness of terms in the sum, along with using (4), we can conclude that the contribution of $\eta_{1,n}$ and $\eta_{2,n}$ to η_n is negligible. Indeed,

$$\mathbf{P}\left(|\eta_{1,n}| > \frac{1}{n^{1/3}}\right) \le \frac{n^{2/3}\mathbf{E}\left(\sum\limits_{1\le i<j<k\le n} (X_{ij} - p_{ij,n})(X_{jk} - p_{jk,n})(X_{ik} - p_{ik,n})\right)^2}{Z(n)}$$
$$= \frac{n^{2/3} \sum\limits_{1\le i<j<k\le n} p_{ij,n}(1 - p_{ij,n})p_{jk,n}(1 - p_{jk,n})p_{ik,n}(1 - p_{ik,n})}{Z(n)}$$
$$\le \frac{27n^{2/3}\binom{n}{3}}{n^4\rho^6} \le \frac{9}{2\rho^6 n^{1/3}}. \tag{5}$$

We will similarly derive a bound for $\eta_{2,n}$. To simplify the calculation, instead of considering $\eta_{2,n}$ directly, we apply Chebyshev's inequality to the following three terms that 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, we obtain

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

We utilize the following theorem to obtain the convergence rate in the CLT for $\eta_{3,n}$.

Theorem 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 (4) and inequality $Q_{ij,n}\le n$, we obtain for $\delta\in(0,1]$,

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbf{E}|(X_{ij}-p_{ij,n})Q_{ij,n}|^{2+\delta}\le\binom{n}{2}n^{2+\delta}\le\frac{n^{4+\delta}}{2}, \tag{7}$$

$$(Z(n))^{1+\delta/2}\ge\frac{n^{4+2\delta}\rho^{6+3\delta}}{3^{3+3\delta/2}}. \tag{8}$$

From (7), (8), and by Theorem 2 with $\delta=\frac{1}{3}$ and $m=\binom{n}{2}$ it follows that

$$\sup_{x\in\mathbb{R}}|\mathbf{P}(\eta_{3,n}<x)-\Phi(x)|\le A\frac{3^{7/2}n^{13/3}}{2n^{14/3}\rho^7}=A\frac{3^{7/2}}{2\rho^7n^{1/3}}. \tag{9}$$

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 $\delta>0$, $x\in\mathbb{R}$,

$$0<\Phi(x+\delta)-\Phi(x)\le\frac{\delta}{\sqrt{2\pi}}. \tag{10}$$

Let denote $\tilde{\eta}_n := |\eta_{1,n}| + |\eta^{(1)}_{2,n}| + |\eta^{(2)}_{2,n}| + |\eta^{(3)}_{2,n}|$. Using inequalities (5), (6), (9) and (10), we obtain for all $x \in \mathbb{R}$

$$\begin{aligned}\mathbf{P}(\eta_n < x) - \Phi(x) &\le \mathbf{P}\left(\eta_n < x, \tilde{\eta}_n \le \frac{4}{n^{1/3}}\right) + \mathbf{P}\left(\tilde{\eta}_n > \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\le \mathbf{P}\left(\eta_{3,n} < x + \frac{4}{n^{1/3}}\right) + \frac{18}{2\rho^6 n^{1/3}} \pm \Phi\left(x + \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\le A\frac{3^{7/2}}{2\rho^7 n^{1/3}} + \frac{18}{\rho^6 n^{1/3}} + \frac{4}{\sqrt{2\pi} n^{1/3}}. \end{aligned} \tag{11}$$

Applying (5), (6), (9), (10), together with the inequality $\mathbf{P}(A \cap B) \ge \mathbf{P}(A) - \mathbf{P}(\overline{B})$, we obtain

$$\begin{aligned}\mathbf{P}(\eta_n < x) - \Phi(x) &\ge \mathbf{P}\left(\eta_n < x, \tilde{\eta}_n \le \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\ge \mathbf{P}\left(\eta_{3,n} < x - \frac{4}{n^{1/3}}\right) - \mathbf{P}\left(\tilde{\eta}_n > \frac{4}{n^{1/3}}\right) \pm \Phi\left(x - \frac{4}{n^{1/3}}\right) - \Phi(x) \\ &\ge \mathbf{P}\left(\eta_{3,n} < x - \frac{4}{n^{1/3}}\right) - \Phi\left(x - \frac{4}{n^{1/3}}\right) - \frac{36}{2\rho^6 n^{1/3}} - \frac{4}{\sqrt{2\pi} n^{1/3}} \\ &\ge -A\frac{3^{7/2}}{2\rho^7 n^{1/3}} - \frac{18}{\rho^6 n^{1/3}} - \frac{4}{\sqrt{2\pi} n^{1/3}}, \end{aligned} \tag{12}$$

for all $x \in \mathbb{R}$. From (11) and (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 1. *The proof of Theorem 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

$$\mathcal{X} := \left\{ x = x(n) : \lim_{n\to\infty} x(n) = \infty, \quad \lim_{n\to\infty} \frac{x(n)}{\sqrt{\ln n}} = 0 \right\}. \tag{13}$$

Let

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

Corollary 2. *Theorem 1 implies the following moderate deviation principle for the sequence* ξ_n. *Let* $x \in \mathcal{X}$ *then, for any Borel 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 *respectively.*

Proof. Theorem 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

$$\frac{2\varepsilon}{\sqrt{2\pi}}e^{-\frac{x^2(\max(|\alpha-\varepsilon|,|\alpha+\varepsilon|))^2}{2}} + \frac{C}{n^{1/3}} \geq \mathbf{P}\left(\xi_n \in (\alpha)_\varepsilon\right)$$
$$\geq \frac{2\varepsilon}{\sqrt{2\pi}}e^{-\frac{x^2(\min(|\alpha-\varepsilon|,|\alpha+\varepsilon|))^2}{2}} - \frac{C}{n^{1/3}}. \tag{14}$$

Using the condition (13) and inequality (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 establishes the moderate large deviation principle for the sequence ξ_n.

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 1 and condition (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 tightness imply the large deviation principle for ξ_n (see, for example, [7, Lemma 4.1.23]). □

Remark 1. *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 (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

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

which will make the bound (10) worse.

Remark 2. *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 (1). Suppose a subgraph contains k vertices. Then the decomposition (3) will include 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 compared to the variance of the final sum of independent variables:

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

We now examine the following condition (15), weaker than (1), where the parameters $p_{\min} = p_{\min,n} > 0$ and $p_{\max} = p_{\max,n} < 1$ are allowed to vary with the parameter $n \in \mathbb{N}$:

$$0 < p_{\min,n} \le p_{ij,n} \le p_{\max,n} < 1, \tag{15}$$

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

Remark 3. *Note that in Theorem 1 condition (1) can be replaced by weaker condition (15). Theorem 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 (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.

In the next remark, we will discuss the following conditions:

$$\lim_{n\to\infty} \frac{\sum\limits_{1\le i<j<k\le n} \max(p_{ij,n}, 1-p_{ij,n})^2 \max(p_{jk,n}, 1-p_{jk,n})^2 \max(p_{ik,n}, 1-p_{ik,n})^2}{\sum\limits_{i=1}^{n-1}\sum\limits_{j=i+1}^{n} p_{ij,n}(1-p_{ij,n})Q_{ij,n}^2} = 0, \tag{16}$$

$$\lim_{n\to\infty}\frac{1}{Z(n)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbf{E}\left(Q_{ij,n}^2(X_{ij,n}-p_{ij,n})^2\mathbf{I}\big(|Q_{ij,n}(X_{ij,n}-p_{ij,n})|>\varepsilon\sqrt{Z(n)}\big)\right)=0, \tag{17}$$

where $\mathbf{I}(\cdot)$ is an indicator of the set, and $Q_{ij,n}$ was defined by (2).

Remark 4. *Note that if we are interested in the CLT without estimating the rate of convergence, then conditions (16) and (17) are sufficient. Indeed, we used Chebyshev inequality to bound the variability of terms $\eta_{1,n}$ and $\eta_{2,n}$, and condition (16) will maintain the contribution of these terms negligible. For the CLT to hold for a sum of independent random variables, we can, for instance, impose Lindeberg's condition (17). (see, for example, [6]).*

Rewriting (16) and (17) for the homogeneous case, where $p_n = p_{ij,n}$ for all $i, j \in [n]$, we obtain the following condition for the CLT to hold:

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

Observe that this condition coincides 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 Improving the convergence rate

We demonstrated that the contribution of bounds like (5) is of the order $n^{-1/3}$. Note that we utilized the second moments in these bounds. The only way to improve this rate is by employing moments greater than two in inequality (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

$$\sup_{x\in\mathbb{R}}|\mathbf{P}(\eta_n<x)-\Phi(x)|<\frac{C_\alpha}{n^\alpha}. \tag{18}$$

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

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

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

$$(X_{i_{v_l}j_{v_l}}-p_{i_{v_l}j_{v_l},n})(X_{j_{v_l}k_{v_l}}-p_{j_{v_l}k_{v_l},n})(X_{i_{v_l}k_{v_l}}-p_{i_{v_l}k_{v_l},n}), \tag{19}$$

we established a one-to-one correspondence with the triangle $(i_{v_l}, j_{v_l}, k_{v_l})$, where 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 (19) appears at least twice in the product Π_l. In other words, every side of a triangle shares at least two triangles included in Π_l. This implies that the 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 provides us with the bound

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

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

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

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

$$n^{3r}(3r)^{6r}. \tag{20}$$

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

$$\mathbf{E}\left(\sum_{1\leq i<j<k\leq n}(X_{ij}-p_{ij,n})(X_{jk}-p_{jk,n})(X_{ik}-p_{ik,n})\right)^{2r} \leq n^{3r}\overline{f}(r). \tag{21}$$

Now, we will show that a lower bound for the expectation has the same order as the upper bound (21). Note that $2r$ triangles can cover each of the $3r$ sides and vertices precisely twice when they form r pairs of coincident triangles. Let U and $|U|$ be the set and number of such configurations, respectively. It is easy to see that

$$\begin{aligned}|U| &= \binom{n}{3r}\frac{\binom{3r}{3}\binom{3r-3}{3}\dots\binom{6}{3}\binom{3}{3}}{r!}\binom{2r}{2}\binom{2r-2}{2}\dots\binom{4}{2}\binom{2}{2}\\ &= \frac{n(n-1)...(n-3r+1)(2r)!}{12^r r!}.\end{aligned} \tag{22}$$

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

$$|\{\Pi_l : \Pi_l \notin U, \mathbf{E}\Pi_l \neq 0\}| \leq n^{3r-3}(3r-3)^{6r} \leq n^{3r-3}\overline{f}(r). \tag{23}$$

Using (22) and (23), we can conclude that there exists $\underline{f}(r,\rho) > 0$ such that for sufficiently large n

$$\begin{aligned}n^{3r}\underline{f}(r,\rho) \leq \mathbf{E}&\left(\sum_{1\leq i<j<k\leq n}(X_{ij}-p_{ij,n})(X_{jk}-p_{jk,n})(X_{ik}-p_{ik,n})\right)^{2r}\\ &\leq n^{3r}\overline{f}(r).\end{aligned} \tag{24}$$

Utilizing (24) and Markov inequality for $2r$-th moment, we obtain

$$\mathbf{P}\left(|\eta_{1,n}| > \frac{1}{n^\alpha}\right) \le \frac{\overline{f}(r)n^{3r}n^{2\alpha r}}{Z^r(n)} \le \frac{27^r\overline{f}(r)n^{2\alpha r}}{n^r\rho^{6r}}. \tag{25}$$

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

We can now determine the optimal α by solving the equation

$$\frac{n^{2\alpha r}}{n^r} = \frac{1}{n^\alpha}, \quad \alpha = \frac{r}{2r+1}. \tag{26}$$

From (25), (26) 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 (18). Thus, we have proved the following theorem.

Theorem 3. *Suppose condition (1) is satisfied. Then, for any $\alpha \in \left[\frac{1}{3}, \frac{1}{2}\right)$, there exists a constant $C_\alpha > 0$ such that inequality (18) holds for all $n \ge 3$.*

Note that from the lower bound in inequality (24), it follows that any upper bound of the $2r$-th moment of the number of triangles will have an order of n^{3r} with respect to n, i.e., the same as in the proposed upper bound. Consequently, even if we improve the upper bound, this will only change the constant C_α in Theorem 3, while the value of the parameter α will remain the same.

Theorem 3 provides a more accurate bound for the convergence rate than Theorem 1. Despite this fact, we consider Theorem 1 to be the main result of this work. Our primary goal was to demonstrate how to prove the CLT for the number of triangles using a purely probabilistic approach, without any combinatorics. Theorem 1 and its proof achieve this objective. It is worth emphasizing that the terms in the proposed decomposition are uncorrelated. It is significantly used in the proof. Moreover, many papers prove the CLT without estimating the convergence rate. It is evident that for these purposes, the method proposed in Theorem 1 is preferable.

We thank the anonymous reviewer whose comments and suggestions helped improve and clarify this manuscript.

References

[1] J. Gilmer, S. Kopparty, *A local central limit theorem for triangles in a random graph*, Random Struct. Algorithms, **48**:4 (2016), 732–750. Zbl 1343.05136

[2] C. Lee, D.J. Wilkinson, A review of stochastic block models and extensions for graph clustering, Appl. Netw. Sci. , **4** (2019), Article ID 122.

[3] B. Bollobàs, S. Janson, O. Riordan, The phase transition in inhomogeneous random graphs, Random Struct. Algorithms, **31**:1 (2007), 3–122. Zbl 1123.05083

[4] A. Ruciński, When are small subgraphs of a random graph normally distributed?, Probab. Theory Relat. Fields, **78**:1 (1988), 1–10. Zbl 0627.60045

[5] A. Sah, M. Sawhney, Local limit theorems for subgraph counts, J. Lond. Math. Soc., II. Ser., **105**:2 (2022), 950–1011. Zbl 1519.05223

[6] V.V. Petrov, Sum of independent random variables, Springer, Berlin etc., 1975. Zbl 0322.60042

[7] A. Dembo, O. Zeitouni, Large deviations techniques and applications, Springer, New York, 1998. Zbl 0896.60013
[8] K. Nowicki, J.C. Wierman, Subgraph counts in random graphs using incomplete U-statistics methods, Discrete Math., **72**:1-3 (1988), 299–310. Zbl 0672.05072
[9] C. Goldschmidt, S. Griffiths, A. Scott, Moderate deviations of subgraph counts in the Erdős-Rényi random graphs $G(n,m)$ and $G(n,p)$, Trans. Am. Math. Soc., **373**:8 (2020), 5517–5585. Zbl 1443.05172
[10] S. Chatterjee, S.R.S. Varadhan, The large deviation principle for the Erdős-Rényi random graph, Eur. J. Comb., **32**:7 (2011), 1000–1017. Zbl 1230.05259
[11] A.V. Logachov, A.A. Mogulskii, Exponential Chebyshev inequalities for random graphons and their applications, Sib. Math. J., **61**:4 (2020), 697–714. Zbl 1448.05180
[12] A.A. Bystrov, N.V. Volod'ko, *Exponential inequalities for the distribution of the number of cycles in the Erdős-Rényi random graphs*, Sib. Adv. Math., **32**:2 (2022), 87–93. Zbl 1523.05040
[13] A.A. Bystrov, N.V. Volodko, Exponential inequalities for the number of subgraphs in the Erdős-Rényi random graph, Stat. Probab. Lett., **195** (2023), Article ID 109763. Zbl 1518.05168
[14] A.A. Bystrov, N.V. Volodko, Exponential inequalities for the tail probabilities of the number of cycles in generalized random graphs, Sib. Adv. Math., **33**:3 (2023), 181-189.
[15] P. Eichelsbacher, B. Rednoß, Kolmogorov bounds for decomposable random variables and subgraph counting by the Stein-Tikhomirov method, Bernoulli, **29**:3 (2023), 1821–1848. Zbl 1515.05159
[16] N. Privault, G. Serafin, Normal approximation for sums of weighted U-statistics–application to Kolmogorov bounds in random subgraph counting. Bernoulli, **26**:1 (2020), 587–615. Zbl 1464.60021
[17] Z.-S. Zhang, Berry-Esseen bounds for generalized U-statistics, Electron. J. Probab., **27** (2022), Paper No. 134. Zbl 1511.60046
[18] W. Hoeffding, A class of statistics with asymptotically normal distribution, In: Kotz, S., Johnson, N.L. (eds), *Breakthroughs in statistics*, Springer, New York, 1992, 308–334.

Artem Vasilhevich Logachov
Lab. of Probability Theory and Math. Statistics, Sobolev Institute of Mathematics,
pr. Koptyuga, 4,
630090, Novosibirsk, Russia
Dep. of Computer Science in Economics, Novosibirsk State Technical University
pr. K. Marksa, 20,
630073, Novosibirsk, Russia
Email address: omboldovskaya@mail.ru

Anatoly Alfredovich Mogulskii
Lab. of Probability Theory and Math. Statistics, Sobolev Institute of Mathematics,
pr. Koptyuga, 4,
630090, Novosibirsk, Russia
Email address: mogul@math.nsc.ru

Anatoly Andreevich Yambartsev
Institute of Mathematics and Statistics, University of São Paulo,
Rua do Matão, 1010,
CEP 05508-090, São Paulo, SP, Brazil
Email address: yambar@usp.br