Network motifs: mean and variance for the count

Catherine Matias, Sophie S. Schbath, Etienne Birmelé, Jean-Jacques J.-J. Daudin, Stephane S. Robin

REVSTAT – Statistical Journal
Volume 4, Number 1, March 2006, 31–51

NETWORK MOTIFS: MEAN AND VARIANCE FOR THE COUNT

Authors: C. Matias
– UMR CNRS 8071, Laboratoire Statistique et Génome,
91000 Evry, France
matias@genopole.cnrs.fr

S. Schbath
– INRA, Unité Mathématique, Informatique et Génome,
78352 Jouy-en-Josas, France
Sophie.Schbath@jouy.inra.fr

E. Birmelé
– UMR CNRS 8071, Laboratoire Statistique et Génome,
91000 Evry, France
birmele@genopole.cnrs.fr

J.-J. Daudin
– UMR ENGREF/INAPG/INRA,
Mathématiques et Informatique Appliquées, INA P-G, 75231 Paris, France
daudin@inapg.inra.fr

S. Robin
– UMR ENGREF/INAPG/INRA,
Mathématiques et Informatique Appliquées, INA P-G, 75231 Paris, France
robin@inapg.inra.fr

Abstract:

- Network motifs are at the core of modern studies on biological networks, trying to encompass global features such as small-world or scale-free properties. Detection of significant motifs may be based on two different approaches: either a comparison with randomized networks (requiring the simulation of a large number of networks), or the comparison with expected quantities in some well-chosen probabilistic model. This second approach has been investigated here. We first provide a simple and efficient probabilistic model for the distribution of the edges in undirected networks. Then, we give exact formulas for the expectation and the variance of the number of occurrences of a motif. Generalization to directed networks is discussed in the conclusion.

Key-Words:

- *network motif; motif count; random graph; sequence of degrees.*

AMS Subject Classification:

- 62P10; 62E15; 05C80.

1. INTRODUCTION

A cellular system can be described by a web of relationships between proteins, genes or more generally metabolites. Studying its basic structural elements, also called **motifs**, is a first step in the understanding of these networks that goes beyond global features (such as the small world or scale-free properties, see [2, 12]). For instance, motifs that occur more frequently than expected in random networks may reveal particular structures corresponding to biological phenomena. Several definitions exist for a network motif. Here we consider the most commonly used: a simple pattern of interconnection in a graph. Detection of significant motifs [7] may be based on two different approaches: either by comparing the observed network with appropriately randomized networks (this requires the simulation of a large number of networks), or by the comparison with expected quantities in some well-chosen probabilistic model. Up to now, only the first approach has been explored ([8], [11], [13]) because no satisfactory probabilistic model has yet been proposed for an analytical approach. The simplest model is the well-known Erdös model, where the probability of appearance of an edge between two different vertices is equal to some fixed $p \in (0, 1)$. This model only concerns undirected networks. Its major drawback lies in the fact that the numbers of edges per vertex, so-called vertex degrees, are distributed according to a Binomial distribution, generally approximated by a Poisson distribution, whereas biological networks appear to be scale-free, meaning a power law for the number of edges per vertex [1] (for more details on random graphs, we refer to [4, 6, 5]). Randomized networks (obtained by simulation, see [10] for instance) rely on the knowledge of the number of (incoming and outcoming, when dealing with directed graphs) edges for each vertex. In the same spirit, we provide a probabilistic model that fits these vertex degrees. Depending on the specified sequence of edges per vertex, our model may describe scale-free networks. This probabilistic model enables us to derive exact formulas for the mean and variance of the number of occurrences of a motif, in a graph specified by a sequence of degrees. One of the advantages of this approach is that we do not need computationally expensive simulations of a large number of graphs, for each fixed sequence of numbers of edges per vertex.

Let us mention another approach developped in [3] where "groups of motifs" are detected using an heuristic algorithm based on a probabilistic model. The main difference between this approach and our work lies in the definition of a motif. Berg and Lässig's motifs are groups of vertices which are highly interconnected in a sparse graph, whereas we consider sets of inter-connected vertices with a given topology.

Section 2 presents the definitions of motifs and their occurrences. To decide whether a given motif $\mathbf{m}$ has an unexpected frequency in a given observed graph, one has first to consider random graphs having some similar properties with the observed graph (Section 3), and then to calculate the expected count of $\mathbf{m}$ in such

random graphs, and eventually its variance (Section 4). Since the derivation of the exact distribution of a motif count is still an open problem, its exact mean and variance can be used to calculate a z-score directly. This avoids heavy simulations used in the literature to evaluate the significance of motif counts [9]. Indeed, from our knowledge, current methods to assess significance of motif counts are based on a large number of simulations *for each* type of graph (namely, a fixed sequence of degrees). Our approach is simple to implement and leads to a generic procedure (valid for any type of graph).

2. MOTIFS AND OCCURRENCES

Recall that, in this paper, a motif $\mathbf{m}$ of size k is simply a connected subgraph with k vertices. We will essentially focus on undirected graphs and motifs, but the generalization to a directed framework will be discussed in the conclusion. Therefore, there are only two motifs of size 3 (triangle and "V") and six motifs of size 4 (see Figure 1).

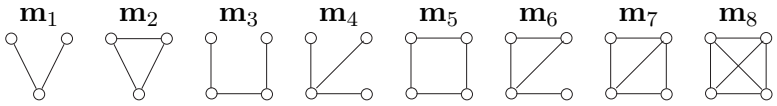

Figure 1: Motifs of size 3 and 4.

Let us fix an undirected graph G with N vertices labelled by $\{1, 2, ..., N\}$. I_k denotes the set of positions $\{i_1, i_2, ..., i_k\}$ in graph G where a motif of size k may occur. Namely, I_k is the set of all subsets of $\{1, 2, ..., N\}$ with cardinality k:

$$I_k = \left\{ \{i_1, i_2, ..., i_k\} \subset \{1, ..., N\}^k \text{ such that } i_j \neq i_\ell, \ \forall 1 \leqslant j \neq \ell \leqslant k \right\}.$$

In the same way, for any subset $J \subset \{1, ..., N\}$, define the sets of positions among the restricted number of vertices $\{1, ..., N\} \backslash J$,

$$I_k(J) = \left\{ \{i_1, i_2, ..., i_k\} \subset \left(\{1, ..., N\} \backslash J\right)^k \text{ such that } i_j \neq i_\ell, \ \forall j \neq \ell \right\}.$$

We say that a given motif $\mathbf{m}$ occurs at position $\alpha = \{i_1, i_2, ..., i_k\} \in I_k$ in G if and only if the sub-graph with vertices $\{i_1, i_2, ..., i_k\}$ in G either has the same topology as $\mathbf{m}$, or contains a subgraph with the same topology as $\mathbf{m}$. For instance, the triangle (motif $\mathbf{m}_2$ from Figure 1) occurs once in the graph in Figure 2 (position $\{2, 3, 4\}$), and the "V" motif ($\mathbf{m}_1$ from Figure 1) occurs 5 times (3 times at position $\{2, 3, 4\}$, once at position $\{1, 2, 3\}$ and once at position $\{1, 2, 4\}$).

To define $N(\mathbf{m})$ the number of occurrences of $\mathbf{m}$ in a graph G, we introduce variables $Y_\alpha(\mathbf{m})$, $\alpha \in I_k$, defined as the number of occurrences of motif $\mathbf{m}$ in the sub-

graph with vertices α. Thus, for any motif of size k, we have $N(\mathbf{m}) = \sum_{\alpha \in I_k} Y_\alpha(\mathbf{m})$. If $\alpha = \{i_1, i_2, ..., i_k\}$, the variable $Y_\alpha(\mathbf{m})$ can be reformulated as $Y_{i_1,i_2,...,i_k}(\mathbf{m})$.

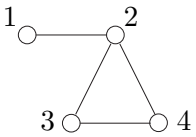

Figure 2: A graph containing 5 occurrences of motif $\mathbf{m}_1$ and 1 occurrence of motif $\mathbf{m}_2$.

3. RANDOM GRAPH MODEL

Undirected graphs are quite properly described by the sequence of the number of edges per node. Let us consider a graph G with N vertices labelled by $\{1, ..., N\}$ and a sequence of integers $(d_1, ..., d_N)$ such that $0 \leqslant d_i \leqslant N-1$. In practice, when analyzing a given graph, d_i is chosen as the observed degree of vertex i. We consider the following probabilistic model for graph G. Random variables Z_{ij} indicating presence/absence of an edge between vertices i and j $(i \neq j)$ are independent Bernoulli variables with mean π_{ij} (they are not identically distributed). Moreover, this probability π_{ij} of appearance of an edge between vertices i and j is related to the observed number of edges at node i and the observed number of edges at node j:

$$\pi_{ij} = \pi_{ji} = \frac{d_i d_j}{C} \qquad \text{and} \qquad \pi_{ii} = 0 \ .$$

C is a normalizing constant such that $\pi_{ij} \in [0, 1]$. For instance, $C = \max_{i \neq j} d_i d_j$. If the degrees are not too large with respect to the total number N of vertices, one may use $C_0 = \sum_{j=1}^{N} d_j(d_+ - d_j)/d_+$ with $d_+ = \sum_i d_i$. With such a choice, the expected number of edges is equal to the observed total number of edges. Moreover, the expected number of edges at node i is almost equal to d_i. Note that we do not allow direct loops from an edge to itself ($\pi_{ii} = 0$).

The advantage of this model is that its parameters are easy to choose from an observed graph, contrary to more general π_{ij}'s, and it almost fits the observed sequence of degrees when choosing C_0 as the normalizing constant. It relies on the same idea of preserving the sequence degrees as the commonly used simulation approach [8]. Our probabilistic model appears as a rigorous formalization of the simulation method. [8] suggest generating graphs that preserve the number of occurrences of all $(k-1)$-node sub-graphs when studying motifs of size k. Taking into account the counts of the $(k-1)$-node sub-graphs would be better than only preserving the sequence of degrees but such a generalization appears to be difficult at this stage.

4. FIRST AND SECOND MOMENTS FOR THE COUNT

Motifs of size 1 or 2 are of no interest here because they are the vertices and the edges, respectively, and their frequencies are set by the graph model. Let $\mathbf{m}$ be a motif of size $k \geqslant 3$. Since the variance of $N(\mathbf{m})$ is equal to $\mathbb{E}N^2(\mathbf{m}) - (\mathbb{E}N(\mathbf{m}))^2$, we will calculate the first and second moments of the count, i.e. $\mathbb{E}N(\mathbf{m})$ and $\mathbb{E}N^2(\mathbf{m})$. As we will see, these moments depend on $\mathbf{m}$, both through its size and its topology. No general formula is provided but we propose a general methodology that can be applied to any topological motifs without theoretical difficulties. Because of technical reasons, we will restrict ourselves to motifs of size 3 and 4. More precisely, for each motif $\mathbf{m}$, we provide a simple description of variable $Y_\alpha(\mathbf{m})$ using indicator random variables (RVs). This description enables us to derive explicit formulas for the moments $\mathbb{E}N(\mathbf{m})$ and $\mathbb{E}N^2(\mathbf{m})$. Before detailing the different cases, we state a common framework that will point out the basic quantities to calculate systematically.

Getting the expected count just requires the calculation of $\mathbb{E}Y_\alpha(\mathbf{m})$ for $\alpha \in I_k$ since we have

$$\mathbb{E}N(\mathbf{m}) = \sum_{\alpha \in I_k} \mathbb{E}Y_\alpha(\mathbf{m}) \ .$$

Getting the second moment is a little more involved. By definition,

$$\mathbb{E}N^2(\mathbf{m}) = \mathbb{E}\left(\sum_{\alpha \in I_k} Y_\alpha(\mathbf{m}) \times \sum_{\beta \in I_k} Y_\beta(\mathbf{m}) \right) = \sum_{\alpha \in I_k} \sum_{\beta \in I_k} \mathbb{E}\big(Y_\alpha(\mathbf{m})\, Y_\beta(\mathbf{m})\big) \ .$$

Let us break down the sums over α and β into $(k+1)$ sums depending on the cardinality of the intersection $\alpha \cap \beta$, denoted by $|\alpha \cap \beta|$. Note that

(i) when $|\alpha \cap \beta| = k$, then $\alpha = \beta$ and $\mathbb{E}(Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})) = \mathbb{E}Y_\alpha^2(\mathbf{m})$,

(ii) when $|\alpha \cap \beta| \leqslant 1$ (disjoint occurrences or a unique vertex in common), then $Y_\alpha(\mathbf{m})$ and $Y_\beta(\mathbf{m})$ are independent random variables, leading to $\mathbb{E}(Y_\alpha(\mathbf{m})\, Y_\beta(\mathbf{m})) = \mathbb{E}Y_\alpha(\mathbf{m})\, \mathbb{E}Y_\beta(\mathbf{m})$.

It gives

$$\begin{aligned} \mathbb{E}N^2(\mathbf{m}) \ = \ & \sum_{|\alpha\cap\beta|=0} \mathbb{E}\big(Y_\alpha(\mathbf{m})\big)\, \mathbb{E}\big(Y_\beta(\mathbf{m})\big) \ + \sum_{|\alpha\cap\beta|=1} \mathbb{E}Y_\alpha(\mathbf{m})\, \mathbb{E}Y_\beta(\mathbf{m}) \\ & + \sum_{2\leqslant n\leqslant k-1} \ \sum_{|\alpha\cap\beta|=n} \mathbb{E}\big(Y_\alpha(\mathbf{m})\, Y_\beta(\mathbf{m})\big) \ + \sum_{\alpha\in I_k} \mathbb{E}Y_\alpha^2(\mathbf{m}) \ . \end{aligned} \tag{4.1}$$

Additionally to quantities $\mathbb{E}Y_\alpha(\mathbf{m})$, we have to calculate terms in the form $\mathbb{E}(Y_\alpha(\mathbf{m})\, Y_\beta(\mathbf{m}))$ when α and β share between 2 and k elements. The next two subsections provide explicit formulas for motifs of size 3 and 4. The generic method is to write $Y_\alpha(\mathbf{m})$ as a sum of Bernoulli RVs whose expectations are straightforward to calculate.

4.1. Motifs of size 3

When $k = 3$, Equation (4.1) reduces to

$$\begin{aligned}\mathbb{E}N^2(\mathbf{m}) \;=\; & \sum_{\{i,j,k\}\in I_3} \sum_{\{\ell,u,v\}\in I_3(ijk)} \mathbb{E}Y_{i,j,k}(\mathbf{m})\,\mathbb{E}Y_{\ell,u,v}(\mathbf{m}) \\ & + \sum_{1\leqslant i\leqslant N} \sum_{\{j,k\}\in I_2(i)} \sum_{\{\ell,u\}\in I_2(ijk)} \mathbb{E}Y_{i,j,k}(\mathbf{m})\,\mathbb{E}Y_{i,\ell,u}(\mathbf{m}) \\ & + \sum_{\{i,j\}\in I_2} \sum_{k\in I_1(ij)} \sum_{\ell\in I_1(ijk)} \mathbb{E}\big(Y_{i,j,k}(\mathbf{m})\,Y_{i,j,\ell}(\mathbf{m})\big) \;+\; \sum_{\{i,j,k\}\in I_3} \mathbb{E}Y^2_{i,j,k}(\mathbf{m})\,.\end{aligned} \tag{4.2}$$

Motif $\mathbf{m}_1$ ("V")

Our approach is based on the split of variable $Y_{i,j,k}(\mathbf{m}_1)$ into the sum of three Bernoulli RVs

$$Y_{i,j,k}(\mathbf{m}_1) = Z_{ij,ik} + Z_{ij,jk} + Z_{ik,jk}\ , \qquad \forall\, i,j,k \in \{1,...,N\}\ ,$$

where $Z_{ij,ik} = 1$ if both edges ij and ik occur, and 0 otherwise. The expectation $\mathbb{E}Z_{ij,ik}$ is the probability $\pi_{ij}\pi_{ik}$. Thus we obtain

$$\begin{aligned}\mathbb{E}Y_{i,j,k}(\mathbf{m}_1) \;&=\; \pi_{ij}\pi_{ik} + \pi_{ij}\pi_{jk} + \pi_{ik}\pi_{jk} \;=\; \frac{d_i d_j d_k}{C^2}\,(d_i + d_j + d_k)\ , \\ \mathbb{E}N(\mathbf{m}_1) \;&=\; \sum_{\{i,j,k\}\in I_3} \frac{d_i d_j d_k}{C^2}\,(d_i + d_j + d_k) \;=\; \sum_{1\leqslant i\leqslant N} \sum_{\{j,k\}\in I_2(i)} \frac{d_i^2 d_j d_k}{C^2}\ .\end{aligned} \tag{4.3}$$

Similarly, we denote by $Z_{ij,ik,jk}$ the indicator RV of the presence of edges ij, jk and ik (note that $Z_{ij,ik}\,Z_{ij,jk} = Z_{ij,ik,jk}$). To calculate $\mathbb{E}(Y_{i,j,k}(\mathbf{m}_1)\,Y_{i,j,\ell}(\mathbf{m}_1))$, we write

$$\begin{aligned}&\mathbb{E}\big(Y_{i,j,k}(\mathbf{m}_1)\,Y_{i,j,\ell}(\mathbf{m}_1)\big) \;= \\ &\quad= \;\mathbb{E}\Big\{\big[Z_{ij,ik} + Z_{ij,jk} + Z_{ik,jk}\big]\,\big[Z_{ij,i\ell} + Z_{ij,j\ell} + Z_{i\ell,j\ell}\big]\Big\} \\ &\quad= \;\pi_{ij}(\pi_{ik} + \pi_{jk})\,(\pi_{i\ell} + \pi_{j\ell} + \pi_{i\ell}\pi_{j\ell}) \;+\; \pi_{ik}\pi_{jk}(\pi_{ij}\pi_{i\ell} + \pi_{ij}\pi_{j\ell} + \pi_{i\ell}\pi_{j\ell}) \\ &\quad= \;\frac{d_i d_j d_k d_\ell}{C^3}(d_i + d_j)^2 \;+\; \frac{d_i^2 d_j^2 d_k d_\ell}{C^4}\Big\{(d_i + d_j)(d_k + d_\ell) + d_k d_\ell\Big\}\ .\end{aligned} \tag{4.4}$$

Now, we focus on the term $\mathbb{E}Y^2_{i,j,k}(\mathbf{m}_1)$. We get

$$\begin{aligned}\mathbb{E}Y^2_{i,j,k}(\mathbf{m}_1) &= \mathbb{E}Z_{ij,ik} + \mathbb{E}Z_{ij,jk} + \mathbb{E}Z_{ik,jk} + 6\,\mathbb{E}Z_{ij,ik,jk} \\ &= \mathbb{E}Y_{i,j,k}(\mathbf{m}_1) + 6\,\pi_{ij}\pi_{ik}\pi_{jk} \\ &= \frac{d_i d_j d_k}{C^2}\,(d_i+d_j+d_k) + 6\,\frac{d_i^2 d_j^2 d_k^2}{C^3}\ .\end{aligned} \tag{4.5}$$

Finally, by using Equations (4.2), (4.3), (4.4) and (4.5), we obtain

$$\begin{aligned}&\mathbb{E}N^2(\mathbf{m}_1) = \\ &= \sum_{\{i,j,k,\ell,u,v\}\in I_6} \frac{d_i d_j d_k d_\ell d_u d_v}{C^4}\,(d_i+d_j+d_k)\,(d_\ell+d_u+d_v) \\ &\quad + \sum_{1\leqslant i\leqslant N}\ \sum_{\{j,k\}\in I_2(i)}\ \sum_{\{\ell,u\}\in I_2(ijk)} \frac{d_i^2 d_j d_k d_\ell d_u}{C^4}\,(d_i+d_j+d_k)\,(d_i+d_\ell+d_u) \\ &\quad + \sum_{\{i,j\}\in I_2}\ \sum_{k\in I_1(ij)}\ \sum_{\ell\in I_1(ijk)} \frac{d_i d_j d_k d_\ell}{C^3}(d_i+d_j)^2 + \frac{d_i^2 d_j^2 d_k d_\ell}{C^4}\big\{(d_i+d_j)(d_k+d_\ell) + d_k d_\ell\big\} \\ &\quad + \sum_{\{i,j,k\}\in I_3} \frac{d_i d_j d_k}{C^2}(d_i+d_j+d_k) + 6\,\frac{d_i^2 d_j^2 d_k^2}{C^3}\ .\end{aligned}$$

Motif $\mathbf{m}_2$ (triangle)

Calculations are simpler for triangles. Motif $\mathbf{m}_2$ occurs at position $\{i,j,k\}$ if and only if the 3 edges ij, jk and ik are present, and $Y_{i,j,k}(\mathbf{m}_2)$ reduces to the indicator RV $Z_{ij,ik,jk}$. Thus we have

$$\mathbb{E}Y_{i,j,k}(\mathbf{m}_2) = \pi_{ij}\pi_{jk}\pi_{ik} = \frac{d_i^2 d_j^2 d_k^2}{C^3}\,; \qquad \mathbb{E}N(\mathbf{m}_2) = \sum_{\{i,j,k\}\in I_3} \frac{d_i^2 d_j^2 d_k^2}{C^3}\ . \tag{4.6}$$

Moreover, the product $Y_{i,j,k}(\mathbf{m}_2)\,Y_{i,j,\ell}(\mathbf{m}_2)$ is equal to the indicator RV $Z_{ij,jk,ik,i\ell,j\ell}$ of presence of the 5 edges ij, jk, ik, $i\ell$ and $j\ell$. Therefore,

$$\mathbb{E}\big(Y_{i,j,k}(\mathbf{m}_2)\,Y_{i,j,\ell}(\mathbf{m}_2)\big) \;=\; \pi_{ij}\pi_{jk}\pi_{ik}\pi_{j\ell}\pi_{i\ell} \;=\; \frac{d_i^3 d_j^3 d_k^2 d_\ell^2}{C^5}\ . \tag{4.7}$$

Since $Y_{i,j,k}(\mathbf{m}_2)$ is an indicator RV, we have $Y^2_{i,j,k}(\mathbf{m}_2) = Y_{i,j,k}(\mathbf{m}_2)$ and $\sum_{\{i,j,k\}\in I_3}\mathbb{E}Y^2_{i,j,k}(\mathbf{m}_2) = \mathbb{E}N(\mathbf{m}_2)$.

By plugging the formulas given by (4.6) and (4.7) in Equation (4.2), we obtain the result.

4.2. Motifs of size 4

When $k = 4$, Equation (4.1) reduces to

$$
\begin{aligned}
\mathbb{E}N^2(\mathbf{m}) \;=\; & \sum_{\{i,j,k,\ell\}\in I_4} \; \sum_{\{u,v,w,x\}\in I_4(ijk\ell)} \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m})\,\mathbb{E}Y_{u,v,w,x}(\mathbf{m}) \\
& + \sum_{1\leqslant i\leqslant N} \; \sum_{\{j,k,\ell\}\in I_3(i)} \; \sum_{\{u,v,w\}\in I_3(ijk\ell)} \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m})\,\mathbb{E}Y_{i,u,v,w}(\mathbf{m}) \\
& + \sum_{\{i,j\}\in I_2} \; \sum_{\{k,\ell\}\in I_2(ij)} \; \sum_{\{u,v\}\in I_2(ijk\ell)} \mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m})\,Y_{i,j,u,v}(\mathbf{m})\big) \\
& + \sum_{\{i,j,k\}\in I_3} \; \sum_{\ell\in I_1(ijk)} \; \sum_{u\in I_1(ijk\ell)} \mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m})\,Y_{i,j,k,u}(\mathbf{m})\big) \\
& + \sum_{\{i,j,k,\ell\}\in I_4} \mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m})\;.
\end{aligned}
\tag{4.8}
$$

Following the approach used for motifs of size 3, we detail how to calculate terms in the form $\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m})$, $\mathbb{E}(Y_{i,j,k,\ell}(\mathbf{m})\,Y_{i,j,u,v}(\mathbf{m}))$, $\mathbb{E}(Y_{i,j,k,\ell}(\mathbf{m})\,Y_{i,j,k,u}(\mathbf{m}))$ and $\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m})$, but only for motif $\mathbf{m}_4$. However, all final formulas are gathered in Tables 1, 2, 3 and 4. Before, we give the split of variables $Y_\alpha(\mathbf{m}_i)$ for $3\leqslant i\leqslant 8$, as sums of indicator RVs (see Equations (4.9) to (4.14)). These splits directly derive from the topology of the motif under consideration. Combined with Equation (4.8), they are the basis for obtaining the final formulas presented in the tables.

There are 12 different occurrences of motif $\mathbf{m}_3$ at position $\{i,j,k,\ell\}$, which correspond to different orders of the nodes:

$$
\begin{aligned}
Y_{i,j,k,\ell}(\mathbf{m}_3) = {} & Z_{ij,jk,k\ell} + Z_{jk,k\ell,\ell i} + Z_{k\ell,\ell i,ij} + Z_{\ell i,ij,jk} + Z_{ik,k\ell,\ell j} + Z_{ij,j\ell,\ell k} \\
& + Z_{\ell j,ji,ik} + Z_{\ell k,ki,ij} + Z_{i\ell,\ell j,jk} + Z_{\ell i,ik,kj} + Z_{ki,i\ell,\ell j} + Z_{ik,kj,j\ell}\,.
\end{aligned}
\tag{4.9}
$$

Different occurrences of motif $\mathbf{m}_4$ appear depending on the *central* node (bottom left node in Fig. 1, motif $\mathbf{m}_4$):

$$
Y_{i,j,k,\ell}(\mathbf{m}_4) \;=\; Z_{ij,ik,i\ell} + Z_{ji,jk,j\ell} + Z_{ki,kj,k\ell} + Z_{\ell i,\ell j,\ell k}\;. \tag{4.10}
$$

There are only 3 different ways for motif $\mathbf{m}_5$ to occur:

$$
Y_{i,j,k,\ell}(\mathbf{m}_5) \;=\; Z_{ij,jk,k\ell,\ell i} + Z_{ij,j\ell,\ell k,ki} + Z_{ik,kj,j\ell,\ell i}\;. \tag{4.11}
$$

Occurrences of motif $\mathbf{m}_6$ are obtained through occurrences of motif $\mathbf{m}_4$. When motif $\mathbf{m}_4$ occurs, there are 3 different ways of adding a vertex in order to obtain motif $\mathbf{m}_6$. This leads to a total of 12 different possible occurrences of motif $\mathbf{m}_6$ at $\{i,j,k,\ell\}$:

$$
\begin{aligned}
Y_{i,j,k,\ell}(\mathbf{m}_6) \;=\; & Z_{ij,ik,i\ell,jk} + Z_{ij,ik,i\ell,j\ell} + Z_{ij,ik,i\ell,k\ell} + Z_{ji,jk,j\ell,ik} \\
& + Z_{ji,jk,j\ell,k\ell} + Z_{ji,jk,j\ell,i\ell} + Z_{ki,kj,k\ell,ij} + Z_{ki,kj,k\ell,i\ell} \\
& + Z_{ki,kj,k\ell,j\ell} + Z_{\ell i,\ell j,\ell k,ij} + Z_{\ell i,\ell j,\ell k,ik} + Z_{\ell i,\ell j,\ell k,jk}\;.
\end{aligned}
\tag{4.12}
$$

Motif $\mathbf{m}_7$ is obtained from motif $\mathbf{m}_5$ by adding a diagonal:

$$Y_{i,j,k,\ell}(\mathbf{m}_7) = Z_{ij,jk,k\ell,\ell i,j\ell} + Z_{ij,jk,k\ell,\ell i,ik} + Z_{ij,j\ell,\ell k,ki,jk} + Z_{ij,j\ell,\ell k,ki,i\ell} + Z_{ik,kj,j\ell,\ell i,ij} + Z_{ik,kj,j\ell,\ell i,k\ell} \,. \tag{4.13}$$

Finally, motif $\mathbf{m}_8$ corresponds to a complete sub-graph on vertices $\{i,j,k,\ell\}$ and is thus equal to an indicator RV:

$$Y_{i,j,k,\ell}(\mathbf{m}_8) = Z_{ij,jk,k\ell,i\ell,ik,j\ell} \,. \tag{4.14}$$

Detailed calculations for motif $\mathbf{m}_4$ (star)

Let us start by calculating the expectation $\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_4)$. We use Equation (4.10) and the fact that $\mathbb{E}Z_{ij,ik,i\ell}$ equals $\pi_{ij}\pi_{ik}\pi_{i\ell} = d_i^3 d_j d_k d_\ell / C^3$. Thus,

$$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_4) = \frac{d_i d_j d_k d_\ell}{C^3}\,(d_i^2 + d_j^2 + d_k^2 + d_\ell^2) \tag{4.15}$$

$$\text{and} \qquad \mathbb{E}N(\mathbf{m}_4) = \sum_{1 \leqslant i \leqslant N} \sum_{\{j,k,\ell\} \in I_3(i)} \frac{d_i^3 d_j d_k d_\ell}{C^3} \,. \tag{4.16}$$

We now calculate $\mathbb{E}(Y_{i,j,k,\ell}(\mathbf{m}_4)\, Y_{i,j,u,v}(\mathbf{m}_4))$ by using the product of the sums of indicator RVs:

$$\begin{aligned}
\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_4)\, Y_{i,j,u,v}(\mathbf{m}_4)\big) &= \\
&= \pi_{ij}\,(\pi_{ik}\pi_{i\ell} + \pi_{jk}\pi_{j\ell})\,(\pi_{iu}\pi_{iv} + \pi_{ju}\pi_{jv} + \pi_{iu}\pi_{ju}\pi_{uv} + \pi_{iv}\pi_{jv}\pi_{uv}) \\
&\quad + \pi_{k\ell}\,(\pi_{ik}\pi_{jk} + \pi_{i\ell}\pi_{j\ell})\,(\pi_{ij}\pi_{iu}\pi_{iv} + \pi_{ij}\pi_{ju}\pi_{jv} + \pi_{iu}\pi_{ju}\pi_{uv} + \pi_{iv}\pi_{jv}\pi_{uv}) \\
&= \frac{d_i d_j d_k d_\ell d_u d_v}{C^5} \\
&\quad \times \left\{ (d_i^2 + d_j^2)^2 + \frac{d_i d_j}{C}\Big((d_k^2 + d_\ell^2)\,(d_u^2 + d_v^2) + (d_i^2 + d_j^2)\,(d_u^2 + d_v^2 + d_k^2 + d_\ell^2) \Big) \right\}.
\end{aligned} \tag{4.17}$$

In the same way, we have

$$\begin{aligned}
\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_4)\, Y_{i,j,k,u}(\mathbf{m}_4)\big) &= \frac{d_i d_j d_k d_\ell d_u}{C^4} \left\{ d_i^2 \left(d_i + \frac{d_j^2 d_k}{C} + \frac{d_j d_k^2}{C} \right) \right. \\
&\quad \left. + d_j^2 \left(\frac{d_i^2 d_k}{C} + d_j + \frac{d_i d_k^2}{C} \right) + d_k^2 \left(\frac{d_i^2 d_j}{C} + \frac{d_i d_j^2}{C} + d_k \right) \right\} \\
&\quad + \frac{d_i^2 d_j^2 d_k^2 d_\ell d_u}{C^6} \left\{ (d_i^2 + d_j^2 + d_k^2)\,(d_u^2 + d_\ell^2) + d_\ell^2 d_u^2 \right\} .
\end{aligned} \tag{4.18}$$

We finally compute expectation $\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_4)$:

$$\begin{aligned}\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_4) \;&=\; \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_4) \\ &\quad + 2\,\frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^5}\left(d_i d_j + d_i d_k + d_i d_\ell + d_j d_k + d_j d_\ell + d_k d_\ell\right), \end{aligned} \tag{4.19}$$

$$\sum_{\{i,j,k,\ell\}\in I_4} \mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_4) \;=\; \mathbb{E}N(\mathbf{m}_4) \;+\; 2\sum_{\{i,j\}\in I_2}\;\sum_{\{k,\ell\}\in I_2(ij)} \frac{d_i^3 d_j^3 d_k^2 d_\ell^2}{C^5}\,.$$

Finally, the second moment $\mathbb{E}N^2(\mathbf{m}_4)$ is obtained by plugging the expressions given by (4.15), (4.16), (4.17), (4.18) and (4.19) in Equation (4.8).

5. CONCLUSION

We provide a rigorous probabilistic model for undirected graphs which fits the vertex degrees of an observed graph and thus partially describes real-world networks. This model allows us to derive explicit formulas for the mean and variance of the number of occurrences of the 2 motifs of length 3 and the 6 motifs of length 4. Here, a motif is a simple pattern of interconnexion in a graph. Our methodology can be extended to longer motifs through straightforward calculations. Indeed, one just needs to describe the motif as a sum of indicator variables of Z-type (see decomposition (4.9)–(4.14) for instance). Then the second moment $\mathbb{E}N^2(\mathbf{m})$ given in equation (4.1) reduces to sums of products of expectations of independent Binomial random variables (the Z_{ij}'s for single edges (ij)), easy to compute. Heavy simulations are usually done so far to study over-representation of motifs. Thus, our formulas are of great interest in practice.

We think that no general formula depending only on the total numbers of edges and vertices of the motif exists; additional topological information on the motif is required ($\mathbf{m}_3$ and $\mathbf{m}_4$ both have 4 vertices and 3 edges, but they clearly have different expected counts).

Our methodology can also be generalized to directed motifs and directed graphs. This is an important issue when analyzing biological networks where the orientation of the edges may be known (direction of a reaction in metabolic networks or activation/regulation in gene interaction networks). This will be the matter of a forthcoming paper. Briefly, the probability π_{ij} that an edge goes from i toward j is proportional to the product $\epsilon_i\,\rho_j$ where ϵ_i is chosen as the observed outcoming degree of vertex i and ρ_j is chosen as the observed incoming degree of vertex j. Therefore, this model fits to the incoming and outcoming vertex degrees. Note that this expression for π_{ij} has already been considered by [3] as part of a more general model to detect groups of highly inter-connected vertices which share some similarity.

Finally, one may be interested in counting exact occurrences of a motif $\mathbf{m}$ in graph G. For instance, no "V" motif is counted in a triangle. Our results can be easily extended by defining new indicator RV $X_{i_1,\ldots,i_k}(\mathbf{m})$ which is equal to 1 if the sub-graph with vertices $\{i_1,\ldots,i_k\}$ has exactly the same topology as $\mathbf{m}$ and 0 otherwise. We then write $X_{i_1,\ldots,i_k}(\mathbf{m})$ as a linear combination of ad-hoc edge indicators Z. For instance, if $\mathbf{m}$ is the "V"motif, we just write $X_{i,j,k}(\mathbf{m}_1) = Z_{ij,ik}(1-Z_{jk}) + Z_{ij,jk}(1-Z_{ik}) + Z_{ik,jk}(1-Z_{ij})$.

Table 1: Mean count $\mathbb{E}N(\mathbf{m})$ for non oriented motifs of size 4.

$\mathbf{m}$	$\mathbb{E}N(\mathbf{m})$
$\mathbf{m}_3$ 	$\mathbb{E}N(\mathbf{m}_3) = 2\,C^{-3} \sum_{\{i,j\}\in I_2} \sum_{\{k,\ell\}\in I_2(ij)} d_i^2 d_j^2 d_k d_\ell$
$\mathbf{m}_4$ 	$\mathbb{E}N(\mathbf{m}_4) = C^{-3} \sum_{i=1}^{N} \sum_{\{j,k,\ell\}\in I_3(i)} d_i^3 d_j d_k d_\ell$
$\mathbf{m}_5$ 	$\mathbb{E}N(\mathbf{m}_5) = 3\,C^{-4} \sum_{\{i,j,k,\ell\}\in I_4} d_i^2 d_j^2 d_k^2 d_\ell^2$
$\mathbf{m}_6$ 	$\mathbb{E}N(\mathbf{m}_6) = C^{-4} \sum_{1\leqslant i\leqslant N} \sum_{j\neq i} \sum_{\{k,\ell\}\in I_2(ij)} d_i^3 d_j d_k^2 d_\ell^2$
$\mathbf{m}_7$ 	$\mathbb{E}N(\mathbf{m}_7) = C^{-5} \sum_{\{i,j\}\in I_2} \sum_{\{k,\ell\}\in I_2(ij)} d_i^3 d_j^3 d_k^2 d_\ell^2$
$\mathbf{m}_8$ 	$\mathbb{E}N(\mathbf{m}_8) = C^{-6} \sum_{\{i,j,k,\ell\}\in I_4} d_i^3 d_j^3 d_k^3 d_\ell^3$

Table 2: Formulas giving $\mathbb{E}(Y_{ijk\ell}(\mathbf{m})\, Y_{ijuv}(\mathbf{m}))$ for non oriented motifs of size 4.

$\mathbf{m}$	$\mathbb{E}\big(Y_{ijk\ell}(\mathbf{m})\, Y_{ijuv}(\mathbf{m})\big)$
$\mathbf{m}_3$ 	$C^{-5} d_i d_j d_k d_\ell d_u d_v \Big[(d_i+d_j)^2\,(d_k+d_\ell)\,(d_u+d_v)\,\{1+3\,C^{-1} d_i d_j\}$ $+\, 2\,d_i d_j (d_i+d_j)\Big\{(d_u+d_v)\,\big(2+C^{-1}(d_i d_j+2\,d_k d_\ell)\big)$ $+\,(d_k+d_\ell)\,\big(2+C^{-1}(d_i d_j+2\,d_u d_v)\big)\Big\}$ $+\, 4\,d_i^2 d_j^2\,\big(1+C^{-1}(d_u d_v+d_k d_\ell)\big)+4\,C^{-1} d_i d_j d_u d_v d_k d_\ell\Big]$
$\mathbf{m}_4$ 	see formula (4.17)
$\mathbf{m}_5$ 	$4\,C^{-7} d_i^3 d_j^3 d_k^2 d_\ell^2 d_u^2 d_v^2 + 5\,C^{-8} d_i^4 d_j^4 d_k^2 d_\ell^2 d_u^2 d_v^2$
$\mathbf{m}_6$ 	$\frac{d_i d_j d_k d_\ell d_u d_v}{C^7}\Big[d_i^2 d_j^2 (d_i+d_j)^2\,(d_k+d_\ell)\,(d_u+d_v)\,\Big(1+\frac{d_k d_\ell}{C}+\frac{d_u d_v}{C}\Big)$ $+\, d_i^2 d_j^2 (d_i+d_j)\Big\{(d_k^2+d_\ell^2)\,(d_u+d_v)\,\Big(1+\frac{d_u d_v}{C}\Big)$ $+\,(d_u^2+d_v^2)\,(d_k+d_\ell)\,\Big(1+\frac{d_k d_\ell}{C}\Big)\Big\}$ $+\, d_i d_j (d_i+d_j)\,(d_i^2+d_j^2)\Big\{d_k d_\ell (d_u+d_v)\,\Big(1+\frac{d_u d_v}{C}\Big)$ $+\, d_u d_v (d_k+d_\ell)\,\Big(1+\frac{d_k d_\ell}{C}\Big)\Big\}$ $+\, d_i d_j (d_i^2+d_j^2)\,\{d_k d_\ell (d_u^2+d_v^2)+d_u d_v (d_k^2+d_\ell^2)\}+d_i^2 d_j^2 (d_k^2+d_\ell^2)(d_u^2+d_v^2)$ $+\,(d_i^2+d_j^2)^2 d_k d_\ell d_u d_v + d_i d_j (d_i+d_j)^2 d_k d_\ell d_u d_v (d_k+d_\ell)(d_u+d_v)\,/\,C\Big]$
$\mathbf{m}_7$ 	$C^{-7} d_i^3 d_j^3 d_k^2 d_\ell^2 d_u^2 d_v^2 \Big\{ C^{-2}(d_i+d_j)^2\,(d_u+d_v)\,(d_k+d_\ell)$ $+\, C^{-2} d_i d_j (d_i+d_j)\Big[\Big(1+\frac{d_u d_v}{C}\Big)(d_k+d_\ell)+\Big(1+\frac{d_k d_\ell}{C}\Big)(d_u+d_v)\Big]$ $+\, C^{-3} d_i^2 d_j^2 (d_u d_v+d_k d_\ell)+C^{-3} d_i d_j d_k d_\ell d_u d_v\Big\}$
$\mathbf{m}_8$ 	$C^{-11} d_i^5 d_j^5 d_k^3 d_\ell^3 d_u^3 d_v^3$

Table 3: Formulas giving $\mathbb{E}(Y_{ijk\ell}(\mathbf{m})\, Y_{ijku}(\mathbf{m}))$ for non oriented motifs of size 4.

$\mathbf{m}$	$\mathbb{E}\big(Y_{ijk\ell}(\mathbf{m})\, Y_{ijku}(\mathbf{m})\big)$
$\mathbf{m}_3$ 	$\frac{d_i d_j d_k d_\ell d_u}{C^4}\Big[6\, d_i d_j d_k \Big\{1 + (d_i+d_j+d_k)\,(d_\ell+d_u)\,/\,C + (d_i d_j + d_i d_k + d_j d_k + d_\ell d_u)\,/\,C + \frac{d_i d_j d_k}{C^2}(d_\ell+d_u)\Big\} + (d_i^2 d_j + d_i d_j^2 + d_i^2 d_k + d_i d_k^2 + d_j d_k^2 + d_j^2 d_k)\Big(1+\frac{d_\ell d_u}{C}+\frac{d_i d_j d_k}{C^2}(d_\ell+d_u)\Big) + 2\,\frac{d_i^2 d_j^2 + d_i^2 d_k^2 + d_j^2 d_k^2}{C}\,(d_u+d_\ell) + 2(d_i^2+d_j^2+d_k^2)\,\frac{d_i d_j d_k}{C}\Big(1+\frac{d_\ell d_u}{C}\Big) + 6\, d_i d_j d_k d_\ell d_u / C^2 (d_i d_k + d_i d_j + d_j d_k)\Big]$
$\mathbf{m}_4$ 	see formula (4.18)
$\mathbf{m}_5$ 	$C^{-6} d_i^2 d_j^2 d_k^2 d_\ell^2 d_u^2 \Big\{ d_i d_j + d_i d_k + d_j d_k + 2\, C^{-1} d_i d_j d_k (d_i+d_j+d_k)\Big\}$
$\mathbf{m}_6$ 	$\frac{d_i^2 d_j^2 d_k^2 d_\ell d_u}{C^5}\Big[(d_i+d_j+d_k)^2 + 2\,\frac{d_u d_\ell}{C^2}\,(d_i^2 d_j^2 + d_i^2 d_k^2 + d_j^2 d_k^2) + 2(d_i+d_j+d_k)\,(d_i d_j + d_i d_k + d_j d_k)\,\frac{(d_u+d_\ell)}{C} + 2\,\frac{d_u d_\ell}{C^2}\,(d_u+d_\ell)\Big\{d_i^2(d_j+d_k) + d_j^2(d_i+d_k) + d_k^2(d_i+d_j)\Big\}\Big] + \frac{d_i^3 d_j^3 d_k^3 d_\ell d_u}{C^7}\Big[3(d_i+d_j+d_k)\,(d_u+d_\ell)^2 + 2\,\frac{d_\ell d_u}{C}\,(d_u+d_\ell)\,(d_i d_j + d_j d_k + d_i d_k)\Big] + \frac{d_i d_j d_k d_\ell^2 d_u^2}{C^6}\Big[d_i^3(d_j+d_k)^2 + d_j^3(d_i+d_k)^2 + d_k^3(d_i+d_j)^2\Big]$
$\mathbf{m}_7$ 	$C^{-6} d_i^2 d_j^2 d_k^2 d_\ell^2 d_u^2 \Big\{3C^{-2}(d_i d_j + d_i d_k + d_j d_k)\, d_i d_j d_k\,(d_\ell+d_u) + C^{-2}(d_i+d_j+d_k)\, d_i d_j d_k d_u d_\ell + 6\, C^{-3} d_i^2 d_j^2 d_k^2 d_\ell d_u + C^{-1}(d_i d_j + d_i d_k + d_j d_k)^2\Big\}$
$\mathbf{m}_8$ 	$C^{-9} d_i^4 d_j^4 d_k^4 d_\ell^3 d_u^3$

Table 4: Formulas giving $\mathbb{E}Y^2_{ijk\ell}(\mathbf{m})$ for non oriented motifs of size 4.

$\mathbf{m}$	$\mathbb{E}Y^2_{ijk\ell}(\mathbf{m})$
$\mathbf{m}_3$ 	$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_3) + C^{-4}d_i^2d_j^2d_k^2d_\ell^2\Big[12\,(3+C^{-2}d_id_jd_kd_\ell)$ $+\,10\,C^{-1}\big(d_id_j+d_id_k+d_id_\ell+d_jd_k+d_jd_\ell+d_kd_\ell\big)\Big]$ $+\,2\,d_id_jd_kd_\ell C^{-4}\Big\{d_id_jd_k(d_i+d_j+d_k)+d_id_jd_\ell(d_i+d_j+d_\ell)$ $+\,d_id_kd_\ell(d_i+d_k+d_\ell)+d_jd_kd_\ell(d_j+d_k+d_\ell)\Big\}$
$\mathbf{m}_4$ 	see formula (4.19)
$\mathbf{m}_5$ 	$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_5) + 6\,C^{-6}d_i^3d_j^3d_k^3d_\ell^3$
$\mathbf{m}_6$ 	$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_6) + 12\,C^{-5}d_i^2d_j^2d_k^2d_\ell^2\Big\{5\,C^{-1}d_id_jd_kd_\ell$ $+\,d_id_j+d_id_k+d_id_\ell+d_jd_k+d_jd_\ell+d_kd_\ell\Big\}$
$\mathbf{m}_7$ 	$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_7) + 30\,\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_8)$
$\mathbf{m}_8$ 	$C^{-6}d_i^3d_j^3d_k^3d_\ell^3$

ACKNOWLEDGMENTS

This work has been supported by the French Action Concertée Incitative *Nouvelles Interfaces des Mathématiques*.

REFERENCES

[1] Barabási, A.-L. and Bonabeau, E. (2003). Scale-free networks, *Scientific American*, 50–59.

[2] Barbour, A.D. and Reinert, G. (2001). Small worlds, *Random Struct. Alg.*, **19**, 54–74.

[3] Berg, J. and Lässig (2004). Local graph alignment and motif search in biological networks, *PNAS*, **101**, 14689–14694.

[4] Bollobas, B. (2001). *Random Graphs*, Cambridge University Press.

[5] Durrett, R. (2006). *Random Graph Dynamics*, Cambridge University Press.

[6] Janson, S.; Rucinski, A. and Luczak, T. (2000). *Random Graphs*, Wiley.

[7] Koyutürk, M.; Grama, A. and Szpankowski, W. (2004). An efficient algorithm for detecting frequent subgraphs in biological networks, *Bioinformatics*, **20**, i200–i207.

[8] Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D. and Alon, U. (2002). Networks motifs: simple building blocks of complex networks, *Science*, **298**, 824–827.

[9] Milo, R.; Itzkovitz, S.; Kashtan, N.; Levitt, R.; Shen-Orr, S.; Ayzenshtat, I.; Sheffer, M. and Alon, U. (2004). Superfamilies of evolved and designed networks, *Science*, **303**, 1538–1542.

[10] Newman, M.E.J.; Strogatz, S.H. and Watts, D.J. (2001). Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E*, **64**, 026118.

[11] Shen-Orr, S.; Milo, R.; Mangan, S. and Alon, U. (2002). Network motifs in the transcriptional regulation network of Eschericha Coli, *Nature Genetics*, **31**, 64–68.

[12] Stumpf, M.P.; Wiuf, C. and May, R.M. (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks, *PNAS*, **102**, 4221–4224.

[13] Zhang, L.; King, O.; Wong, S.; Goldberg, D.; Tong, A.; Lesage, G.; Andrews, B.; Bussey, H.; Boone, C. and Roth, F. (2005). Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network, *Journal of Biology*, **4**, 6.

APPENDIX

This appendix contains some indications in order to prove the formulas put in Tables 1, 2, 3 and 4 for the motifs $\mathbf{m}_3$, $\mathbf{m}_5$, $\mathbf{m}_6$, $\mathbf{m}_7$ and $\mathbf{m}_8$ of length 4.

Motif $\mathbf{m}_3$

We use the split of $Y_{i,j,k,\ell}(\mathbf{m}_3)$ into the sum of 12 different terms. Some symmetrical terms appear and we obtain

$$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_3) \;=\; 2\,\frac{d_i d_j d_k d_\ell}{C^3}\left(d_i d_j + d_i d_k + d_i d_\ell + d_j d_k + d_j d_\ell + d_k d_\ell\right)$$

$$\text{and} \qquad \mathbb{E}N(\mathbf{m}_3) \;=\; 2\sum_{\{i,j\}\in I_2}\;\sum_{\{k,\ell\}\in I_2(i,j)}\frac{d_i^2 d_j^2 d_k d_\ell}{C^3}\;.$$

Let us now compute $\mathbb{E}(Y_{i,j,k,\ell}(\mathbf{m}_3)Y_{i,j,u,v}(\mathbf{m}_3))$. This is a big product but a large number of terms may be grouped together and we have

$$\begin{aligned}
&\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_3)\,Y_{i,j,u,v}(\mathbf{m}_3)\big) \;=\\
&\quad=\; \frac{d_i d_j d_k d_\ell d_u d_v}{C^5}\left\{(d_i+d_j)\,(d_u+d_v)\left(1+\frac{d_i d_j}{C}\right)+2\,d_i d_j\left(1+\frac{d_u d_v}{C}\right)\right\}\\
&\qquad\times\left\{(d_i+d_j)(d_k+d_\ell)+2\,d_i d_j\right\}\\
&\qquad+\;2\,\frac{d_i^2 d_j^2 d_k d_\ell d_u d_v}{C^6}\left\{(d_i+d_j)\,(d_u+d_v)+d_u d_v+d_i d_j\right\}\\
&\qquad\times\left\{(d_i+d_j)(d_k+d_\ell)+2\,d_k d_\ell\right\}\;.
\end{aligned}$$

After some simplifications, we obtain,

$$\begin{aligned}
&\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_3)\,Y_{i,j,u,v}(\mathbf{m}_3)\big) \;=\\
&\quad=\; \frac{d_i d_j d_k d_\ell d_u d_v}{C^5}\left[(d_i+d_j)^2\,(d_k+d_\ell)\,(d_u+d_v)\left\{1+3\frac{d_i d_j}{C}\right\}+2d_i d_j(d_i+d_j)\right.\\
&\qquad\times\left\{(d_u+d_v)\left(2+\frac{d_i d_j}{C}+2\,\frac{d_k d_\ell}{C}\right)+(d_k+d_\ell)\left(2+\frac{d_i d_j}{C}+2\,\frac{d_u d_v}{C}\right)\right\}\\
&\qquad\left.+\;4\,d_i^2 d_j^2\left(1+\frac{d_u d_v}{C}+\frac{d_k d_\ell}{C}\right)+4\,\frac{d_i d_j d_u d_v d_k d_\ell}{C}\right].
\end{aligned}$$

To get $\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_3)$, we write

$$\begin{aligned}
Y^2_{i,j,k,\ell}(\mathbf{m}_3) \;=\; & Y_{i,j,k,\ell}(\mathbf{m}_3) \\
& + 2\Big\{ 6\, Z_{ij,jk,k\ell,\ell i} + 6\, Y_{i,j,k,\ell}(\mathbf{m}_8) + 6\, Z_{ij,j\ell,\ell k,ki} + 6\, Z_{i\ell,\ell j,jk,ki} \\
& + 5\, Z_{ik,i\ell,jk,j\ell,k\ell} + 5\, Z_{ij,i\ell,jk,j\ell,k\ell} + 5\, Z_{ij,ik,jk,j\ell,k\ell} + 5\, Z_{ij,ik,i\ell,j\ell,k\ell} \\
& + 5\, Z_{ij,ik,i\ell,jk,k\ell} + 5\, Z_{ij,ik,i\ell,jk,j\ell} + Z_{i\ell,jk,j\ell,k\ell} + Z_{ik,i\ell,jk,k\ell} \\
& + Z_{ij,i\ell,jk,j\ell} + Z_{ij,ik,i\ell,jk} + Z_{ij,ik,i\ell,k\ell} + Z_{ij,ik,jk,k\ell} + Z_{ij,i\ell,j\ell,k\ell} \\
& + Z_{ij,ik,jk,k\ell} + Z_{ik,i\ell,j\ell,k\ell} + Z_{ik,jk,j\ell,k\ell} + Z_{ij,ik,i\ell,j\ell} + Z_{ij,ik,jk,j\ell} \Big\} \,.
\end{aligned}$$

This leads to

$$\begin{aligned}
\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_3) \;=\; & \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_3) \\
& + \frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^4} \Big[12 \Big(3 + \frac{d_i d_j d_k d_\ell}{C^2} \Big) \\
& + \frac{10}{C} \left(d_i d_j + d_i d_k + d_i d_\ell + d_j d_k + d_j d_\ell + d_k d_\ell \right) \Big] \\
& + 2\, \frac{d_i d_j d_k d_\ell}{C^4} \Big\{ d_i d_j d_k (d_i + d_j + d_k) + d_i d_j d_\ell (d_i + d_j + d_\ell) \\
& + d_i d_k d_\ell (d_i + d_k + d_\ell) + d_j d_k d_\ell (d_j + d_k + d_\ell) \Big\} \,.
\end{aligned}$$

Motif $\mathbf{m}_5$ (square)

First, let us calculate the probability $\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_5)$ that the motif $\mathbf{m}_5$ occurs at position $\{i,j,k,\ell\}$. Write $Y_{i,j,k,\ell}(\mathbf{m}_5) = Z_{ij,jk,k\ell,\ell i} + Z_{ij,j\ell,\ell k,ki} + Z_{ik,kj,j\ell,\ell i}$. Each one of these indicator RVs has same expectation equal to $d_i^2 d_j^2 d_k^2 d_\ell^2 / C^4$. Therefore, we have

$$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_5) = 3\, \frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^4} \qquad \text{and} \qquad \mathbb{E}N(\mathbf{m}_5) = 3 \sum_{\{i,j,k,\ell\} \in I_4} \frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^4} \,.$$

We now calculate $\mathbb{E}(Y_{i,j,k,\ell}(\mathbf{m}_5)\, Y_{i,j,u,v}(\mathbf{m}_5))$ like $\mathbb{E}\{(Z_{ij,jk,k\ell,\ell i} + Z_{ij,j\ell,\ell k,ki} + Z_{ik,kj,j\ell,\ell i})\, (Z_{ij,ju,uv,vi} + Z_{ij,jv,vu,ui} + Z_{iu,uj,jv,vi})\}$. We get

$$\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_5)\, Y_{i,j,u,v}(\mathbf{m}_5)\big) \;=\; 4\, \frac{d_i^3 d_j^3 d_k^2 d_\ell^2 d_u^2 d_v^2}{C^7} + 5\, \frac{d_i^4 d_j^4 d_k^2 d_\ell^2 d_u^2 d_v^2}{C^8} \,. \tag{5.1}$$

Now we provide the calculation of $\mathbb{E}(Y_{i,j,k,\ell}(\mathbf{m}_5)\, Y_{i,j,k,u}(\mathbf{m}_5))$.

$$\begin{aligned}\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_5)\, Y_{i,j,k,u}(\mathbf{m}_5)\big) \;&=\\ &= \pi_{ij}\pi_{jk}\pi_{k\ell}\pi_{i\ell}\left\{\pi_{ku}\pi_{iu}+\pi_{ju}\pi_{ku}\pi_{ik}+\pi_{ik}\pi_{ju}\pi_{iu}\right\}\\ &\quad+\pi_{ij}\pi_{j\ell}\pi_{k\ell}\pi_{ik}\left\{\pi_{jk}\pi_{ku}\pi_{iu}+\pi_{ju}\pi_{ku}+\pi_{jk}\pi_{ju}\pi_{iu}\right\}\\ &\quad+\pi_{ik}\pi_{jk}\pi_{j\ell}\pi_{i\ell}\left\{\pi_{ij}\pi_{ku}\pi_{iu}+\pi_{ij}\pi_{ju}\pi_{ku}+\pi_{ju}\pi_{iu}\right\}\\ &= \frac{d_i^2d_j^2d_k^2d_\ell^2d_u^2}{C^6}\left\{d_id_j+d_id_k+d_jd_k+2\,\frac{d_id_jd_k}{C}(d_i+d_j+d_k)\right\}.\end{aligned} \tag{5.2}$$

Easy computation of $\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_5)$ is allowed since all 3 products of two different indicator RVs appearing in $Y_{i,j,k,\ell}(\mathbf{m}_5)$ are equal to $Z_{ij,jk,k\ell,\ell i,j\ell,ik}$ (indicator RV of the complete graph with vertices $\{i,j,k,\ell\}$), whose expectation equals $d_i^3d_j^3d_k^3d_\ell^3/C^6$

$$\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_5) \;=\; \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_5)+6\,\frac{d_i^3d_j^3d_k^3d_\ell^3}{C^6}\;. \tag{5.3}$$

Motif $\mathbf{m}_6$

According to the split of $Y_{i,j,k,\ell}(\mathbf{m}_6)$ into the sum of 12 terms with symetrical expectations in the form $d_i^3d_jd_k^2d_\ell^2/C^4$, we have,

$$\sum_{\{i,j,k,\ell\}\in I_4}\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_6) \;=\; C^{-4}\sum_{i=1}^{N}\sum_{\{i,j\}\in I_2}\sum_{\{k,\ell\}\in I_2(ij)} d_i^3d_jd_k^2d_\ell^2\;.$$

Concerning $\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_6)$, we have

$$\begin{aligned}Y^2_{i,j,k,\ell}(\mathbf{m}_6) \;=\; Y_{i,j,k,\ell}(\mathbf{m}_6)+2\Big\{&30\,Y_{ijk\ell}(\mathbf{m}_8)+6\Big(Z_{ik,i\ell,jk,j\ell,k\ell}+Z_{ij,i\ell,jk,j\ell,k\ell}\\ &+Z_{ij,ik,jk,j\ell,k\ell}+Z_{ij,ik,i\ell,j\ell,k\ell}+Z_{ij,ik,i\ell,jk,k\ell}+Z_{ij,ik,i\ell,jk,j\ell}\Big)\Big\}\;.\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_6) \;=\; \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_6)+12\,\frac{d_i^2d_j^2d_k^2d_\ell^2}{C^5}\Big\{&5\,\frac{d_id_jd_kd_\ell}{C}\\ &+d_id_j+d_id_k+d_id_\ell+d_jd_k+d_jd_\ell+d_kd_\ell\Big\}\;.\end{aligned}$$

Motif $\mathbf{m}_7$

Using the split $Y_{i,j,k,\ell}(\mathbf{m}_7) = Z_{ij,ik,i\ell,jk,j\ell} + Z_{ij,ik,i\ell,kj,k\ell} + Z_{ij,ik,i\ell,\ell j,\ell k} + Z_{ji,jk,j\ell,ik,i\ell}$, we obtain

$$\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_7) = \frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^5}\left(d_i d_j + d_i d_k + d_i d_\ell + d_j d_k + d_j d_\ell + d_k d_\ell\right),$$

$$\mathbb{E}N(\mathbf{m}_7) = \sum_{\{i,j\}\in I_2}\ \sum_{\{k,\ell\}\in I_2(ij)} \frac{d_i^3 d_j^3 d_k^2 d_\ell^2}{C^6}.$$

$$\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_7)\, Y_{i,j,u,v}(\mathbf{m}_7)\big) =$$

$$\begin{aligned}
&= \frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^4}\left(\frac{d_j d_\ell}{C} + \frac{d_i d_k}{C} + \frac{d_j d_k}{C} + \frac{d_i d_\ell}{C} + \frac{d_i d_j}{C}\right)\\
&\quad\times \frac{d_i d_j d_u^2 d_v^2}{C^3}\left(\frac{d_j d_v}{C} + \frac{d_i d_u}{C} + \frac{d_j d_u}{C} + \frac{d_i d_v}{C} + \frac{d_i d_j}{C} + \frac{d_i d_j d_u d_v}{C^2}\right)\\
&\quad+ \frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^4}\times\frac{d_k d_\ell}{C}\times\frac{d_i^2 d_j^2 d_u^2 d_v^2}{C^4}\left(\frac{d_j d_v}{C}+\frac{d_i d_u}{C}+\frac{d_j d_u}{C}+\frac{d_i d_v}{C}+\frac{d_i d_j}{C}+\frac{d_u d_v}{C}\right).
\end{aligned}$$

After simplifications, we have

$$\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_7)\, Y_{i,j,u,v}(\mathbf{m}_7)\big) =$$

$$\begin{aligned}
&= \frac{d_i^3 d_j^3 d_k^2 d_\ell^2 d_u^2 d_v^2}{C^7}\left\{(d_i+d_j)^2\,\frac{(d_u+d_v)\,(d_k+d_\ell)}{C^2}\right.\\
&\quad+ \frac{d_i d_j}{C^2}\,(d_i+d_j)\left[\left(1+\frac{d_u d_v}{C}\right)(d_k+d_\ell) + \left(1+\frac{d_k d_\ell}{C}\right)(d_u+d_v)\right]\\
&\quad\left.+ \frac{d_i^2 d_j^2}{C^3}\,(d_u d_v + d_k d_\ell) + \frac{d_i d_j d_k d_\ell d_u d_v}{C^3}\right\}.
\end{aligned}$$

Now we focus on $\mathbb{E}(Y_{i,j,k,\ell}(\mathbf{m}_7)\, Y_{i,j,k,u}(\mathbf{m}_7))$.

$$\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_7)\, Y_{i,j,k,u}(\mathbf{m}_7)\big) =$$

$$\begin{aligned}
&= \frac{d_i^2 d_j^2 d_k^2 d_\ell^2}{C^5}\left\{\frac{d_j d_\ell d_i d_k d_u^2}{C^3}\left[d_j d_u + d_i d_k + d_j d_k + \frac{d_i d_j d_k d_u}{C} + d_i d_j + \frac{d_i d_j d_k d_u}{C}\right]\right.\\
&\quad+ \frac{d_i d_\ell d_j d_k d_u^2}{C^3}\left[\frac{d_i d_j d_k d_u}{C} + d_i d_k + d_j d_k + d_i d_u + d_i d_j + \frac{d_i d_j d_k d_u}{C}\right]\\
&\quad+ \frac{d_k d_\ell d_i d_j d_u^2}{C^3}\left[\frac{d_i d_j d_k d_u}{C} + d_i d_k + d_j d_k + \frac{d_i d_j d_k d_u}{C} + d_i d_j + d_k d_u\right]\\
&\quad+ \left(d_i d_k + d_j d_k + d_i d_j\right)\frac{d_u^2}{C^2}\\
&\quad\left.\times\left[\frac{d_i d_j d_k d_u}{C} + d_i d_k + d_j d_k + \frac{d_i d_j d_k d_u}{C} + d_i d_j + \frac{d_i d_j d_k d_u}{C}\right]\right\}.
\end{aligned}$$

After simplifications, we have

$$\begin{aligned}\mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_7)\, Y_{i,j,k,u}(\mathbf{m}_7)\big) \;&=\\ &= \frac{d_i^2 d_j^2 d_k^2 d_\ell^2 d_u^2}{C^6}\left\{3\,(d_i d_j + d_i d_k + d_j d_k)\,\frac{d_i d_j d_k}{C^2}\,(d_\ell + d_u)\right.\\ &\quad \left. +\,(d_i + d_j + d_k)\,\frac{d_i d_j d_k d_u d_\ell}{C^2} + 6\,\frac{d_i^2 d_j^2 d_k^2 d_\ell d_u}{C^3} + \frac{(d_i d_j + d_i d_k + d_j d_k)^2}{C}\right\}.\end{aligned}$$

Now, we compute $\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_7)$. Any product of two different indicator RVns appearing in $\mathbf{m}_7$ is equal to indicator RV of the complete graph on $\{i,j,k,\ell\}$. Thus,

$$\mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_7) \;=\; \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_7) + 30\,\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_8)\;,$$

where $\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_8)$ is given below.

Motif $\mathbf{m}_8$

Motif $\mathbf{m}_8$ corresponds to a totally connected subgraph. In particular, $Y_{i,j,k,\ell}(\mathbf{m}_8)$ is an indicator RV, which simplifies calculations. We have

$$\begin{aligned}\mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_8) \;&=\; \frac{d_i^3 d_j^3 d_k^3 d_\ell^3}{C^6}\;,\\ \mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_8)\, Y_{i,j,u,v}(\mathbf{m}_8)\big) \;&=\; \frac{d_i^5 d_j^5 d_k^3 d_\ell^3 d_u^3 d_v^3}{C^{11}}\;,\\ \mathbb{E}\big(Y_{i,j,k,\ell}(\mathbf{m}_8)\, Y_{i,j,k,u}(\mathbf{m}_8)\big) \;&=\; \frac{d_i^4 d_j^4 d_k^4 d_\ell^3 d_u^3}{C^9}\;,\\ \mathbb{E}Y^2_{i,j,k,\ell}(\mathbf{m}_8) \;&=\; \mathbb{E}Y_{i,j,k,\ell}(\mathbf{m}_8)\;.\end{aligned}$$