Methods for Graphlets Frequency Computation in Networks

Martin Nehéz[1,*], Teodora Šimon[1]

[1] *Institute of Computer Science and Mathematics, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava*
Ilkovičova 3, 841 04 Bratislava, Slovak Republic

Abstract
In the article, we address various methods for counting the occurrence of graphlets in networks. In addition to algorithmic methods, we will focus on combinatorial and probabilistic methods. We estimate the counts of graphlets for triangle chain graphs and sparse random Erdős-Rényi graphs, respectively.

Keywords
Graphlet, Orbit, Triangle Chain Graphs, Random graphs, Probabilistic method

1. Introduction

The analysis of large real-world or complex networks differs significantly in many respects from research in graph theory. Specifically, real-world networks are understood as macroscopic objects, and statistical methods are often useful for describing them. In contrast, graph theory is mostly deterministic and deals with graphs whose structure is exactly defined. Moreover, although graph theory often deals with symmetric graphs (e.g. Cayley graphs), real-world networks rarely exhibit a symmetric nature.

An analogous situation arises when searching for anomalies or similarities between two given real-world networks. Standard graph isomorphism is, in fact, inapplicable to such a problem (see below).

The problem of network similarity has recently gained significant importance, driven by research questions arising in fields such as bioinformatics [1] and social networks [2]. For example, protein-protein interaction networks can have a similar structure even if they differ in terms of cardinality [1, 3] (i.e., they may have different numbers of vertices and edges). That is one of the reasons why the graph isomorphism is unusable. On the other hand, methods based on a machine learning approach appear more suitable [4]. The principal idea is based on the assumption that the local structure of networks can influence their global properties. More precisely, the frequency of occurrence of small motifs or subgraphs forms a vector that can characterize a given network. Subsequently, instead of comparing entire networks, only these vectors are compared. A key task in this approach is determining the frequency of the relevant subgraphs.

In many research papers concerning network similarity, small relevant subgraphs are commonly referred to as *graphlets*—i.e., connected induced subgraphs [1, 3]. They typically have between 2 and 5 vertices. (Paper [5] suggests that 4 vertices are sometimes sufficient.) Although they are small, determining their number in a large network poses a challenging problem. Paper [6] represents arguably the best general solution available to date. It is based on an algorithm that uses a system of enumeration equations for graphlets and so-called orbits. Corresponding freeware is available under the name "ORCA" [6].

However, the approach based on the ORCA algorithm is not the only one. Some earlier methods used, for example, less accurate randomized algorithms [7]. Since algorithmic methods have their limitations, the question arises whether it is possible to use suitable enumeration formulas. Partially affirmative

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia
*Corresponding author.
martin.nehez@stuba.sk (M. Nehéz); teodora.simon@stuba.sk (T. Šimon)
https://uim.fei.stuba.sk/en/pracovnici/nehez/ (M. Nehéz)
0000-0002-3033-9019 (M. Nehéz); 0009-0006-1911-6056 (T. Šimon)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

CEUR-WS.org/Vol-4323/paper34.pdf

answer to the above question is possible due to Kirchhoff's matrix tree theorem. For graphlets that are trees, it is theoretically possible to derive the corresponding formulas. On the other hand, focusing on certain graph classes can also yield some results. Yet another option would be to test generative AI for this goal.

In this paper, we present several analytical methods for enumerating graphlet frequencies. We show that certain known results from graph theory can be applied for this purpose. We focus primarily on triangle chain graphs and Erdős-Rényi random graphs, respectively. We also outline the use of algorithmic methods.

The organization of the paper is as follows. Sect. 2 contains definitions and preliminary facts. The software tools for graphlet enumeration are described in Sect. 3. Our main results are presented in Sect. 4 and Sect. 5, respectively. An analytical estimate of graphlet counts for triangle chain graphs are expressed in Sect. 4. An analogical result for sparse Erdős-Rényi random graphs is presented in Sect. 5.

2. Definitions and Preliminaries

The asymptotic notation such as o, O, Θ [8] is used in usual way. Moreover, for two sequences (or equivalently functions) $a = (a_n)_{n=0}^{\infty}$ and $b = (b_n)_{n=0}^{\infty}$, we will write $a_n \ll b_n$ if $a_n \geq 0$ and $a_n = o(b_n)$.

Throughout this paper, all graphs are simple, undirected, and without weights. Standard notions of the graph theory are used without definitions and further comments. We address [9] for references. Nevertheless, some necessary notions are explained bellow.

A *graph* $G = (V, E)$ consists of a set of vertices (or nodes), denoted by V, and set of edges (or links, connections), denoted by E. Usually $|V(G)| = n$, where $|.|$ stands for the cardinality of a given set. For a vertex $v \in V(G)$, its degree (the number of adjacent vertices of v) is denoted by $\deg(v)$. A *path* is a sequence of connected vertices in which no vertex and no edge appears more than once. A path with n vertices is usually denoted by P_n.

Two graphs $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ are said to be *isomorphic* if there is a one-to-one mapping $\sigma : V_1 \to V_2$, called *isomorphism*, such that for all pairs $u, v \in V_1$ it holds $\{u, v\} \in E_1$ if and only if $\{\sigma(u), \sigma(v)\} \in E_2$. An isomorphism of a graph G onto itself (i.e. $\sigma : G \to G$) is referred to as *automorphism*. Given a graph G, let $Aut(G)$ denote a group of all automorphisms $\sigma : G \to G$. Given a vertex $v \in V$, the *orbit* of v is defined by

$$Orb(v) = \{\, \sigma(v) \mid \sigma \in Aut(G) \,\}\,. \tag{1}$$

The *stabilizer* of a vertex v, denoted by $Stab(v)$, is a subgroup of $Aut(G)$ such that

$$Stab(v) = \{\, \sigma \in Aut(G) \mid \sigma(v) = v \,\}\,. \tag{2}$$

A graph is said to be *connected* if for any pair of its vertices u and v, there exists a sequence of edges that form a path between u and v. Given a graph $G = (V, E)$, a graph $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$ is referred to as a *subgraph* of G. Let $G = (V, E)$ be a graph and let $S \subseteq V(G)$, an *induced subgraph* $G[S]$ is the graph (S, E') where $E' = \{\, \{u, v\} \in E \mid u \in S \text{ and } v \in S \,\}$.

Let $G = (V, E)$ be a graph, a *graphlet* is a connected induced subgraph of G. Two graphlets are same if there exists an isomorphism such that it maps one graphlet to the other one. (Two different occurrences of the same graphlet are usually referred to as its copies.) There are totally 30 graphlets with $2, 3, 4$ and 5 vertices. List of all graphelts can be found e.g. in [6], p. 2. Ordering (and labeling) of graphlets with at most 4 vetices is shown in Fig. 1. Specifically, the 2-vertex graph with one edge is denoted by g_0, the 3-vertex path corresponds to g_1, etc. The last graphlet in Fig. 1 (i.e. g_8) is the clique K_4.

The number of all orbits is 73 [6]. In this paper, they are denoted by $o_0, o_1, \ldots, o_{72}$. Recall that, given a graph G, each orbit is a set of vertices that are mapped to each other in any automorphism of G. As for the graphlet g_0, both vertices belong to the same orbit o_0. The graphlet g_1 has two orbits: o_1 contains two vertices with degree 1 and o_2 contains only the single "central" vertex. Graphlet g_2 has only one orbit o_3, since all vertices belong to o_3, etc. The numbering of orbits (from o_0 to o_{14}) and their assignment to graphlets are presented in Tab. 1.

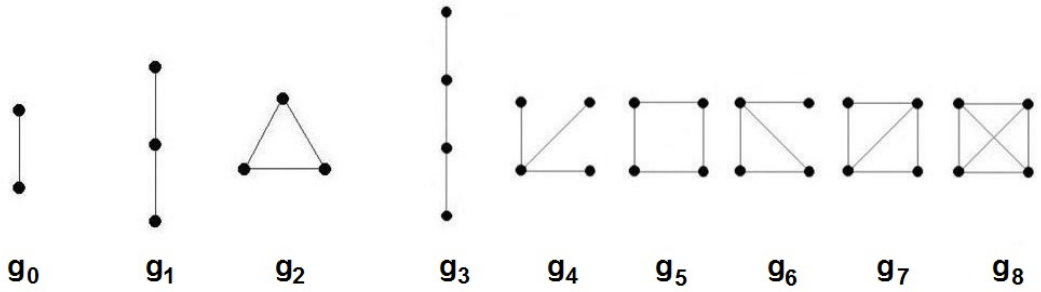


Figure 1: All graphlets with 2, 3 and 4 vertices.

Table 1
Numbers of vertices occurred in orbit o_j and assignment of orbits to graphlets. Numbering of orbits is the same as used in [6], Fig. 1, p. 2.

Graphlet g_k	Orbit o_j	No. of vertices occurred in o_j with resp. to g_k
g_0	o_0	2
g_1	o_1	2
g_1	o_2	1
g_2	o_3	3
g_3	o_4	2
g_3	o_5	2
g_4	o_6	3
g_4	o_7	1
g_5	o_8	4
g_6	o_9	1
g_6	o_{10}	2
g_6	o_{11}	1
g_7	o_{12}	2
g_7	o_{13}	2
g_8	o_{14}	4

In general, the relationship between the numbers of automorphisms, stabilizers, and orbits is determined by the orbit stabilizer theorem.

The Orbit Stabilizer Theorem. If $Aut(G)$ is a group of all automorphisms of G, then for every $v \in V(G)$, we have:

$$|Aut(G)| = |Orb(v)| \cdot |Stab(v)| \,. \tag{3}$$

The number of automorphism for small graphlets is usually simple. For example, the graphlet g_6 has 3 orbits: two with cardinality 1 (namely o_9 and o_{11}) and one with cardinality 2 (o_{10}). The number of automorphisms of g_6 is simply 2 because only two vertices of o_{10} can be mutually mapped to each other.

On the other hand, each clique K_n (i.e., graphlets g_2 and g_8) has only one orbit comprising all the vertices. Because K_n has all possible edges, so $|Aut(K_n)| = n!$, since these are all permutations of an n-elements set. The numbers of all automorphisms for graphlets $g_0, \ldots, g_8$ are listed in Tab. 5.

3. Software Tools for Graphlet Enumeration

Searching for the exact number of all graphlets in a given network is algorithmically demanding problem. Thanks to the equations for node-orbit counts [6], pp. 17–24, the brute-force approach can be abandoned in favor of a more efficient method. A very effective software tool for this purpose (and likely the best at the moment) is the Orbit Counting Algorithm, shortly ORCA, developed by T. Hočevar [6].

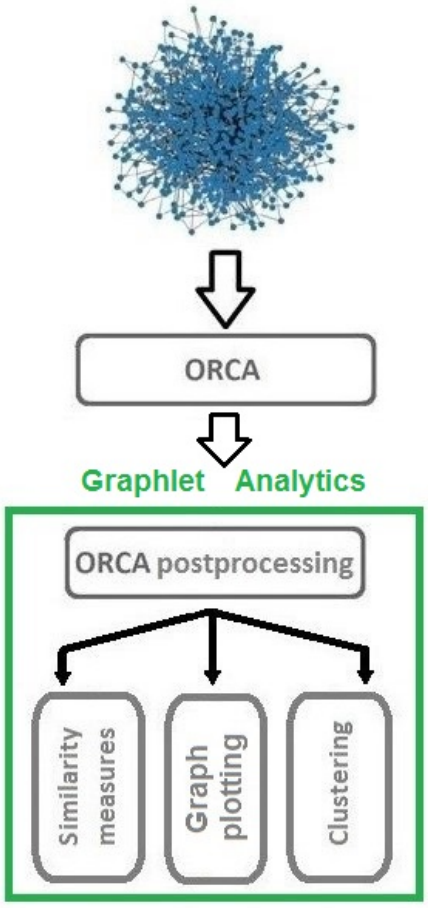


Figure 2: Process of Graphlet-based Analysis of Networks.

ORCA is a freeware console application which can operate in 2 modes: processing graphlets either with at most 4 vertices or with at most 5 vertices. Upon the latter mode, and for an input graph with n vertices, the output is the file with a matrix of $n \times 73$ integers. The value $a_{i,j}$ of the matrix is the frequency of ith vertex in the jth orbit, where $i = 0, \ldots, n-1$ and $j = 0, \ldots, 72$. It follows that to obtain the number of graphlet occurrences, it is necessary to sum numbers in columns occurred in the output from ORCA. Each such sum represents the number of all orbits in an input graph. In order to determine the number of graphlets, the sums for each column must be divided by the number of vertices occurring in each orbit. For each orbit, this number is listed in the third column of Tab. 1.

However, counting the number of graphlet occurrences is not part of the ORCA software. We will refer such a procedure to as *ORCA-postprocessing*.

For this reason, a system named "Graphlet Analytics" (shortly GA) was developed at our institute by the second author of this paper. It comprises of two parts:

- ORCA-postprocessing,
- functionality for graphlet-based analytics (it includes graph plotting, the calculation of network similarity measures, and k-means clustering).

The entire process of graphlet-based analysis is shown in Fig. 2. Its first phase is executed by ORCA, and the second phase by GA. The "Graphlet Analytics" software is shown in the green box in Fig. 2.

4. Graphlets in Triangle Chain Graphs

Let k be a positive integer, a *k-triangle chain* is a graph consisting of k triangles, where each consecutive pair of triangles shares exactly one vertex. It is denoted by TC_k.

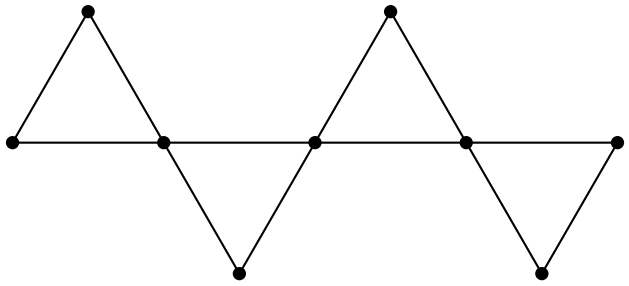


Figure 3: Graph TC_4.

Inductively, TC_1 is a triangle (or the cycle) K_3. For $k \geq 2$, TC_k is formed from TC_{k-1} by attaching a triangle K_3 to the TC_{k-1} in such a way that triangle K_3 and an arbitrary end-triangle (outermost) of TC_{k-1} share exactly one vertex u, originally with $\deg(u) = 2$.

Graph TC_4 is shown in Fig. 3. The class of k-triangle chain graphs is defined by

$$TC = \{\, TC_k \mid k \geq 1 \,\} .$$

Each k-triangle chain TC_k has $n = 2k + 1$ vertices, $3k$ edges, k triangles and $k - 1$ vertices with degree 4. The TC_k consists of $k - 2$ *internal triangles* and 2 *end-triangles* (or *outermost triangles*).

In this section, we determine the number of graphlets in class TC using both analytical and computational methods. The analytical method is based on a pure combinatorial argument. To verify the analytical results, we use a computational method that employs the software tools described in Sect. 3.

Given a graph H, let $N_i(H)$ denote the number of all occurrences of the graphlet g_i in H for $i = 0, \ldots, 8$. It is easy to see that graphlets g_4, g_5, g_7 and g_8 do not occur in TC_k, see also Fig. 3. It follows that

$$N_j(TC_k) = 0 \qquad \text{for } j \in \{4, 5, 7, 8\} .$$

The numbers of graphlets occurring in TC_k with positive frequencies are determined in the following statement.

Lemma 1. *If $k \geq 1$, then it holds:*

$$N_0(TC_k) = 3k \,, \tag{4}$$

$$N_1(TC_k) = 4(k-1) \,, \tag{5}$$

$$N_2(TC_k) = k \,, \tag{6}$$

$$N_3(TC_k) = \begin{cases} 0 & \text{if } k \in \{1, 2\} \,, \\ 4(k-2) & \text{if } k \geq 3 \,, \end{cases} \tag{7}$$

$$N_6(TC_k) = 4(k-1) \,. \tag{8}$$

Proof. The graphlet g_0 is an edge; therefore, $N_0(TC_k)$ represents the number of edges in TC_k. It follows that $N_0(TC_k) = 3k$.

The graphlet g_1 is a path P_3. Since g_1 is an induced subgraph, the central vertex of g_1 must correspond to exactly one of the vertices of TC_k that have degree 4. There are $(k - 1)$ ways to choose a vertex of degree 4 in TC_k. If such a vertex is chosen, then there are $2 \cdot 2$ ways to chose two end-vertices. Thus, $N_1(TC_k) = 4(k - 1)$.

The graphlet g_2 is a triangle, and clearly $N_2(TC_k) = k$ holds.

The graphlet g_3 is a path P_4. Since g_3 is an induced subgraph, the central edge of g_3 must correspond exactly to any edge connecting the two vertices of degree 4 in TC_k. Fig. 4 illustrates such a situation; the central edge of g_3 is denoted by e_2. Note that g_3 does not occur in either TC_1 or TC_2. If $k \geq 3$, then there are $(k - 2)$ ways to choose the placement of e_2. If e_2 is chosen, then the remaining two edges of g_3 can be chosen in $2 \cdot 2 = 4$ ways; see Fig. 4. It follows that $N_3(TC_k) = 4(k - 2)$ if $k \geq 3$.

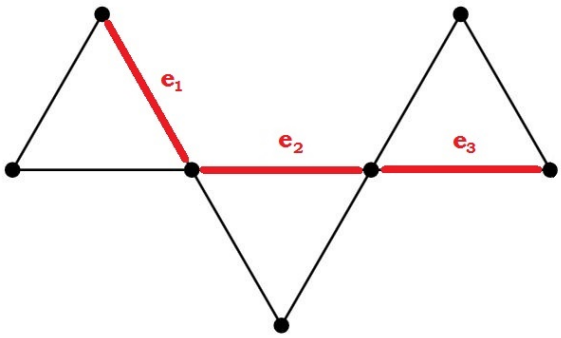


Figure 4: One possible placement of the graphlet g_3 in TC_3. Edges of g_3 are shown in red.

The graphlet g_6 is a triangle with one additional edge. The graph TC_k contains k triangles; however, it is necessary to distinguish between internal triangles and end-triangles. If we chose an internal triangle (there are $(k-2)$ ways to chose it), then the vertex of g_6 that has degree 1 can be chosen in 4 ways in TC_k. Thus, in this case, we have $4(k-2)$ possibilities. Otherwise, if we chose an end-triangle (2 options), then the vertex of g_6 with degree 1 can be chosen in 2 ways in TC_k. Thus, we have $2 \cdot 2$ options. In total, we get:

$$N_6(TC_k) = 4(k-2) + 4 = 4k - 4 = 4(k-1)\,.$$

□

If n is very large, approximate graphlet counts can also be useful. For this reason, the following statement also presents asymptotic estimates of the number of graphlets.

Corollary 1. *The asymptotic estimations of numbers of graphtlets for TC_k is listed in Tab. 2.*

Table 2
Asymptotic graphlet counts for TC_k.

i	$N_i(TC_k)$
0	$\frac{3}{2}n + O(1)$
1	$2n + O(1)$
2	$n/2 + O(1)$
3	$2n + O(1)$
6	$2n + O(1)$

Proof. Each TC_k has $n = 2k + 1$ vertices. By substitution of $k = (n-1)/2$ into equations (4)-(8), we get resulting asymptotic formulas in Tab 2. □

As we see in Tab. 2, the number of all graphlets grows linearly with respect to n. This fact is relatively straightforward, as TC_k has a simple structure and a linear number of edges. Nevertheless, k-triangle chain graphs are an interesting example of a graph class in which some graphlets have zero occurrences, while the count of others grows linearly. This provides a rough idea of how to construct graphs with the desired number of graphlets.

Manually counting graphlet frequencies is challenging in more complex cases. Therefore, the formulas given in Lemma 1 apply only to graphlets with at most 4 vertices. Consequently, counting the number of graphlets using software applications remains relevant.

The second part of this section focuses on experiments. We generated graphs TC_k for $k \in \{4, 10, 100, 1000\}$ in an input format suitable for processing with the ORCA software. Subsequently, we

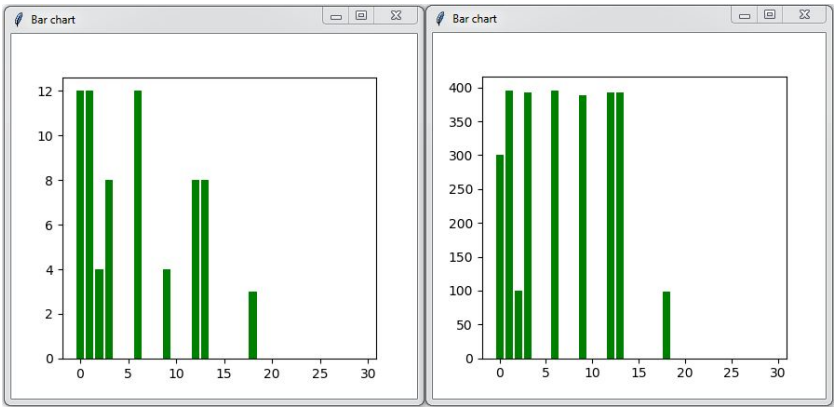


Figure 5: Number of graphlets in the graph TC_4 (left) and in the graph TC_{100} (right). In both charts, the x-axis shows the graphlet numbers, and the y-axis shows their corresponding counts. The counts are displayed for all graphlets with 2 to 5 vertices.

processed these datasets using the software-based analysis described in Sect. 3 — namely: ORCA → GA. The obtained graphlet counts are listed in Tab. 3. A visual representation of the corresponding values for the two instances, i.e., TC_4 and TC_{100}, is shown in Fig. 5. Both charts in the figure were generated using GA. The counts are displayed for all graphlets with 2 to 5 vertices.

To verify the results in Tab. 3, we also utilized generative AI models of the LLM type—specifically ChatGPT and Calude Opus 4.6 Thinking, respectively. The former one made several errors, whereas the latter one yielded correct results. The ChatGPT was able to correct itself after we pointed it out.

Table 3
Numbers of graphlets $g_0, \ldots, g_6$ calculated for TC_k.

k	$N_0(TC_k)$	$N_1(TC_k)$	$N_2(TC_k)$	$N_3(TC_k)$	$N_4(TC_k)$	$N_5(TC_k)$	$N_6(TC_k)$
4	12	12	4	8	0	0	12
10	30	36	10	32	0	0	36
100	300	396	100	392	0	0	396
1000	3000	3996	1000	3992	0	0	3996

5. Graphlets in Random Graphs

The Erdös-Rényi graphs are well known, among other things, for the use of probabilistic methods in proofs. While combinatorial arguments were used in the calculation of graphlets counts in the previous section, in this section we will use probabilistic methods.

The *Erdös-Rényi random graph model* (shortly *E-R model*) [8] is introduced as follows.

Let n be a positive integer and let $p \in \mathbb{R}$ be a constant such that $0 < p < 1$. Consider that for n labeled vertices of $V(G)$, each unordered pair of vertices introduces one *slot* available for an edge. Clearly, the total number of slots is $\binom{n}{2}$. Each edge exists in G independently and with the probability p, thus $\Pr[\{u, v\} \in E(G)] = p$, for all $u, v \in V(G)$.

Let $(\Omega, \mathbb{F}, \Pr)$ be a probability space where the sample space Ω consists of all (labeled) graphs G of the order n. If G has $|E(G)|$ edges, $0 \leq |E(G)| \leq \binom{n}{2}$, then the probability of obtaining G as a result of random edge generation process is given by:

$$\Pr[G] = p^{|E(G)|}(1-p)^{\binom{n}{2}-|E(G)|} . \tag{9}$$

The probability space $(\Omega, \mathbb{F}, \Pr)$ is denoted by $\mathbb{G}(n, p)$ or $G_{n,p}$ and called *the probability space of all random graphs* or *E-R random graph model.*

A statement about a random graph from $\mathbb{G}(n,p)$ is said to hold *asymptotically almost surely (a.a.s.)* if it holds with probability approaching 1 as $n \to \infty$. A graph G with n vertices is said to be *dense* if it has $\Theta(n^2)$ edges (i.e. asymptotically equal to n^2) and G is said to be *sparse* if it has $o(n^2)$ edges (i.e. asymptotically less than n^2).

Threshold functions are usually defined for monotone properties. In this paper, we use them only for increasing properties. For further details see [8]. Let K_n be a clique with $n > 0$ vertices and let 2^{K_n} denote the set of all subgraphs of K_n with n vertices. Let Q be an increasing property of graphs from 2^{K_n}. A function $\hat{p} = \hat{p}(n)$ is called *threshold function* for Q iff the following condition holds:

$$\Pr[\ \mathbb{G}(n,p) \text{ has } Q\ ] \to \begin{cases} 0 & \text{if} \quad p \ll \hat{p}\,, \\ 1 & \text{if} \quad p \gg \hat{p}\,. \end{cases}$$

Let X be a random variable on $\mathbb{G}(n,p)$. Expectation of X is denoted by $\mathbb{E}(X)$ and variance of X by $Var(X)$.

5.1. Previous results on random graphs

In this subsection, we list the results concerning random graphs that are necessary for our analysis in the next subsection. The threshold function for graphlets are listed in Tab. 4. They were derived in [8, 10].

Table 4
Threshold functions of nontrivial graphlets [10].

Graphlet g_i	Description	Threshold function
g_0	Edge	n^{-2}
g_1	Path P_3	$n^{-3/2}$
g_2	Triangle $\triangle$	n^{-1}
g_3	Path P_4	$n^{-4/3}$
g_4	3-star	$n^{-4/3}$
g_5	Cycle C_4	n^{-1}
g_6	$\triangle$+edge	n^{-1}
g_7	Chordal-C_4	$n^{-4/5}$
g_8	Clique K_4	$n^{-2/3}$

Let $X_{n,i}$ denote the random variable on $\mathbb{G}(n,p)$ such that $X_{n,i}$ is the number of copies of graphlet g_i in a random graph for $i = 0, \ldots, 8$.

Lemma 2. *([8]) For $i = 0, \ldots, 8$, let n_i (m_i) denote the number of vertices (edges) of g_i. The expectation of the random variable $X_{n,i}$ is given by*

$$\mathbb{E}(X_{n,i}) = \frac{n_i!}{aut(g_i)} \binom{n}{n_i} p^{m_i} (1-p)^{\binom{n_i}{2} - m_i}\,, \tag{10}$$

where $aut(g_i)$ denote the number of automorphisms of g_i.

Let us define the random variable $Y_{n,I}$ as follows:

$$Y_{n,i} = \frac{X_{n,i} - \mathbb{E}(X_{n,i})}{\sqrt{Var(X_{n,i})}}\,. \tag{11}$$

Let

$$r(G) = \max\left\{ \frac{|E(H)|}{|V(H)|}\ ;\ H \subseteq G, |V(H)| > 0 \right\}. \tag{12}$$

Lemma 3 ([11]). *Let g_i be the i-th graphlet and let $r_i = r(g_i)$. Then $Y_{n,i}$ is asymptotically normally distributed if and only if*

1. $np^{r_i} \to \infty$, *and*
2. $n^2(1-p) \to \infty$.

5.2. The New Results

It is well known that real-world networks are usually sparse, sometimes they even have linear number of edges. It is easy to see that if $p = cn^{-1}$ for some constant $c > 0$, then random graphs from $\mathbb{G}(n, cn^{-1})$ have asymtotically linear number of edges. In such a case, however, according to Tab. 4, graphlets g_7 and g_8 do not appear in $\mathbb{G}(n, cn^{-1})$ a.a.s. This is because their threshold function is asymptotically greater than cn^{-1}. Therefore, in this subsection, we set $p = n^{-1/2}$. Such an assumption allows $\mathbb{G}(n, p)$ to remain sparse, albeit denser than in the case of $p = cn^{-1}$. More precisely, if $p = n^{-1/2}$, then we have $|E| = p\binom{n}{2} = \Theta(n^{-1/2} \cdot n^2) = \Theta(n^{3/2})$. Thus $|E| = o(n^2)$, hence the random graphs are sparse a.a.s.

By Lemma 2, each $\mathbb{E}(X_{n,i})$ can be rewritten into the form of the 4^{th} column of the Tab. 5. A better explanation is provided in the following example.

Table 5
Graphlets, numbers of their automorphisms, values of r_i, and values of $\mathbb{E}(X_{n,i})$.

Graphlet g_i	$aut(g_i)$	$r_i = r(g_i)$	$\mathbb{E}(X_{n,i})$
g_0	2	1/2	$\binom{n}{2}p$
g_1	2	2/3	$\frac{3!}{2}\binom{n}{3}p^2(1-p)$
g_2	6	1	$\frac{3!}{6}\binom{n}{3}p^3$
g_3	2	3/4	$\frac{4!}{2}\binom{n}{4}p^3(1-p)^3$
g_4	6	3/4	$\frac{4!}{6}\binom{n}{4}p^3(1-p)^3$
g_5	8	1	$\frac{4!}{8}\binom{n}{4}p^4(1-p)^2$
g_6	2	1	$\frac{4!}{2}\binom{n}{4}p^4(1-p)^2$
g_7	4	5/4	$\frac{4!}{4}\binom{n}{4}p^5(1-p)$
g_8	24	3/2	$\frac{4!}{24}\binom{n}{4}p^6$

Example 1. We explain the calculation of $\mathbb{E}(X_{n,6})$ in Tab. 5 for the graphlet g_6. We shall express all the quantities appearing in eq. (10) of Lemma 2. The graphlet g_6 has:

- $n_6 = 4$ vertices,
- $m_6 = 4$ edges,
- 2 empty slots with no edges present (i.e. 2 edges missing to get K_4).

The number of empty slots is calculated as the exponent for the term $(1-p)$ as follows:

$$\binom{n_6}{2} - m_6 = \binom{4}{2} - 4 = 6 - 4 = 2\,. \tag{13}$$

The number of automorphisms is $aut(g_6) = 2$, since two vertices of the orbit o_{10} can be mapped to each other and the remaining orbits (o_9 and o_{11}) each have a single element, see Tab. 1 and Tab. 5. By gathering all these values and substituting them into (10), we get the resulting equation:

$$\mathbb{E}(X_{n,6}) \;=\; \frac{n_6!}{aut(g_6)}\binom{n}{n_6}p^{m_6}(1-p)^{\binom{n_6}{2}-m_6} \;=\; \frac{4!}{2}\binom{n}{4}p^4(1-p)^2\,. \tag{14}$$

Our main result is the following statement.

Theorem 1. *Let $X_{n,i}$ and $Y_{n,i}$ denote random variables for a graphlet g_i as above. In E-R random graph $\mathbb{G}(n,p)$, where $p = n^{-1/2}$, expectations of all random variables $X_{n,i}$ satisfy asymptotic equations as listed in Tab. 6. Moreover all $Y_{n,i}$ are asymptotically normally distributed for $i = 0, \ldots, 8$.*

Table 6
Values of $\mathbb{E}(X_{n,i})$.

Graphlet g_i	$\mathbb{E}(X_{n,i})$
g_0	$\frac{1}{2}n^{3/2}$
g_1	$\frac{1}{2}n^{2}$
g_2	$\frac{1}{6}n^{3/2}$
g_3	$\frac{1}{2}n^{5/2}$
g_4	$\frac{1}{6}n^{5/2}$
g_5	$\frac{1}{8}n^{2}$
g_6	$\frac{1}{2}n^{2}$
g_7	$\frac{1}{4}n^{3/2}$
g_8	$\frac{1}{24}n$

Proof. It follows from Lemma 2 and Lemma 3. Asymptotic estimations are based on the fact that $p = n^{-1/2}$ and on the following two equations.

1. If $\alpha \in \{3,4\}$ and $\beta \in \{3,4,5,6\}$, then:

$$\binom{n}{\alpha} p^{\beta} \;=\; \frac{n^{\alpha}}{\alpha!} \cdot n^{-\beta/2} \left(1 + O(n^{-1})\right) \;=\; \frac{n^{\alpha-\beta/2}}{\alpha!} \left(1 + O(n^{-1})\right) . \tag{15}$$

2. If $p \to 0$ and $k \in \{1,2,3\}$, then:

$$(1-p)^k = 1 - kp + O(p^2) . \tag{16}$$

By substituting $p = n^{-1/2}$ into (16), we get:

$$(1-p)^k = 1 - kn^{-1/2} + O(n^{-1}) = 1 + O(n^{-1/2}) . \tag{17}$$

If p, α, β, k are as above, then by (15) and (17), we get:

$$\binom{n}{\alpha} p^{\beta} (1-p)^k \;=\; \frac{n^{\alpha-\beta/2}}{\alpha!} \left(1 + O(n^{-1/2})\right) . \tag{18}$$

Finally, to obtain the expressions for $\mathbb{E}(X_{n,i})$ in Tab. 6, it suffices to simply substitute the specific values for p, α, β, k into (18).

The second part of proof follows from Lemma 3. It is necessary to verify two conditions of the lemma.

1. For the validity of condition $np^{r_i} \to \infty$, it is sufficient to verify it only for the maximum value of r_i. As can be seen from Tab. 5, the maximum value of r_i is reached at $i = 8$, i.e. for the clique K_4. For $r_8 = 3/2$ and $p = n^{-1/2}$, we therefore get:

$$np^{r_i} = n \cdot n^{-3/4} = n^{1/4} \to \infty \tag{19}$$

as $n \to \infty$. The condition is valid.

2. The second condition is easy to verify. For $p = n^{-1/2}$, we get $n^2(1-p) = n^2(1 - n^{-1/2}) \to \infty$ as $n \to \infty$, as required.

Therefore, random variables $Y_{n,i}$ are asymptotically normally distributed for all graphlets with at most 4 vertices. □

6. Conclusion

Currently, the most widely used method for determining the number of graphlets in networks is based on an algorithmic approaches. ORCA is arguably the best software tool to date for the accurate enumeration of graphlets.

We have shown that for certain graph classes, the given problem can be solved exclusively by analytical methods. We focused on triangle chain graphs and sparse E-R random graphs. Corresponding formulas were derived for both classes of graphs in Sect. 4 and Sect. 5, respectively. In the first case, combinatorial argument was used. In the second case, probabilistic methods were used in conjunction with known results.

There are a relatively large number of open problems.

First, whether analytical methods allow for the estimation of graphlets for classes of graphs such as random geometric graphs or the Chung-Lu random graph model, etc.

Secondly, what are the limitations of the ORCA software, and how does its performance compare to that of generative AI? In our opinion, generative AI does not yet reach the level of accuracy provided by ORCA, which calculates precise results.

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback and insightful comments. The authors also gratefully acknowledge the financial support provided by the head of the Institute of Computer Science and Mathematics, prof. M. Vojvoda.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to minor grammar and spelling check. Further, the authors also used ChatGPT to supplement the methods mentioned in Section 1 and to locate the paper [11] on the web. Finally, ChatGPT and Claude Opus 4.6 Thinking were used to compare the results presented in Tab. 3.

After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] N. Pržulj, D. G. Corneil, and I. Jurišica. "Modeling Interactome: Scale-Free or Geometric?" In: *Bioinformatics* 20.18 (2004), pp. 3508–3515.

[2] R. Charbey and C. Prieur. "Stars, Holes, or Paths Across Your Facebook Friends: A Graphlet-Based Characterization of Many Networks". In: *Network Science* 7.4 (2019), pp. 476–497.

[3] N. Pržulj. "Biological Network Comparison Using Graphlet Degree Distribution". In: *Bioinformatics* 23.2 (2007), pp. 177–183.

[4] I. Jurišica and D. Wigle. *Knowledge Discovery in Proteomics*. Boca Raton: Chapman & Hall/CRC Press, 2006.

[5] J. Janssen, M. Hurshman, and N. Kalyaniwalla. "Model Selection for Social Networks Using Graphlets". In: *Internet Mathematics* 8.4 (2012), pp. 338–363.

[6] T. Hočevar and J. Demšar. "Computation of Graphlet Orbits for Nodes and Edges in Sparse Graphs". In: *Journal of Statistical Software* 71.10 (2016), pp. 1–24.

[7] D. Marcus and Y. Shavitt. "RAGE - A rapid graphlet enumerator for large networks". In: *Comput. Networks* 56.2 (2012), pp. 810–819.

[8] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. New York: John Wiley & Sons, 2000.

[9] J. L. Gross and J. Yellen. *Handbook of Graph Theory*. CRC Press, 2003.

[10] M. Nehéz. "Probabilistic Analysis of Graphlets in Sparse E-R Random Graphs". In: *Proceedings of the 18th Conference Information Technologies – Applications and Theory (ITAT 2018)*. Vol. 2203. CEUR Workshop Proceedings. CEUR-WS.org, 2018, pp. 171–175.

[11] A. Ruciński. "When Are Small Subgraphs of a Random Graph Normally Distributed?" In: *Probability Theory and Related Fields* 78 (1988), pp. 1–10.