# Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig


## Random Graphs with Motifs

(revised version: September 2011)

by


*Anatol Wegner*



Preprint no.: 61 2011


# Random Graphs with Motifs

Anatol E. Wegner
*Max Planck Institute for Mathematics in the Sciences and Inselstr. 7, Leipzig, Germany*
(Dated: September 21, 2011)

We introduce and analyze a particular generalization of the Erdös-Rényi random graph model that is based on adding not only edges but also copies of small graphs onto the nodes of a graph. The resulting model is analytically tractable and can generate random graphs with local structures that are not tree like. First, we introduce the simplest generalization called the triplet model corresponding to the addition of three node subgraphs and investigate some of its key properties. In the undirected model we obtain a random graph with non-zero clustering and assortativity while in the directed case we obtain a graph with triangular motifs. Then, we formulate our model in its full generality by allowing the addition of graphs of arbitrary size and generalize the results obtained for the triplet model.

## I. INTRODUCTION

Complex networks are studied across disciplines as many real world systems can be modeled as networks where nodes represent interacting elements and edges interactions between them. Some of the most studied properties networks are the small world property, clustering, assortativity, power law degree distributions and network motifs [1–3]. Consequently, the construction of mathematical models that capture one or more of these properties of of real world complex networks while being analytically tractable is of much interest. The most studied of such models is the Erdös-Rényi model[4]. Although its study has provided essential insights into complex networks it is considered to be a rather unrealistic model for real networks as it fails to account for the characteristics of real world-complex networks listed above except the small world property. This is mainly due to its narrow-Poisson type degree distribution and also the fact that it is locally tree like and therefore fails to produce densely connected local structures. Our approach in this paper is based on constructing graphs by adding not only edges but also small subgraphs on to the nodes of a graph. The parameters of the model are directly related to the probability of adding a certain subgraph on to a set of nodes in analogy with the Erdös-Rényi model where edges are independently added between pairs of nodes with a certain probability. These random graph models can produce sparse graphs with high expectation values for counts of highly connected subgraphs and can be solved analytically for much of their properties. Consequently, graphs with local structures that are not tree like can be generated and their general properties can be investigated within the framework of the model.

The motivation for constructing our model is the observation that properties of complex networks like motifs [5], clustering[6, 7], assortativity[8] and heavy tailed degree distributions [8] are interrelated through subgraph counts. Whereas the clustering coefficient is determined by the number of triangles and two stars[1], the assortativity can be expressed in terms subgraph counts [8] and the n-th moments of the degree distribution are determined by the counts of the star shaped subgraphs up to size n[8]. Therefore a model that is able to match subgraph counts of a given network would also reproduce properties that are statistics of the subgraph counts.

There are various definitions for network motifs but essentially they all depend on comparing subgraph counts of a given network with a certain null-model, usually some version of the configuration model[5]. In this paper we construct networks using motifs as building blocks, rather than then assessing the significance of network motifs compared to a null-model.

A similar and more elaborate model that is also based on the addition of small subgraphs has been proposed by Bollobas et al. [9] and some special instances of the model presented here are equivalent to particular cases of the model in [9].

The paper is structured as follows: First we formulate the Erdös-Rényi model in order to clarify the nature of the generalization and briefly review some of its essential properties for comparison with the generalized model. We then introduce the triplet model where three node subgraphs are added on to the nodes of a graph as an example for the generalized model. We analyze the triplet model in the undirected and directed cases in details. In the next section we generalize our model to include building blocks of arbitrary size and generalize the results obtained for the triplet model. We then also briefly discuss the hierarchy between multiplet models and its connection to complexity.

---

[1] A n-star consists of n edges attached to a central node

## II. THE ERDÖS-RÉNYI MODEL

The Erdös-Rényi model is arguably the most extensively studied random graph model [4]. In the Erdös-Rényi model edges occur independently with a fixed probability p and for undirected graphs it can be formulated starting with a the set of labeled nodes $V = \{1, 2, 3..N\}$ and the set $E = \{\{i, j\} : i, j \in V\}$ of pairs of nodes. For each of these pairs of nodes an edge is present between the pair with probability $p$ or absent with probability $(1 - p)$. The state space for pairs of nodes can be taken as $S = \{0, 1\}$-0 denoting the absence and 1 the presence of an edge. Since edges are assumed to occur independently, the probability distribution over the configuration space $X = \{0, 1\}^{C_2^N}$ is given by:

$$P(x) = p^{e(x)}(1 - p)^{C_2^N - e(x)} \tag{1}$$

[2] for all $x \in X$, where $e(x)$ is the number of edges in the configuration $x$. Each such configuration in $X$ corresponds to a unique labeled graph of which the adjacency matrix can be obtained by the following projection:

$$\Phi : X \longrightarrow A^{N\times N}, \qquad \Phi(x)_{\alpha\beta} = S(\{\alpha, \beta\}, x) \tag{2}$$

Where $S(\{\alpha, \beta\}, x)$ is the state of the pair $\{\alpha, \beta\}$ in the configuration x.

We recall some well known properties of the Erdös Rényi model, for a more comprehensive account we refer to [10]. The first quantity of interest which is the degree distribution is given by:

$$P(k) = C_k^{N-1} p^k (1 - p)^{N-1-k} \tag{3}$$

For asymptotically large graphs with a fixed mean degree $\kappa = pN$ this approaches a Poisson distribution:

$$P(k) \simeq \frac{\kappa^k e^{-\kappa}}{k!} \tag{4}$$

The degree distribution of the Erdös-Rényi random graph is narrowly concentrated around the mean degree which is one of the main reasons that the model is considered to be a rather poor model for real world networks with heavy tailed degree distributions.

The probability that a n-node connected graph H with e(H) edges is induced[3] on set of n nodes is given by:

$$P(H) = \Lambda(H) p^{e(H)} (1 - p)^{C_2^n - e(H)} \simeq \Lambda(H) \left(\frac{\kappa}{N}\right)^{e(H)} \tag{5}$$

where $\Lambda(H)$ is the number graphs isomorphic to H. The last part of the equation is valid for large graphs with a fixed mean degree $\kappa = pN$. In this case only subgraphs with $e(H) < n$ [4] have a high density and the clustering coefficient is $C = \kappa/N$ which tends to zero for large N. Here by high density we mean that $< n(H) > /N$ is nonzero as $N \to \infty$. Where $< n(H) >$ is the expected subgraph count of H.

The extension of the Erdös-Rényi model to the directed case is straight forward. To include directions one can either consider ordered pairs of nodes while keeping the state space as $S = \{0, 1\}$ or one can stick to the unordered pairs and enlarge the state space to $S = \{00, 01, 10, 11\}$ where 00 stand for absence while 01,10 for the two possible directed edges and 11 for a mutual edge. In order to preserve the symmetry between the nodes $p_{01}$ and $p_{10}$ have to be equal. In the first generalization the probability of a mutual edge is $p^2$, while it is a free parameter in the second. The second description is the one that fits better into the framework of our model, since our approach will be based on considering all possible subgraphs that can be realized on a set of nodes of size n.

## III. THE TRIPLET MODEL

### A. The Undirected Triplet Model

In this section, we will introduce and analyze the undirected triplet model which is the simplest generalization of the Erdös-Rényi model within the framework of our model. The undirected triplet model is based on assigning

---

[2] $C_k^N = \frac{N!}{(N-k)!k!}$

[3] An induced subgraph of a graph G is a subset of nodes together with all respective edges present in G.

[4] For connected graphs this is equivalent to H being a tree.

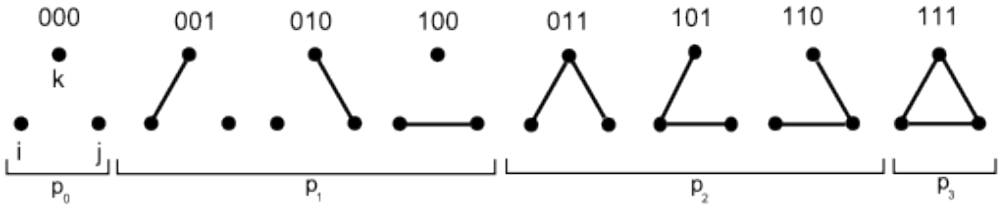


FIG. 1: The 8 states of a triplet.

each triplet of nodes in $T_3^N = \{\{i,j,k\} : i,j,k \in \{1,2,...N\}\}$ one of eight states, each corresponding to a 3 node subgraph. The graphical representations of the triplet states in $S_3 = \{0,1\}^3$ are shown in Fig.1. The triplets are independently assigned one of the states in S with the following probabilities: $P(000) = p_0, P(001) = P(010) = P(100) = p_1, P(011) = P(101) = P(110) = p_2, P(111) = p_3$. Isomorphic states (the 3 one edged states and the 3 V-shaped states) are assigned equal probabilities in order to preserve the symmetry between nodes. Following these assumptions the resulting probability distribution over the configuration space $X = S_3^{C_3^N}$ is:

$$P(x) = \prod_{t \in T_3^N} P(S(t,x)) = p_0^{n_0(x)} p_1^{n_1(x)} p_2^{n_2(x)} p_3^{n_3(x)} \tag{6}$$

where $S(t,x)$ is the state of t in the configuration x and $n_0(x)$, $n_1(x)$,$n_2(x)$ and $n_3(x)$ are the number of triplets in the empty, single edged, V-shaped and triangle states respectively. Written in terms of the state counts the probability distribution is the multinomial distribution:

$$P(n_0,n_1,n_2,n_3) = \left( \begin{array}{c} C_3^N \\ n_0,n_1,n_2,n_3 \end{array} \right) p_0^{n_0} (3p_1)^{n_1} (3p_2)^{n_2} p_3^{n_3} \tag{7}$$

[5] In order to obtain a probability distribution over labeled graphs of size N we define a random variable $\Phi(x)$ that maps the elements of the configuration space to adjacency matrices. $\Phi(x)$ can be thought of as projection from the latent configuration space X onto the edges of a graph. The projection is equivalent to adding an edge to the graph whenever the edge is present in the state of a triplet in the configuration x:

$$\Phi : X \longrightarrow A^{N \times N}, \qquad \Phi(x)_{\alpha\beta} = \sum_{t \in T_3 | \alpha,\beta \in t} A_{\alpha\beta}[S(t,x)] \tag{8}$$

where $A[S(t,x)]$ is the adjacency matrix of the state of the triplet t in the configuration x.

Defined in this way $\Phi$ induces a probability distribution over the space of labeled graphs of size N:

$$P_G(A) = P_X(\Phi^{-1}(A)) \tag{9}$$

Where A is an adjacency matrix with integer weights smaller than N-1. Since every edge is contained in N-2 triplets some edges will get a nonzero contribution from the state of more than one triplet producing an edge with integer weight larger than one. Whether to include the weights or not in the model is a matter of choice and if all weights are to be set to one the above formula has to be modified as $\Phi(x)_{\alpha\beta} = min(1, \sum_{t \in T_3 | \alpha,\beta \in t} A_{\alpha\beta}[S(t,x)])$. Allowing multiple edges facilitates calculations in some instances therefore in this section we perform calculations keeping multiple edges unless stated otherwise. Later we will see that in the case of sparse graphs multiple edges are rare and therefore their inclusion makes little difference.

In general, the inversion of the projection is quite complicated for large graphs since the configuration space is latent and in most cases there is large number of configurations that produce the same graph. On the other hand if one is interested in local properties of the model the inversion turns our to be relatively easy.

A more restricted (micro canonical) version of the triplet model can also be constructed by fixing the number of triplets in each state and assigning to all such configurations equal probability. Most results obtained below can be generalized to this case in a straightforward manner.

---

[5] $\left( \begin{array}{c} N \\ n_1,n_2 \ldots n_k \end{array} \right) = \frac{N!}{n_1! n_2! \ldots n_k!}$ is the multinomial coefficient. Whenever used it implies $n_1 + n_2 \cdots + n_k = N$.

### 1. The Weight and Degree Distributions

The first quantity of interest is $P(A_{ij} = m)$, the probability that a randomly chosen pair i-j, of nodes is connected by an edge of weight m. For an edge to have weight m, exactly m of the N-2 triplets that contain i and j have to be in one of the states 100, 110, 101 or 111 which contribute 1 to $A_{ij}$ and the remaining N-2-m of these triplets have to be in one of the states 000, 001, 010 or 011 which do not contribute to $A_{ij}$. Therefore the probability that a randomly chosen pair of nodes is connected by an edge of weight m is:

$$P(m) = C_m^{N-2}(p_1 + 2p_2 + p_3)^m(1 - (p_1 + 2p_2 + p_3))^{N-2-m} \tag{10}$$

To obtain the probability distribution for the degree of a randomly picked node one has to take in to account the states of all triplets that contain the node. There are $C_2^{N-1}$ such triplets. States 111 and 101 contribute two and states 100, 001, 011 and 110 contribute one to the degree of node i. While the remaining states 000 and 010 do not contribute to the degree. Let $n_2, n_1$ and $n_0$ be the number of triplets in states that contribute 2, 1 and 0 to the degree of node i respectively. Then $P(n_2, n_1, n_0)$ is a multinomial distribution and $P(k)$ is the sum of $P(n_2, n_1, n_0)$'s that satisfy $n_1 + 2n_2 = k$:

$$P(k) = \sum_{n_1+2n_2=k} \begin{pmatrix} C_2^{N-1} \\ n_0, n_1, n_2 \end{pmatrix} (p_0 + p_1)^{n_0}(2p_1 + 2p_2)^{n_1}(p_2 + p_3)^{n_2} \tag{11}$$

### 2. Subgraph Probabilities

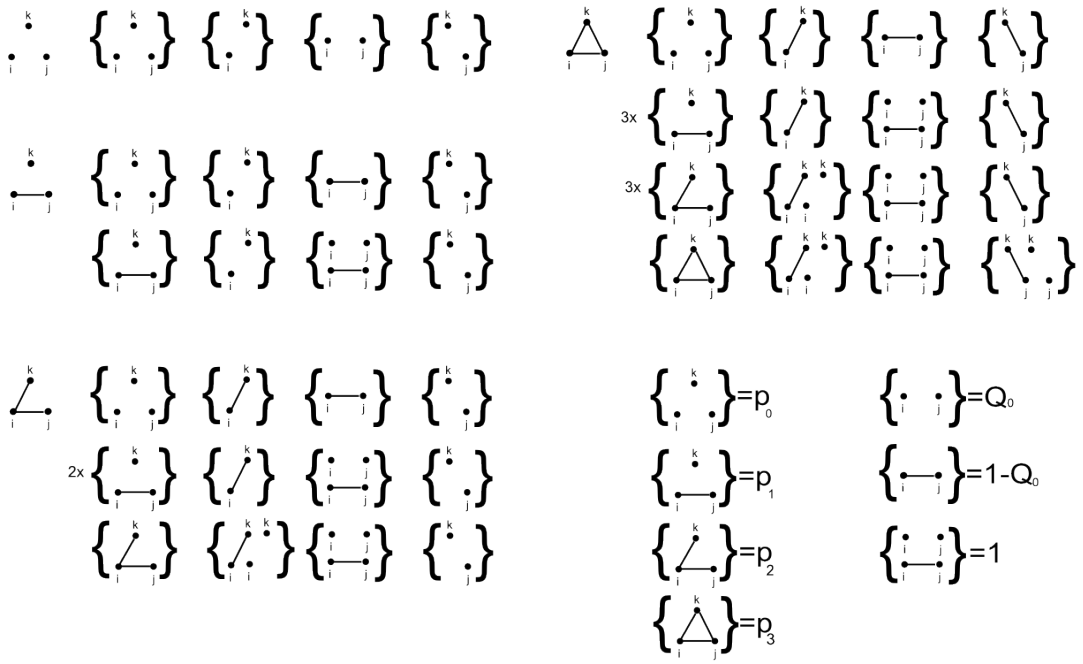


FIG. 2: The decomposition of the tree node subgraphs. The first column indicates the state of the triplet {i,j,k} and the other columns indicate the contributions of the other triplets that contain ik, ij and jk respectively. The probabilities corresponding to each of the factors are also given.

As stated in the introduction our main motivation for constructing our model is to obtain graphs with local structures that are not only tree like which should be reflected in the expectation values of subgraph counts. In this section we will calculate subgraph probabilities for the triplet model and show that the triplet model can produce sparse graphs with a large number of triangles.

To calculate the probability that a certain subgraph H of size m is induced on a set of nodes the states of all triplets that can contribute an edge between these nodes have to be taken into account. Here we will calculate induced

subgraph probabilities because the expressions are in general shorter and the ordinary subgraph probabilities can be obtained directly from these. Moreover when considering subgraphs we ignore edge weights and only consider patterns of attachment.

The general strategy while calculating the probability that a certain graph H is induced on a set of m nodes is to first fix the state of the triplets that are fully contained in the set. We call these base triplets. The base triplets can be in any combination of states that is compatible with H, that is none of the states should contain an edge where H has none[6]. Once the states of the base triplets are fixed, this imposes conditions on the triplets that only contain two nodes from the set. The first condition being that whenever the states of the base triplets do not contain an edge i-j where H actually has one this edge has to be present in at least one of the other triplets that contain nodes i and j while the second condition is that these triplets should not contain an edge where H has none. In this way given H the induced subgraph probability can decomposed into a sum of terms of which each corresponds to such a decomposition of H into states of the base triplets and a set of conditions on the triplets that contain only two nodes of H. Fig. 2 shows these decompositions for all tree node subgraphs. For instance for the graph induced on nodes {i,j,k} to be empty the triplet {i,j,k} has to in the state 000 and the states of the other triplets containing only two of the nodes should also not contain any edge between i-j, j-k or k-i.

The first subgraph probability of interest is the edge probability $P_e$. Each pair {i,j} of nodes is contained in N-2 triplets. For there to be no edge between i and j all the N-2 triplets have to be in one of the states 000, 010, 001 or 011. Therefore the edge probability is given by $P_e = 1 - (p_0 + 2p_1 + p_2)^{N-2} = 1 - (1 - (p_1 + 2p_2 + p_3))^{N-2}$

Now we will calculate the induced subgraph probabilities of 3 node subgraphs of which the decompositions are given in Fig.2. Once the state of the base triplet is fixed we have to calculate the possible contributions from the triplets that only contain two of the nodes. For each pair there are N-3 such triplets. As in the calculation of the edge probability, the probability that these N-3 triplets do not produce an edge between these nodes is $Q_0 = (1 - (p_1 + 2p_2 + p_3))^{N-3}$. Conversely the probability that these triplets produce an edge is $1 - Q_0$. Then, using the decomposition of the tree node subgraphs given in Fig.2, the subgraph probabilities for the three node subgraphs are:

$$\begin{aligned} P_{\therefore} &= p_0 Q_0^3 \\ P_{\text{-}} &= 3[p_0 Q_0^2 (1 - Q_0) + p_1 Q_0^2] \\ P_{\vee} &= 3[p_0 Q_0 (1 - Q_0)^2 + 2p_1 Q_0 (1 - Q_0) + p_2 Q_0] \\ P_{\triangle} &= p_0 (1 - Q_0)^3 + 3p_1 (1 - Q_0)^2 + 3p_2 (1 - Q_0) + p_3 \end{aligned} \tag{12}$$

The factors of three in the expressions for the single edged and V-shaped come from the symmetry of these subgraphs. The given probabilities are equivalent to the expectation values of the indicator functions for the occurrence of induced subgraphs and therefore the ensemble averages for the induced subgraph counts can be directly obtained by multiplying the above probabilities with $C_3^N$ . To calculate probabilities for a subgraph H of size m the same method can be applied the only difference being that one has to consider the states of $C_3^m$ base triplets. For large subgraphs the number of combinations compatible with H can be quite large but in the case of sparse graphs the number of configurations that have to be taken into account reduces significantly.

### *3. The Sparse Triplet Model*

The majority of real world networks are sparse, that is the number of edges is of the same order as the number nodes[1, 2]. For the triplet model to be sparse the expectation value of the number of edges (including edge weights), has to be of order N:

$$< e >= C_3^N (3p_1 + 6p_2 + 3p_3) \sim O(N) \tag{13}$$

This condition is equivalent to the $p_i$'s being of order $1/N^2$. We set $p_i = \beta_i / N^2$ (i=1,2,3) for convenience. In the sparse case most expressions can be expanded in orders of $1/N$ and in the large graph limit (fixing $\beta_1, \beta_2$, $\beta_3$ and $N_{\to\infty}$) multinomial distributions can be approximated by combinations of Poisson distributions. For instance the weight and degree distributions can be approximated by:

$$P(m) \simeq \left(\frac{\kappa}{N}\right)^m + O(N^{-m-1}) \tag{14}$$

---

[6] Ordinarily subgraph probabilities can be calculated by lifting this condition.

where $\kappa \simeq \beta_1 + 2\beta_2 + \beta_3$ is the average degree.This shows that multiple edges are quite rare when the model is sparse. Moreover in the sparse case the degree distribution can be approximated by the following combination of Poisson distributions:

$$P(k) \simeq \sum_{n_1+2n_2=k} \frac{e^{-\lambda_1}\lambda_1^{n_1}}{n_1!}\frac{e^{-\lambda_2}\lambda_1^{n_2}}{n_2!} \tag{15}$$

Where $\lambda_1 = \beta_1 + \beta_2$ and $\lambda_2 = (\beta_2 + \beta_3)/2$.Which shows that the degree distribution decreases no faster than $(k/2)!^{-1}$ which is considerably slower compared to $k!^{-1}$ in the Erdös-Rényi model. Moreover the equation also shows that graphs that have the same degree distribution but different local structures can be produced by the model. Similarly, the subgraph probabilities can be expanded in orders of $N^{-1}$.In the sparse case $1 - Q_0 \simeq \kappa/N$ and the probabilities of the three node subgraphs can be expanded up to order $N^{-3}$:

$$\begin{array}{l} P_{\cdot\!\cdot\!\cdot} = 1 - \frac{3\beta_1+6\beta_2+3\beta_3}{N} + \frac{9}{2}\frac{(\beta_1+2\beta_2+\beta_3)^2}{N^2} \\ P_{\cdot\!-} = 3[\frac{\beta_1+2\beta_2+\beta_3}{N} + \frac{\beta_1}{N^2} - \frac{5}{2}\frac{(\beta_1+2\beta_2+\beta_3)^2}{N^2}] \\ P_{\vee} = 3[\frac{\beta_2}{N^2} + \frac{(\beta_1+2\beta_2+\beta_3)^2}{N^2}] \\ P_{\triangle} = \frac{\beta_3}{N^2} \end{array} \tag{16}$$

Starting from these probabilities one can define the clustering for the ensemble, to be 3 times the expectation value of the number of triangles divided by the expectation number of connected triples in the ensemble.[7]

$$C = \frac{3\times < n_{\triangle} >}{< n_{\vee} >} = \frac{3P_{\triangle}}{3P_{\triangle} + P_{\vee}} \tag{17}$$

In the sparse case the the above formulas can be approximated as follows:

$$C \simeq \frac{\beta_3}{\beta_3 + \beta_2 + (\beta_1 + 2\beta_2 + \beta_3)^2} \leq \frac{1}{1+\kappa} \tag{18}$$

Where $\kappa = \beta_1 + 2\beta_2 + \beta_3$ is the average degree. The last inequality shows that there is a limit to the clustering that can be obtained by randomly adding triangles to a sparse graph. In order to obtain higher values for the clustering one has to consider models that use larger, more densely connected building blocks such as complete graphs. In the sparse case the probabilities for larger subgraphs are easy to calculate up to leading order. The inclusion of each non empty base triplet adds a factor of $1/N^2$ and each edge that is not generated by the states of the base triplets adds a factor of $1 - (1 - \kappa/N^2)^{N-m} \simeq \kappa/N$ to the probability. Therefore the leading order terms will be produced by combinations of base triplet states that cover large parts of the subgraph with the minimum number of non empty base triplet states.For instance the probability for the three star and 3 chain in the triplet model can be calculated in this way:

$$\begin{array}{l} P_{3*} = 4\left[\frac{\kappa^3}{N^3} + 3\frac{\kappa(\beta_2+\beta_3)}{N^3}\right] \\ P_{\sqcup} = 12\left[\frac{\kappa^3}{N^3} + 2\frac{\kappa(\beta_2+\beta_3)}{N^3}\right] \end{array} \tag{19}$$

The first terms in the expressions correspond to the case when all 4 base triplets are in the empty state and the second terms to the case when two of the edges are produced by one of the 4 base triplets while the other are empty. We omit the cases when more than one of the base triplets are in nonempty states since their contribution will be of order $1/N^4$ or lower. These together with the 3 node subgraph probabilities can be used to calculate the the assortativity [8]:

$$r^2 = \frac{3n_{\triangle} + n_{\sqcup} - \frac{n_{\vee}^2}{n_{\cdot\!\cdot\!\cdot}}}{n_{\vee} + 3n_{3*} - \frac{n_{\vee}^2}{n_{\cdot\!-}}} = \frac{\beta_3\beta_1 - \beta_2^2}{(\beta_1+\beta_2)(\beta_2+\beta_3) + (\beta_1+2\beta_2+\beta_3)^3 + (\beta_1+2\beta_2+\beta_3)^2(\beta_2+\beta_3)} \tag{20}$$

Which shows that depending on the parameters the triplet model can generate graphs with both positive and negative assortativity. As for the clustering coefficient, upper and lower bounds for the assortativity can also derived in terms of the average degree $\kappa$. The above formula can also be interpreted in terms of the effects of subgraphs on the degree

---

[7] Note that this is different from calculating the average of the clustering coefficient.

of the nodes that they connect. For instance the V shaped subgraph connects a a central node contributing 2 to its degree with two other nodes to which it only contributes one edge. Therefore on average it will connect a high degree central node to two lower degree peripheral nodes. Thus the addition of V shaped subgraphs will decrease the assortativity. On the other hand the single edge and triangle subgraphs contribute equally to the degree of the nodes they connect therefore they fail to produce any effect on the assortativity on their own but when combined the triangles will produce high degree nodes that the single edged states connect resulting in an increased assortativity.

#### *4. The Connected Component Phase Transition*

One of the most interesting results for the Erdös-Rényi random graph is the emergence of a giant connected component at a critical probability $pN = \kappa = 1$ as N tends to infinity. Here following a heuristic argument we derive the condition for a giant component to emerge in the triplet model. For this we assume that a fraction u of the nodes is not contained in the giant component. Then consistency requires that if a node is not in the giant component the states of the triplets containing the node should connect it only to nodes that are also not in the giant component. States 000, 010 do not connect the node i to any nodes while 100 and 001 connect it to one and 110,011,101 and 111 connect it to two nodes. Therefore the probability distribution $P(t_1, t_2)$ that $t_1$ triplets connect the node to one and $t_2$ to two nodes as N tends to infinity has to satisfy the following condition:

$$u = \sum_{t_1,t_2}^{\infty} P(t_1, t_2) u^{t_1+2t_2} \tag{21}$$

Where:

$$P(t_1, t_2) = \begin{pmatrix} C_2^{N-1} \\ t_0, t_1, t_2 \end{pmatrix} (p_0 + p_1)^{t_0} (2p_1)^{t_1} (3p_2 + p_3)^{t_2} \tag{22}$$

for large N this can be approximated as:

$$P(t_1, t_2) \simeq \frac{e^{-\frac{\beta_1}{2}} \left(\frac{\beta_1}{2}\right)^{t_1}}{t_1!} \frac{e^{-\frac{3\beta_2+\beta_3}{2}} \left(\frac{3\beta_2+\beta_3}{2}\right)^{t_2}}{t_2!} \tag{23}$$

Therefore the fraction S of nodes in the giant component is given by the solution of the equation:

$$1 - u = S = 1 - e^{-\beta_1(S) - \frac{3\beta_2+\beta_3}{2}(2S-S^2)} \tag{24}$$

For which a nonzero solution exists only if:

$$\beta_1 + 3\beta_2 + \beta_3 > 1 \tag{25}$$

This reduces to the condition $\kappa > 1$ for Erdös-Rényi graphs when $\beta_2, \beta_3 \simeq 0$ (see next section) and when $\beta_2 = 0$ to the result obtained in [11]for graphs where the number of triangles and edges connected to a node is distributed according to a product of Poisson distributions.

#### *5. The triplet model equivalent to the Erdös Rényi random graph*

In order to obtain a triplet model equivalent to the Erdös Rényi model we assume that edges in the triplet states are independent and occur with probability p. Then the probabilities for the triplet states are $p_0 = (1-p)^3, p_1 = p(1-p)^2, p_2 = p^2(1-p)$ and $p_3 = p^3$. Because the edges are already independent in the state space of the triplet model edges also occur independently with probability $p' = 1 - (1-p)^{N-2}$ in the graph.Consequently, the model reduces to the Erdös Rényi model when edge weights are set to one.

### B. The Directed Triplet Model

To generalize the triplet model to the directed case the triplet state space has to be expanded to include all directed 3 node graphs. The set of the directed states $S_{3\rightarrow} = \{00, 01, 10, 11\}^3$ contains 64 directed 3 node subgraphs that fall

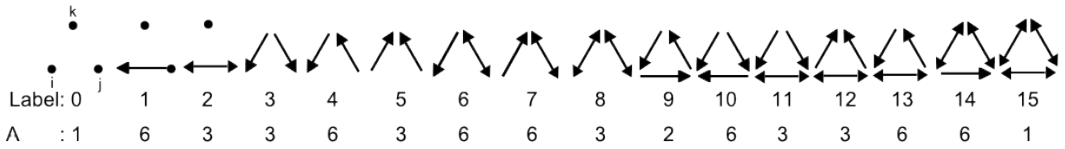


FIG. 3: The 16 isomorphism classes of the directed 3 node subgraphs and their graphical representations. The first row indicates the label of the classes while the second row indicates the number graphs in each class.

into 16 isomorphism classes[Fig.3]. As in the undirected case we assume that isomorphic states equiprobable and that the assignment of states to triplets is i.i.d. Then the resulting probability distribution over the configuration space $X = S_{3\rightarrow}^{C_3^N}$ is:

$$P(x) = \prod_{t \in T_3} p_{S(t,x)} = \prod_{i=0}^{15} p_i^{n_i(x)} \tag{26}$$

where $S(t,x)$ is the state of triplet t and the $n_i(x)$'s are the number of triplets that are in state i in the configuration x.

As in the undirected case we define a random variable $\Phi(x)$ that corresponds to the projection of configurations on to adjacency matrices. The projection $\Phi(x)$ is equivalent to adding an edge from node i to node j whenever such an edge is present in the state of a triplet in the configuration x:

$$\Phi : X \longrightarrow A^{N\times N}, \qquad \Phi(x)_{\alpha\beta} = \sum_{t \in T_3 | \alpha,\beta \in t} A_{\alpha\beta}[S(t,x)] \tag{27}$$

where $A[S(t,x)]$ is the adjacency matrix of the state of the triplet t in the configuration x. Methods used in the undirected case can then be generalized to the directed case in a straightforward manner.

For instance to calculate the degree distribution $P(k_{in}, k_{out})$ the contribution of each triplet state to degree of node i has to be considered. Let $m_{I,O}$ and $q_{I,O}$ denote the number of triplets states that contribute I incoming edges and O outgoing edges and the probabilities to be in such states, respectively. Then the degree distribution is given by:

$$P(k_{in}, k_{out}) = \sum \begin{pmatrix} C_2^{N-1} \\ m_{00}, m_{01}, m_{10}, m_{11}, m_{12}, \\ m_{21}, m_{20}, m_{02}, m_{22} \end{pmatrix} q_{00}^{m_{00}} q_{01}^{m_{01}} q_{10}^{m_{10}} q_{11}^{m_{11}} q_{12}^{m_{12}} q_{21}^{m_{21}} q_{02}^{m_{02}} q_{20}^{m_{20}} q_{22}^{m_{22}} \tag{28}$$

Where the sum is performed over the all $m_{I,O}$ that satisfy the following conditions:

$m_{10} + m_{11} + m_{12} + 2(m_{20} + m_{21} + m_{22}) = k_{in}$

$m_{01} + m_{11} + m_{21} + 2(m_{02} + m_{12} + m_{22}) = k_{out}$

The $q_{I,O}$ can be inferred from Fig[3]:

$q_{00} = p_0 + 2p_1 + p_2$

$q_{10} = 2p_1 + 2p_3 + 2p_4 + 2p_6$

$q_{20} = p_5 + 2p_{10} + p_{12}$

$q_{01} = 2p_1 + 2p_5 + 2p_4 + 2p_7$

$q_{02} = p_3 + 2p_{10} + p_{11}$

$q_{11} = 2p_2 + 2p_4 + 2p_6 + 2p_7 + 2p_8 + 2p_9 + 2p_{10} + 2p_{13}$

$q_{21} = 2p_7 + 2p_{11} + 2p_{13} + 2p_{14}$

$q_{12} = 2p_7 + 2p_{12} + 2p_{13} + 2p_{14}$

$q_{22} = p_8 + 2p_{14} + p_{15}$

One feature of the degree distribution is that depending on the state probabilities it can generate degree distributions that are not symmetric with respect to the in and out degree as well as degree distributions where the in and out degrees are correlated. The degree distribution of in, out and mutual edges which can be calculated in the same way.

### *1. Subgraph Probabilities*

The method for obtaining the probability that a certain graph is induced on a set of nodes in the directed case is essentially the same as in the undirected case. For instance to calculate the subgraph probabilities for the three node

subgraphs, all states of the base triplet$\{i, j, k\}$ that are compatible with the induced subgraph have to be considered. Once the state of the base triplet is fixed this imposes conditions on the states of the triplets that contain only two of the nodes in the triplet. One example of such a decomposition is given in Fig. 4.

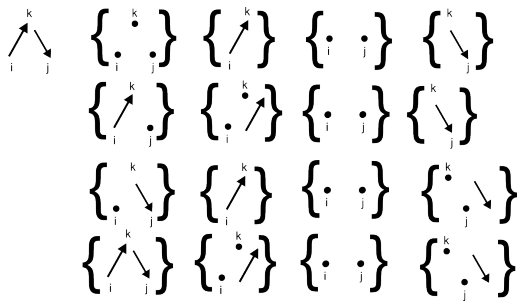

FIG. 4: The decomposition of for subgraph 4. The first column indicates the state of the triplet $\{i, j, k\}$ while the other columns indicate the conditions on the contributions of the triplets containing ik, ij and jk only.

The possible contributions from the N-3 triplets containing only two of the nodes and the corresponding probabilities are: no edge $r_0 = q_0^{N-3}$, a directed edge i to j or no edge $r_{01} = (q_0 + q_1)^{N-3}$, a directed edge i to j $r_1 = r_{01} - r_0$, a mutual edge $r_2 = 1 - 2r_{01} + r_0$ and a mutual edge or a directed edge from i to j $r_{12} = 1 - r_{01}$.

Where, $q_1 = (p_1 + p_3 + 2p_4 + p_5 + p_6 + p_7 + p_9 + 3p_{10} + p_{11} + p_{12} + 2p_{13} + p_{14})$ is the sum of the probabilities of the states that contain a directed edge from i to j and $q_2 = p_2 + 2p_6 + 2p_7 + 2p_8 + p_{11} + p_{12} + 2p_{13} + 4p_{14} + p_{15}$ is the sum of the probabilities of the states that contain a mutual edge between i and j. Then probability that the state does not contain any edge at between i and j is $q_0 = 1 - 2q_1 - q_2$.

From the decomposition in Fig.4 one can then find that the probability that motif 4 is induced on a set of three nodes is :$P_4 = p_0 r_0 r_1^2 + 2p_1 r_0 r_{01} r_1 + p_4 r_0 r_{01}^2$. Probabilities for all other three node directed subgraphs can be obtained using the same method[8]. In the case of sparse graphs ( $p_i = \beta_i/N^2$ for$i = 1, 2, ...15$) $r_1$ and $r_2$ become of order $N^{-1}$. Then probabilities of the V shaped subgraphs up to order $N^{-2}$ have the general form of $P_{\vee i} \simeq \frac{\beta_i}{N^2} + r_1^{e(i)} r_2^{m(i)}$ where e(i) is the number of directed edges and m(i) is the number of mutual edges in subgraph i. The probabilities up to leading order for the triangle shaped subgraphs are: $P_{\triangle i} \simeq \frac{\beta_i}{N^2}$.Therefore provided that the probabilities corresponding to triangular building blocks are of order $N^{-2}$ the expectation values of the corresponding subgraph counts will be of order N and therefore when compared to the configuration model using the method of [5] the triplet model will have triangular motifs. Since the configuration model with the same degree distribution as the sparse triplet model has expectation values of order 1 for triangular motif counts [12].

## IV. THE MULTIPLET MODEL

The multiplet model is a generalization of the triplet model (and of the Erdös-Rényi model) that allows the construction of graphs using building blocks of arbitrary size. Using larger building blocks the multiplet model can generate more densely connected local structures than the triplet model and also patterns of attachment between network motifs.

The multiplet model ($M_n^N$) of order n (corresponding to the size of the building blocks) on N nodes is based on considering n-tuples of nodes and their states. The state space for n-tuples $t \in T_n^N = \{\{i_1, i_2, ..., i_n\} : i_1, i_2, ..., i_n \in V\}$ corresponds to all n node graphs that can be realized on n nodes. Therefore the most general state spaces are $S_n = \{0, 1\}^{C_2^n}$ in the undirected and $S_{\to n} = \{00, 01, 10, 11\}^{C_2^n}$ in the directed case. In analogy with the Erdös-Rényi model the states of n-tuples are assumed to be identically distributed and independent. Isomorphic states are assigned equal probabilities in order to conserve the symmetry between nodes. Then the resulting probability distribution over the configuration space $X_n^N = S_n^{C_n^N}$ is:

$$P(x) = \prod_{t \in T_n^N} p_{S(t,x)} \tag{29}$$

[8] A list of al the 3 node induced subgraph probabilities can be found in the appendix.

Where $S(t,x)$ is the state of the n-tuple t in the configuration x and the $p_i$'s are the state probabilities. As in the triplet model a configuration $x \in X$ is projected on to a graph by adding an edge to the graph whenever the edge is present in the state of an n-tuple:

$$\Phi : X \longrightarrow A^{N\times N}, \qquad \Phi(x)_{\alpha\beta} = \sum_{t\in T_n^N | \alpha,\beta\in t} A_{\alpha\beta}[S(t,x)] \tag{30}$$

where $A[S(t,x)]$ is the entry of the adjacency matrix of the state of t in the configuration x.

In the multiplet models calculations are done considering all configurations $x \in X$ that result in the desired graph property. Consequently, the generalization of most of the methods used in the triplet model to higher order models is straightforward.

For instance the degree distribution for an undirected model in the class $M_n^N$ is:

$$P(k) = \sum_{(n-1)m_{n-1}+(n-2)m_{n-2}+\cdots 2m_2+m_1=k} \left( \begin{array}{c} C_{n-1}^{N-1} \\ m_0, m_1 \ldots, m_{n-1} \end{array} \right) q_{n-1}^{m_{n-1}} q_{n-2}^{m_{n-2}} \cdots q_1^{m_1} q_0^{m_0} \tag{31}$$

Where the $m_i$'s are the number of states that contribute i to the degree and the $q_i$'s are the total probabilities of such n-tuple states. In the directed case a similar expression for the in and out degree distribution can be obtained by replacing the $q_i$'s and $m_i$'s with the corresponding $q_{I,O}$ and $m_{I,O}$'s as in the directed triplet model and summing over all configurations that produce the desired in and out degrees. Moreover the above formula shows that using the multiplet model graphs that have the same degree distribution but differ significantly with respect to their local structures can be generated. For instance in the undirected case with n=5 there are 4 equations for the $q_i$'s while there are 33 parameters corresponding to the probabilities of the 34 isomorphism classes.

In order to calculate subgraph probabilities in the multiplet model $M_n^N$ for a subgraph H of size m the subgraph has to be decomposed into the contributions coming from the n-tuples that contain k,k-1,...,3 and 2 of its nodes (k=sup[n,m]). In this decomposition once the states of the n-tuples that contain more than two of the nodes are fixed the edges of H that are not produced by the state of these n-tuples have to be be present in at least one of the n-tuples that contain only 2 of the nodes. In the case of induced subgraph probabilities one has to restrict the states of the n-tuples so that they do not contribute an edge where where H has none.

### A. The Sparse Multiplet Models

For a multiplet model $M_n^N$ to be sparse the $p_i$'s have to be of order $N^{1-n}$ or less. Then for sufficiently large N $n \ll N$ so that $C_N^n \simeq N^n/n!$ the degree distribution can be approximated by a product of Poisson distributions:

$$P(k) \simeq \sum_{(n-1)m_{n-1}+(n-2)m_{n-2}+\cdots 2m_2+m_1=k} \prod_{j=0}^{n-1} \frac{e^{-\lambda_j}\lambda_j^{m_j}}{m_j!} \tag{32}$$

Where $\lambda_i = q_i C_{n-1}^{N-1}$.The above equation shows that the degree distribution decreases as $(k/(n-1))!^{-1}$ for large k. Moreover for $k < n$, P(k) is a function of the $\lambda_i$'s up to i=k only and therefore depending on the $\lambda_i$'s P(k) can fit broad degree distributions in the range $k < n$, including power law type degree distributions with a cutoff near n.

Although in the multiplet model subgraph probabilities for large subgraphs can still be calculated exactly, in the sparse case expressions up to leading order simplify significantly. Consequently, one can obtain expressions for the assortativity and clustering as in the triplet model by calculating the subgraph probabilities.

The results for the emergence of the connected component in the triplet model can also be generalized to multiplet models by considering how many nodes the states in $S_n$ connect to each other.

### B. The Hierarchy of Multiplet Models and Complexity

As there exists a triplet model that is equivalent to the Erdös-Rényi model, for most of the models in a class $M_n^N$ there exists an equivalent higher order model in the class $M_{n+1}^N$. Consequently, there exists a nested hierarchy between multiplet models ordered according to the size of building blocks used so that for every lower order random graph model in a certain subclass, a model of higher order that is equivalent can be found. Within this hierarchy of increasing complexity graph properties traditionally associated with higher complexity such as assortativity, clustering, heavy tailed degree distributions and highly connected network motifs are indeed indicators of higher complexity. The precise form of the hierarchy and its connection to different complexity measures is beyond the scope of this paper and will be treated elsewhere.

## V. DISCUSSION

We presented a model for constructing random graphs using small graphs as building blocks that can be solved exactly for many of its properties.The model provides a method for introducing correlations between edges in a systematic way.In the simplest case, corresponding to the triplet model, we showed that in the undirected case graphs with non zero clustering and assortativity and in the directed case graphs with triangular motifs can be obtained.Then we generalized the model to include building blocks of arbitrary size. We showed that multiplet models of of higher order are able to generate graphs with heavy tailed degree distributions and a multitude of local structures. The multiplet model can be used to investigate how network properties like motifs, clustering, assortativity and the degree distribution are interrelated. The effects of motifs on various processes taking place on networks, including diffusion, percolation, synchronization and information processing can also be studied within the framework of the multiplet model.

We showed that properties that are considered to be characteristics of complex networks are indeed indicators that these networks are better modeled by multiplet models of higher order and thus higher complexity. Moreover the model can generate networks that have the same degree distribution but vary significantly with respect to their local structures. This diversity of local structures increases as the degree distribution gets broader which indicates that networks with broad degree distributions have a higher the capacity to adapt their local structure and might be a hint to why evolving real world networks often have heavy tailed degree distributions.

The multiplet model allows the construction of random graphs that can be called genuinely complex while still being analytically tractable. Although the calculations for higher order models become increasingly difficult to perform by hand their algorithmic implementation is straightforward. The only difficulty being that for higher order models there is no known practical way to partition the state space S in to isomorphism classes but one can still consider higher order models with state spaces less general than $S_n$.

Real world networks that have an underlying bipartite structure such as collaboration networks can be modeled naturally by multiplet models. For instance in collaboration networks complete graphs of size L would correspond to independent occurrences of associations (movie casts, executive boards, scientific publications etc.) of size L. While other building blocks would correspond to preferred attachment patterns between these.

The model can be generalized to cases where the nodes are be partitioned into different classes and the connections within and between these sets follow different rules.This might be useful in modeling networks that have components with different local structures. Another possibility would be the case when the set of nodes is a metric space. In this case the set of n-tuples might be restricted to sets of nodes that are within a certain distance of each other and/or the probabilities might be assumed be functions of distance. Such additional structure might be needed in order to model accurately real networks in which distance plays a role.

### Appendix A: Directed 3-node subgraph probabilities

The induced subgraph probabilities for the 16 directed three node subgraphs in the triplet model:

$P_0 = p_0 r_0^3$
$P_1 = 6[p_1 r_0^2 r_{01} + p_0 r_1 r_0^2]$
$P_2 = 3[p_2 r_0^2 + 2p_1 r_{12} r_0^2 + p_0 r_0^2 r_2]$
$P_3 = 3[p_3 r_0 r_{01}^2 + 2p_1 r_0 r_{01} r_1 + p_0 r_0 r_1^2]$
$P_4 = 6[p_4 r_0 r_{01}^2 + 2p_1 r_0 r_{01} r_1 + p_0 r_0 r_1^2]$
$P_5 = 3[p_5 r_0 r_{01}^2 + 2p_1 r_0 r_{01} r_1 + p_0 r_0 r_1^2]$
$P_6 = 6[p_6 r_0 r_{01} + (p_3 + p_4) r_0 r_{12} r_{01} + p_2 r_0 r_1 + p_1 (r_2 r_{01} + 2 r_1 r_{12}) + p_0 r_2 r_0 r_1]$
$P_7 = 6[p_7 r_0 r_{01} + (p_5 + p_4) r_0 r_{12} r_{01} + p_2 r_0 r_1 + p_1 (r_2 r_{01} + 2 r_1 r_{12}) + p_0 r_2 r_0 r_1]$
$P_8 = 3[p_8 r_0 + 2(p_6 + p_7) r_0 r_{12} + (p_3 + p_5 + 2p_4) r_0 r_{12}^2 + 2p_2 r_0 r_2 + 4p_1 r_0 r_{12} r_2 + p_0 r_2^2]$
$P_9 = 2[p_9 r_{01}^3 + 3p_4 r_{01}^2 r_1 + 3p_1 r_1^2 r_{01} + p_0 r_1^3]$
$P_{10} = 6[p_{10} r_{01}^3 + (p_3 + p_4 + p_5) r_{01}^2 r_1 + 3p_1 r_1^2 r_{01} + p_0 r_1^3]$
$P_{11} = 3[p_{11} r_{01}^2 + 2p_7 r_1 r_{01} + p_3 r_2 r_{01}^2 + 2p_1 (r_{01} r_1 r_2 + r_{12} r_1^2) + p_2 r_1^2 + p_0 r_1^2 r_2 + 2p_{10} r_{01}^2 r_{12} + 2(p_5 + p_4) r_{01} r_{12} r_1]$
$P_{12} = 3[p_{12} r_{01}^2 + 2p_6 r_1 r_{01} + p_5 r_2 r_{01}^2 + 2p_1 (r_{01} r_1 r_2 + r_{12} r_1^2) + p_2 r_1^2 + p_0 r_1^2 r_2 + 2p_{10} r_{01}^2 r_{12} + 2(p_3 + p_4) r_{01} r_{12} r_1]$
$P_{13} = 6[p_{13} r_{01}^2 + (p_{10} + p_9) r_{01}^2 r_{12} + (p_6 + p_7) r_1 r_{01} + (p_3 + p_5) r_{01} r_1 r_{12} + p_4 (2 r_1 r_{01} r_{12} + r_{01}^2 r_2) + p_2 r_1^2 + 2p_1 (r_{01} r_1 r_2 + r_1^2 r_{12}) + p_0 r_2 r_1^2]$

$P_{14} = 6[p_{14}r_{01}+(p_{12}+p_{11}+2p_{13})r_{01}r_{12}+p_9 r_{01}r_{12}^2+2p_{10}r_{01}r_{12}^2+p_8 r_1+(p_6+p_7)(r_2 r_{01}+2r_1 r_{12})+(p_3+p_5)(r_{01}r_{12}r_2+r_{12}^2 r_1)+p_4(2r_1 r_{12}^2+2r_2 r_{01}r_{12})+2p_2 r_2 r_1+p_1(4r_2 r_1 r_{12}+r_{01}r_2^2)+p_0 r_2^2 r_1]$

$P_{15} = p_{15}+6p_{14}r_{12}+(6p_{13}+3p_{12}+3p_{11})r_{12}^2+(2p_9+6p_{10})r_{12}^3+3p_8 r_2+(6p_7+6p_6)r_2 r_{12}+(6p_4+3p_5+3p_3)r_2 r_{12}^2+3p_2 r_2^2+6p_1 r_2^2 r_{12}+p_0 r_2^3$


[1] M.E.J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[2] R. Albert and A.L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[3] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538, 2004.

[4] P. Erdos and A. Rényi. {On the evolution of random graphs}. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[5] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824, 2002.

[6] P.W. Holland and S. Leinhardt. Transitivity in structural models of small groups. *Comparative Group Studies*, 1971.

[7] D.J. Watts and S.H. Strogatz. Collective dynamics of small-worldnetworks. *Nature*, 393(6684):440–442, 1998.

[8] E. Olbrich, T. Kahle, N. Bertschinger, N. Ay, and J. Jost. Quantifying structure in networks. *European Physical Journal B*, 77:239–247, 2010.

[9] B. Bollobás, S. Janson, and O. Riordan. Sparse random graphs with clustering. *Random Structures & Algorithms*, 38(3):269–323, 2011.

[10] B. Bollobás. *Random graphs*, volume 73. Cambridge Univ Pr, 2001.

[11] M.E.J. Newman. Random graphs with clustering. *Physical review letters*, 103(5):58701, 2009.

[12] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon. Subgraphs in random networks. *Physical Review E*, 68(2):026127, 2003.