

# Assessing the exceptionality of coloured motifs in networks

Sophie Schbath<sup>1</sup>, Vincent Lacroix<sup>2</sup>, Marie-France Sagot<sup>3,4</sup>

<sup>1</sup> INRA, UR1077 Mathématique, Informatique et Génome, 78352 Jouy-en-Josas, France

<sup>2</sup> Centre de Regulació Genòmica, Universitat Pompeu Fabra, Dr. Aiguàder, 88, 08003 Barcelona, Spain

<sup>3</sup> Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France

<sup>4</sup> Projet BAMBOO, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France

## Abstract

Various methods have been recently employed to characterise the structure of biological networks. In particular, the concept of network motif and the related one of coloured motif have proven useful to model the notion of a functional/evolutionary building block. However, algorithms that enumerate all the motifs of a network may produce a very large output and methods to decide which motifs should be selected for downstream analysis are needed. A widely used such method is to assess if the motif is exceptional, that is, over- or under- represented with respect to a null hypothesis. Much effort has been put in the last thirty years to derive  $p$ -values for the frequencies of topological motifs, i.e. fixed subgraphs. They rely either on (compound) Poisson and Gaussian approximations for the motif count distribution in Erdős-Rényi random graphs or on simulations in other models. In this paper, we focus on a different definition of graph motifs that corresponds to coloured motifs. A coloured motif is a connected subgraph with fixed vertex colours but unspecified topology. Our work is the first analytical attempt to assess the exceptionality of coloured motifs in networks without any simulation. We first establish analytical formulae for the mean and the variance of the count of a coloured motif in an Erdős-Rényi random graph model. Using simulations under this model, we further show that a Pólya-Aeppli distribution better approximates the distribution of the motif count compared to Gaussian or Poisson distributions. The Pólya-Aeppli distribution, and more generally the compound Poisson distributions are indeed well-designed to model counts of clumping events. Altogether, these results enable to derive a  $p$ -value for a coloured motif, without spending time on simulations.

## 1 Introduction

Descriptions of biological networks serve two main purposes. On the one hand, it enables to address questions related to the evolution of the network, that is, how such a complex

structure has been set up in the course of evolution. On the other hand, structural analysis can be seen as a first necessary step prior to a dynamical analysis which in turn enables to simulate networks and to study their response to perturbation. Usually, three main classes of biological networks are considered (Alm and Arkin (2003)): protein interaction, gene regulatory and metabolic. When analysing their structure, these networks are usually modelled as graphs where vertices represent molecules (metabolites, genes, proteins) and edges (directed or undirected) represent interactions between these molecules (the direction, when it is known, indicating which molecule is acting upon the other). For instance, in the case of a gene regulatory network, vertices correspond to genes and there is a directed edge from a gene coding for a transcription factor to every gene that this transcription factor regulates.

The structure of a biological network may be apprehended by using a variety of measures, such as vertex degree (Jeong *et al.* (2000)), degree correlation (Maslov and Sneppen (2002)) or average shortest path length (Wagner and Fell (2001)).

In this paper, we focus on the concept of motif. A network motif has been initially defined as a pattern of interconnections which occurs unexpectedly often in a network (Milo *et al.* (2002), Shen-Orr *et al.* (2002)). The assumption generally made is that subnetworks sharing the same topology will be functionally similar. Over- (respectively under-) represented subnetworks may therefore correspond to conserved (resp. avoided) and thus important (resp. vital/detrimental) cellular functions. In the context of regulatory networks, simple patterns such as loops may be interpreted as logical circuits controlling the dynamic behaviour of a network. If the over- and under-representation of network motifs is often assessed via simulations of random networks in practice, approximations of the subgraph count distribution in various random graph models have been proposed in the literature. Some of these approximations can be found in the book Janson *et al.* (2000) or in more recent studies such as Stark (2001), Itzkovitz *et al.* (2003), Camacho *et al.* (2007), Picard *et al.* (2008).

A limitation of the notion of topological motif is that in many cases, the same subgraph may in fact correspond to different functions, depending on the nature of the vertices that compose it. This is typically the case for metabolic networks whose fullest representation is in terms of a bipartite graph with two sets of vertices, one corresponding to reactions and the other to chemical compounds those reactions require as input or produce as output. Topological motifs which neglect vertex labels (for the reactions or/and the compounds) may associate completely different chemical transformations, while motifs that took such labels into account but enforced topological isomorphism would miss the fact that some sets of similar transformations may occur in different order. A biological example of the latter is given in the simple case of linear sets of transformations in Figure 1 where rectangles are reactions and circles are compounds. More complex examples are discussed in Lacroix *et al.* (2006).

Moreover, in some situations, as for example in the case of protein interaction networks, the topology of the network is not fully known. Indeed, high-throughput experiments used to obtain large-scale protein interaction data are notoriously noisy, that is, they may detect interactions when there is none (false positive) and they may miss existing interactions

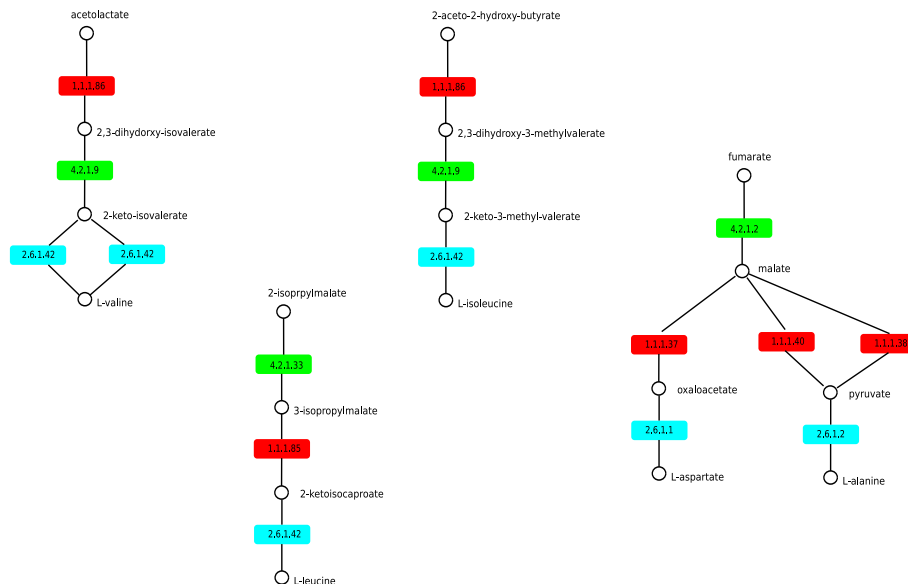


Figure 1: Similar sets of transformations in the metabolic network of the bacterium *Escherichia coli*.

(false negative). In this context, it may be inadequate to look for exact repetitions of a pattern. An alternative definition has thus been proposed, where a motif is defined by using the labels of its vertices and only connectedness of the induced subgraph is required (Lacroix *et al.* (2006)).

A coloured motif is defined as a *multiset* of colours (vertex labels), i.e. a motif may contain colours whose multiplicity are greater than 1. The cardinality of a motif, i.e. of the multiset, will be called the size of a motif. An occurrence of a motif is defined as a connected subgraph whose labels match the motif.

The enumeration of coloured motifs is a non-trivial task which has been the subject of several works (Lacroix *et al.* (2006), Hermelin *et al.* (2007)) which allow to establish the complexity of the problem and provide algorithms to efficiently detect all the occurrences of a motif in a graph. In practice, current methods now allow to enumerate all the motifs of size 7 of a graph representing the metabolic network of a bacterium in less than two hours. Beyond the time complexity of the task, a major challenge that remains open is to make sense of the potentially very large output of such an enumeration procedure, especially when the focus is not on a single motif but on all motifs of a given size. Ideally, one would need a method to rank the motifs according to their biological relevance in order to prioritise a small number of motifs for downstream analysis. However, the notion of biological relevance is generally ill-defined, and a classically used approximation is its statistical significance (or exceptionality).

The exceptionality of a coloured motif, that is the over- or under- representation of the motif with respect to a null model, can be assessed by comparing the observed count of occurrences of a motif to the expected count of the same motif under a null

hypothesis. Up to now, this procedure was performed (for instance, in MOTUS (Lacroix *et al.* (2008), <http://pbil.univ-lyon1.fr/software/motus/>) using simulations: a large number of random graphs were generated and the motif of interest was sought in each one, generating an empirical distribution of the motif count to which the observed count could be compared in order to derive a  $z$ -score and a  $p$ -value. The main limitation of this procedure is that it adds a multiplicative factor to the time complexity of the algorithm. Moreover, it is not trivial to choose the optimal number of simulations to perform in order to get a satisfactory estimation of the  $p$ -value. As a rule of thumb, in order to estimate quite accurately a  $p$ -value of 1 over  $10^i$ , at least  $10^{i+2}$  simulations should be performed.

In this paper, we propose a new approach for assessing the exceptionality of coloured motifs which does not require simulations and therefore circumvents the previously mentioned limitations. We were able to establish exact analytical formulae for the mean and the variance of the count of a coloured motif in an Erdős-Rényi (ER) random graph model. Thanks to these results, one can now derive a  $z$ -score for each motif and therefore rank them according to their exceptionality. We then worked on modelling the complete distribution of the count of a coloured motif in an ER random graph model. To this purpose, we performed a large number of simulations, using different colour frequencies for the motif and different number of vertices and edges for the graph. We could establish that the Poisson distribution was not appropriate whereas the Pólya-Aeppli distribution was a good and better approximation than the commonly used Gaussian distribution. The choice of a Pólya-Aeppli distribution was driven by the following facts: (i) motif occurrences overlap in a network, as shown in Figure 1, (ii) compound Poisson distributions are particularly adapted to model counts of clumping events (Johnson *et al.* (1992), chap. 9), (iii) Pólya-Aeppli approximations are efficient for the count of words in letter sequences (Schbath (1995)). These results can in turn be used to derive a  $p$ -value for each motif, and therefore to introduce a cut-off for deciding which motifs should be selected for downstream analysis.

To our knowledge, there has been no previous work on the significance of coloured motifs in random graphs. This is the reason why we started by focusing on the more general random graph model that is available. We are aware that this may not be the most suitable model to describe the structure of a biological network. However, we argue that this work provides a first necessary basis which can later be extended to richer models, such as the promising mixture of Erdős-Rényi models proposed by Daudin *et al.* (2008).

## 2 Definitions and notations

**Coloured random graph model.** We consider a random graph  $G$  with  $n$  vertices  $\{V_1, \dots, V_n\}$ . We assume that random edges are independent and distributed according to a Bernoulli distribution with parameter  $p \in ]0, 1]$  (so-called Erdős-Rényi model). Moreover, vertices are randomly and independently coloured as follows. Let  $\mathcal{C}$  be a finite set

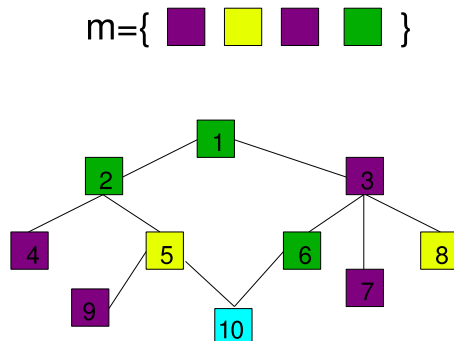


Figure 2: Example of a graph and a motif. The motif  $\mathbf{m}$  occurs three times in the graph, at positions  $\{2, 4, 5, 9\}$ ,  $\{1, 3, 7, 8\}$  and  $\{3, 6, 7, 8\}$ .

of  $r$  different colours and  $f$  a probability measure on  $\mathcal{C}$ :  $f(c)$  is then the probability for a vertex to be coloured with  $c \in \mathcal{C}$ .

In a metabolic network, the colours of reaction vertices can represent classes of chemical transformations; in regulation networks, the colours of gene vertices can represent functional classes. For defining these classes, the EC number hierarchy (<http://www.chem.qmul.ac.uk/iubmb/enzyme/>) or Gene Ontology (<http://www.geneontology.org/GO.doc.shtml>) are classically used.

**Coloured motif.** We consider motifs as introduced in Lacroix *et al.* (2006): a (coloured) motif  $\mathbf{m}$  of size  $k$  is a multiset of  $k$  colours  $\{m_1, \dots, m_k\} \in \mathcal{C}^k$ . Colours from a motif may not be different, i.e. one may have  $m_i = m_j$  for some  $1 \leq i, j \leq k$ ; We then denote by  $s_{\mathbf{m}}(c)$  the multiplicity of the colour  $c$  in  $\mathbf{m}$ . When there is no ambiguity,  $s_{\mathbf{m}}(c)$  will simply be denoted by  $s(c)$ . The notion of multiplicity of a single colour in  $\mathbf{m}$  will be extended to a multiset of colours in Section 3.2.

**Motif occurrences.** We now define an occurrence of such a coloured motif. To this purpose, we introduce the following notation. If  $i_1, i_2, \dots, i_k$  are  $k$  different indices from  $\{1, \dots, n\}$ , then  $G(i_1, i_2, \dots, i_k)$  represents the subgraph of  $G$  induced by the vertices  $\{V_{i_1}, \dots, V_{i_k}\}$ . Let  $I_k$  be the set of all the subsets of size  $k$  from  $\{1, \dots, n\}$ . We say that a motif  $\mathbf{m} = \{m_1, \dots, m_k\}$  occurs at position  $\alpha = \{i_1, \dots, i_k\} \in I_k$  if and only if  $G(\alpha)$  is connected and the colours of  $G(\alpha)$ , denoted by  $C(\alpha)$ , are exactly  $\{m_1, \dots, m_k\}$ .  $I_k$  corresponds then to the set of possible positions for the occurrence of a motif of size  $k$ . Figure 2 gives an example of a motif and its occurrences.

**Number of occurrences.** We introduce the random indicator variable  $Y_\alpha(\mathbf{m})$  which equals one if motif  $\mathbf{m}$  occurs at position  $\alpha \in I_k$  in  $G$  and zero otherwise:

$$Y_\alpha(\mathbf{m}) = \mathbb{I}\{\mathbf{m} \text{ occurs at position } \alpha\}.$$

$Y_\alpha(\mathbf{m})$  is then a Bernoulli random variable whose expectation is denoted by  $\mu(\mathbf{m})$ :

$$\mu(\mathbf{m}) = \mathbb{E}Y_\alpha(\mathbf{m}) = \mathbb{P}(\mathbf{m} \text{ occurs at position } \alpha).$$

The probability  $\mu(\mathbf{m})$  for  $\mathbf{m}$  to occur at position  $\alpha$  will be given in Section 3.1.

The number of occurrences of the motif  $\mathbf{m}$  in the graph  $G$ , denoted by  $N(\mathbf{m})$ , is defined by:

$$N(\mathbf{m}) = \sum_{\alpha \in I_k} Y_\alpha(\mathbf{m}). \quad (1)$$

### 3 Mean and variance for the count

This section will provide analytical formulae for the mean and the variance of the number of occurrences of a coloured motif in a random graph. It involves the computation of some probabilities of connectedness. The generalization to the number of occurrences of a set a coloured motifs will be done in the supplementary material.

#### 3.1 Mean number of occurrences

The mean number of occurrences of the motif  $\mathbf{m}$  in the graph  $G$  simply follows from the count expression (1):

$$\mathbb{E}N(\mathbf{m}) = \sum_{\alpha \in I_k} \mathbb{E}Y_\alpha(\mathbf{m}) = \binom{n}{k} \mu(\mathbf{m})$$

where  $\mu(\mathbf{m})$  is the occurrence probability of the motif and is given below by equation (3).

**Occurrence probability.** The probability  $\mu(\mathbf{m})$  for  $\mathbf{m}$  to occur at position  $\alpha = (i_1, \dots, i_k)$  is simply equal to the product of two probabilities: the probability that  $G(\alpha)$  is connected and the probability to assign colours  $\{m_1, \dots, m_k\}$  to vertices  $\{V_{i_1}, \dots, V_{i_k}\}$ . The latter, denoted by  $\gamma(\mathbf{m})$ , follows from the multinomial distribution:

$$\gamma(\mathbf{m}) = \frac{k!}{\prod_{c \in \mathcal{C}} s(c)!} \prod_{i=1}^k f(m_i) \quad (2)$$

leading to

$$\mu(\mathbf{m}) = g(k, p) \times \gamma(\mathbf{m}) \quad (3)$$

where  $g(k, p)$  denotes the probability for a random graph (Erdős-Rényi model) with  $k$  vertices and edge probability  $p$  to be connected (by definition,  $0! = 1$ ).

**Connectivity probability** The probability  $g(k, p)$  is calculated recursively (Gilbert (1959)) as follows:

$$g(k, p) = 1 - \sum_{i=1}^{k-1} \binom{k-1}{i-1} g(i, p) (1-p)^{i(k-i)} \quad (4)$$

where  $g(1, p) = 1$ . For instance, for  $2 \leq k \leq 5$ , which is typically the range for the motif size in practice, we have:

$$\begin{aligned} g(2, p) &= p, \\ g(3, p) &= 3p^2 - 2p^3, \\ g(4, p) &= 16p^3 - 33p^4 + 24p^5 - 6p^6, \\ g(5, p) &= 125p^4 - 528p^5 + 970p^6 - 980p^7 + 570p^8 - 180p^9 + 24p^{10}. \end{aligned}$$

### 3.2 Variance of the number of occurrences

Getting the variance is much more involved. We start from  $\text{Var}N(\mathbf{m}) = \mathbb{E}N^2(\mathbf{m}) - (\mathbb{E}N(\mathbf{m}))^2$  and we have to compute the moment of order two:

$$\mathbb{E}N^2(\mathbf{m}) = \sum_{\alpha \in I_k} \sum_{\beta \in I_k} \mathbb{E}[Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})].$$

First, the sums over  $\alpha$  and  $\beta$  are calculated according to the number  $\ell$  of vertices shared by the subgraphs  $G(\alpha)$  and  $G(\beta)$ :

$$\mathbb{E}N^2(\mathbf{m}) = \sum_{\ell=0}^k \sum_{|\alpha \cap \beta|=\ell} \mathbb{E}[Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})].$$

Second, we use the fact that  $Y_\alpha(\mathbf{m})$  and  $Y_\beta(\mathbf{m})$  are indicator variables which leads to  $\mathbb{E}[Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})] = \mathbb{P}(Y_\alpha(\mathbf{m}) = 1 \text{ and } Y_\beta(\mathbf{m}) = 1)$ . These random variables are not independent but the above probability can be written as

$$\mathbb{E}[Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})] = K(\alpha, \beta) \times Q_{\mathbf{m}}(\alpha, \beta) \quad (5)$$

with

$$\begin{aligned} K(\alpha, \beta) &= \mathbb{P}(G(\alpha) \text{ and } G(\beta) \text{ are connected}) \\ Q_{\mathbf{m}}(\alpha, \beta) &= \mathbb{P}(C(\alpha) = C(\beta) = \{m_1, \dots, m_k\}). \end{aligned}$$

The terms  $K(\alpha, \beta)$  and  $Q_{\mathbf{m}}(\alpha, \beta)$  are now separately calculated.

**Computation of  $Q_{\mathbf{m}}(\alpha, \beta)$**  Let  $\ell = |\alpha \cap \beta|$ ; the subgraphs  $G(\alpha)$  and  $G(\beta)$  have thus  $\ell$  vertices in common, with  $0 \leq \ell \leq k$ . Let  $\mathbf{m}^* \subset \mathbf{m}$  such that  $|\mathbf{m}^*| = \ell$  and denote  $\mathbf{m}^- = \mathbf{m} \setminus \mathbf{m}^*$ ;  $\mathbf{m}^*$  represents the colours of the  $\ell$  vertices shared by  $G(\alpha)$  and  $G(\beta)$ . The multiplicity of colour  $c \in \mathcal{C}$  in  $\mathbf{m}^*$  (respectively in  $\mathbf{m}^-$ ) is denoted by  $s^*(c)$  (resp.  $s^-(c)$ ). To calculate  $\mathbb{P}(C(\alpha) = C(\beta) = \mathbf{m})$ , we start by choosing the  $\ell$  colours  $\mathbf{m}^*$  of  $G(\alpha) \cap G(\beta)$  [event with probability  $\gamma(\mathbf{m}^*)$ ], then the  $(k-\ell)$  remaining colours  $\mathbf{m}^-$  are spread over both  $G(\alpha) \setminus (G(\alpha) \cap G(\beta))$  [event with probability  $\gamma(\mathbf{m}^-)$ ] and  $G(\beta) \setminus (G(\alpha) \cap G(\beta))$  [event with probability  $\gamma(\mathbf{m}^-)$ ]. Finally, one just has to sum over all possible *different*  $\mathbf{m}^* \subset \mathbf{m}$  which is equivalent to summing over all  $\mathbf{m}^* \subset \mathbf{m}$  and dividing each term by the multiplicity of  $\mathbf{m}^*$  in  $\mathbf{m}$ . This leads to

$$Q_{\mathbf{m}}(\alpha, \beta) = \sum_{\mathbf{m}^* \subset \mathbf{m}} \frac{\gamma(\mathbf{m}^*)[\gamma(\mathbf{m}^-)]^2}{s(\mathbf{m}^*)} \quad (6)$$

where  $s(\mathbf{m}^*) = s_{\mathbf{m}}(\mathbf{m}^*)$  is the multiplicity of  $\mathbf{m}^*$  in  $\mathbf{m}$ . For instance, if  $\mathcal{C} = \{1, 2, 3\}$ ,  $\mathbf{m} = \{1, 3, 1, 2\}$  and  $\ell = 2$ , then the multiplicity of  $\mathbf{m}^* = \{1, 3\}$  in  $\mathbf{m}$  equals 2 whereas the multiplicity of  $\mathbf{m}^* = \{1, 1\}$  equals 1.

**Computation of  $K(\alpha, \beta)$**  Let again  $\ell = |\alpha \cap \beta|$ . If  $\ell = 0$  (i.e.  $G(\alpha)$  and  $G(\beta)$  are disjoint) or  $\ell = 1$  (i.e.  $G(\alpha)$  and  $G(\beta)$  have a unique vertex in common) then the events  $\{G(\alpha) \text{ is connected}\}$  and  $\{G(\beta) \text{ is connected}\}$  are independent leading to

$$K(\alpha, \beta) = g^2(k, p), \quad \text{if } \ell = 0 \text{ or } 1. \quad (7)$$

Another easy case is when  $\ell = k$  because it means that  $\beta = \alpha$  and therefore

$$K(\alpha, \beta) = g(k, p), \quad \text{if } \ell = k. \quad (8)$$

For the other cases, no general formulae have been found so far but, for small values of  $k$ , one can automatically enumerate all the solutions thanks to the edge binary tree, as described below. As an illustration, the case  $k = 3$  (and  $\ell = 2$ ) will be detailed.

The principle is to work conditionally to the subgraph  $G(\alpha) \cap G(\beta)$ :

$$\begin{aligned} \mathbb{P}(G(\alpha) \text{ and } G(\beta) \text{ are connected}) = \\ \sum_{G'} \mathbb{P}(G(\alpha) \cap G(\beta) = G') \times [\mathbb{P}(G(\alpha) \text{ connected} \mid G(\alpha) \cap G(\beta) = G')]^2 \end{aligned} \quad (9)$$

where  $G'$  is any subgraph of  $\ell$  vertices. Since  $k$  is typically small, both probabilities can be computed by enumerating all possible subgraphs  $G'$  and  $G(\alpha)$ . This can be done by traversing the complete edge binary tree associated to the  $k(k-1)/2$  potential edges of  $G(\alpha)$  that is, to the binary tree whose branches are labelled according to the presence or absence of edges in the subgraph  $G(\alpha)$ . This tree is composed of  $k(k-1)/2$  levels, one for each potential edge and each internal vertex in this tree has two sons: the left one corresponds to the presence of the corresponding edge in the graph whereas the right one corresponds to its absence. It follows that each path from the root to a leaf



corresponds to one of the  $2^{k(k-1)/2}$  possible graphs of size  $k$ . Figure 3 gives an example for  $k = 3$ . Vertices are labelled  $\{i, j, u\}$ , the higher level corresponds to the edge  $(i, j)$ , the middle one corresponds to the edge  $(i, u)$  and the lower level corresponds to the edge  $(j, u)$ . Leaves corresponding to connected graphs are drawn with a square. In practice, the connectedness of a graph can be checked thanks to its adjacency matrix to the power  $k - 1$ . Indeed a graph of size  $k$  with adjacency matrix  $A$  is connected if and only if  $A^{k-1}$  contains no zero (every vertex can be reached from any vertex in at most  $k - 1$  steps). Additionally, the binary tree is built such that all pairs of common vertices between  $G(\alpha)$  and  $G(\beta)$  are at the top levels. The probability of each connected graph of size  $k$  can then be easily calculated when traversing the tree and likewise for both probabilities appearing in Equation (9).

As an illustration, we now detail the computation for  $k = 3$  and  $\ell = 2$ . Let  $i$  and  $j$  be the two common vertices between  $G(\alpha)$  and  $G(\beta)$ , and let  $u$  be the third vertex of  $G(\alpha)$  ( $\alpha = \{i, j, u\}$ ). The edge binary tree is given by Figure 3. In this case, there are only two subgraphs  $G'$  with  $\ell = 2$  vertices: either  $i$  and  $j$  are connected (probability  $p$ ) or they are not connected (probability  $1 - p$ ). In Figure 3, we indicate with a dashed horizontal line the separation between edges in  $G'$  (the conditioning event) and edges in  $G(\alpha) \setminus G'$ . Overall, with  $k = 3$ , there are four possible connected subgraphs  $G(\alpha)$ : the triangle (labelled by 'a') and the three possible 'V' (labelled by 'b', 'c', and 'd'). The probability that  $G(\alpha)$  is connected given  $i \leftrightarrow j$  is obtained from cases 'a' (proba.  $p^2$ ), 'b' (proba.  $p(1 - p)$ ) and 'c' (proba.  $p(1 - p)$ ):

$$\mathbb{P}(G(\alpha) \text{ connected} \mid i \leftrightarrow j) = p^2 + 2p(1 - p) = 2p - p^2.$$

The probability that  $G(\alpha)$  is connected given that  $i$  is not connected with  $j$  is obtained from case 'd' (proba.  $p^2$ ), leading to

$$\begin{aligned} \mathbb{P}(G(\alpha) \text{ and } G(\beta) \text{ are connected}) &= p \times [2p - p^2]^2 + (1 - p) \times [p^2]^2 \\ &= 4p^3 - 3p^4. \end{aligned}$$

Using this algorithm, we find the following results for  $k = 3$  and  $k = 4$  ( $k = 2$  can be processed with the trivial formulae (7) or (8)):

$$\begin{aligned} k = 3 \text{ and } \ell = 2 & : K(\alpha, \beta) = 4p^3 - 3p^4 \\ k = 4 \text{ and } \ell = 2 & : K(\alpha, \beta) = 64p^5 - 160p^6 + 100p^7 + 77p^8 - 136p^9 + 68p^{10} - 12p^{11} \\ k = 4 \text{ and } \ell = 3 & : K(\alpha, \beta) = 27p^4 - 60p^5 + 46p^6 - 12p^7. \end{aligned}$$

Finally we obtained analytical formulae for the variance.

## 4 Towards the motif count distribution: a simulated approach

**Aim.** No theoretical results exist so far on the distribution of coloured motifs in random graphs. In this paper, we propose an approximation for this distribution. Thanks

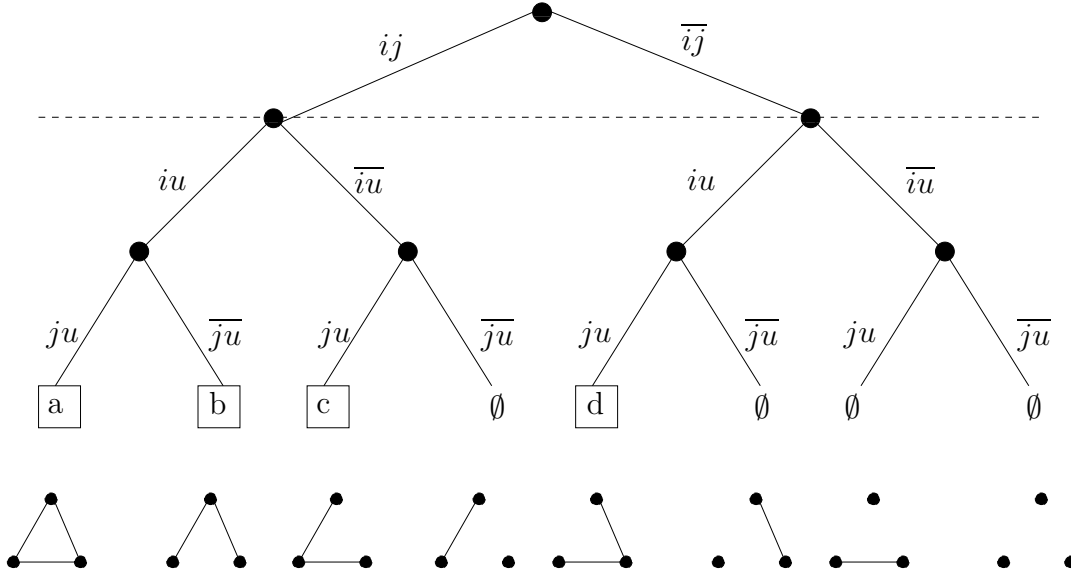


Figure 3: Complete edge binary tree for vertices  $i$ ,  $j$  and  $u$ . Branches are labelled according to the presence or absence of edges: label  $ij$  for instance means that  $i$  and  $j$  are connected, whereas  $\overline{ij}$  means the opposite. Leafs which correspond to connected subgraphs are represented by a square.

to simulations, we first studied the quality of the normal approximation which is classically assumed, especially when using  $z$ -scores (Milo *et al.* (2002), Lacroix *et al.* (2006)). However, network motif occurrences tend to overlap in networks. It is well-known from probability theory that compound Poisson distributions are more relevant than Gaussian distributions to model the count of rare and clumping events. Besides, a compound Poisson approximation for the count of particular subgraphs (topological network motifs) has been proposed by Stark (2001) under certain asymptotic conditions on the ER random graph model. Moreover, by analogy with pattern occurrences in letter sequences (Schbath (1995)), Picard *et al.* (2008) recently investigated a particular compound Poisson approximation, namely a Pólya-Aeppli approximation, and concluded that this distribution fits well the count of topological network motifs. The Pólya-Aeppli distribution (denoted by  $\mathcal{PA}$ ) with parameters  $(\lambda, a)$  is the distribution of  $\sum_{c=1}^C K_c$  where the number of clumps  $C$  is Poisson distributed ( $C \sim \mathcal{P}(\lambda)$ ) and the size  $K_c$  of the clumps is geometrically distributed ( $\mathbb{P}(K_c = k) = (1 - a)a^k$ ). Its mean is equal to  $\lambda/(1 - a)$  and its variance equals  $\lambda(1 + a)/(1 - a)^2$ . We have then also considered the Pólya-Aeppli approximation. We did not investigate the Poisson approximation because, as we can see on Table 1, the variance of the count (whatever the coloured motif) is quite different from the mean count.

**Simulation design.** We have simulated 10,000 Erdős-Rényi random graphs with  $n$  vertices ( $n \in \{100, 500, 1000\}$ ) and edge probability  $p \in \{0.05, 0.01, 0.005\}$ . Vertices have

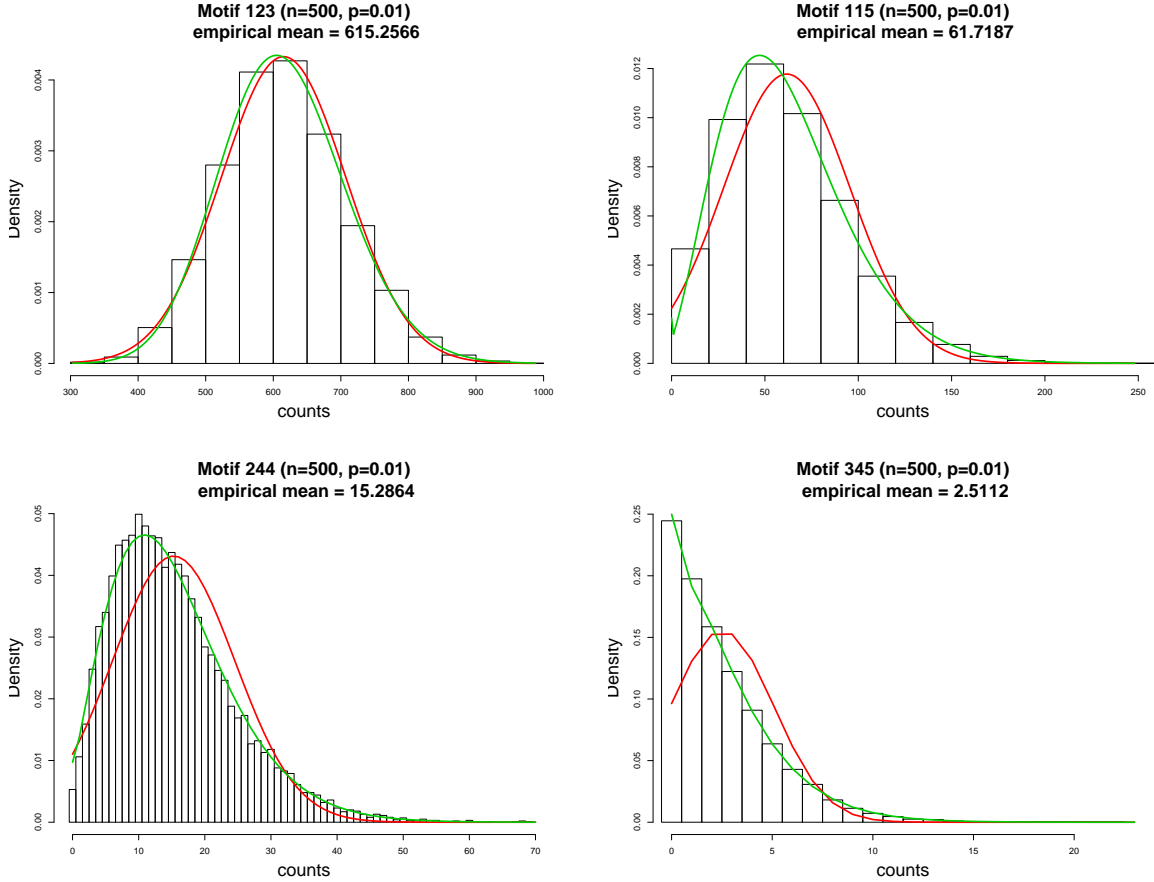


Figure 4: Empirical distributions for the count of motifs  $\{1, 2, 3\}$ ,  $\{1, 1, 5\}$ ,  $\{2, 4, 4\}$  and  $\{3, 4, 5\}$  in random graphs with  $n = 500$  and  $p = 0.01$ . The empirical means are respectively 615, 61, 15 and 2. The red (respectively green) curves correspond to the ad-hoc normal distributions (resp. Pólya-Aeppli distributions).

been randomly coloured with 5 colours ( $\mathcal{C} = \{1, 2, 3, 4, 5\}$ ) and according to the following colour frequencies:  $f = (50, 25, 10, 5, 1)/91$ . These choices for  $n$ ,  $p$  and  $f$  allow to get coloured motifs of size 3 with a wide range of expected counts. We have then selected 14 motifs of size 3 to cover both this variety of counts and different multiplicity pattern:  $\{1, 1, 1\}$ ,  $\{1, 2, 2\}$ ,  $\{1, 2, 3\}$ ,  $\{1, 1, 4\}$ ,  $\{1, 3, 4\}$ ,  $\{1, 1, 5\}$ ,  $\{2, 4, 4\}$ ,  $\{4, 4, 4\}$ ,  $\{2, 4, 5\}$ ,  $\{3, 4, 5\}$ ,  $\{1, 5, 5\}$ ,  $\{3, 5, 5\}$ ,  $\{4, 5, 5\}$  and  $\{5, 5, 5\}$ .

For each motif and each couple  $(n, p)$ , we then obtained an empirical distribution which has been compared with both the normal distribution  $\mathcal{N}(\widehat{\mathbb{E}N(\mathbf{m})}, \widehat{\text{Var}N(\mathbf{m})})$  and the Pólya-Aeppli distribution  $\mathcal{PA}(\widehat{\lambda}, \widehat{a})$  with  $\widehat{\lambda} = (1 - a)\widehat{\mathbb{E}N(\mathbf{m})}$  and  $\widehat{a} = [\widehat{\text{Var}N(\mathbf{m})} - \widehat{\mathbb{E}N(\mathbf{m})}] / [\widehat{\text{Var}N(\mathbf{m})} + \widehat{\mathbb{E}N(\mathbf{m})}]$  (see Figure 4 for 4 representative examples).

**Quality of approximation.** To measure this quality, we adopted two criteria: (1) the Kolmogorov-Smirnov distance which measures the maximal difference between the empirical cumulative distribution function (cdf)  $\hat{F}$  and the cdf of the normal or the Pólya-Aeppli distribution. The closer to 0 the KS distance, the better the approximation. (2) 1 minus the empirical cdf calculated at the 99% and 99.9% quantiles of the normal or of the Pólya-Aeppli distribution. The closer to 1% and 0.1% these values, the better the approximation.

**Results.** Results for different values of  $n$  and  $p$  are very similar. We only present here the ones corresponding to  $n = 500$  and  $p = 0.01$  because these values are very close to those observed in real cases such as the metabolic network of *Escherichia coli* as considered in Lacroix *et al.* (2006). Nevertheless, all results are presented in the supplementary material.

We can first notice just by eye (see Figure 4) that the normal distribution seems satisfactory for frequent motifs but the rarer the motif, the worse the goodness-of-fit. The Pólya-Aeppli distribution seems to fit quite correctly the count distribution whatever the motif. These initial impressions are emphasised when we look at the Kolmogorov-Smirnov distances (see Table 1): the ones for the Pólya-Aeppli distribution are always smaller than those for the normal distribution and sometimes much smaller. In fact, the distance to the normal distribution is quite large for very rare motifs (typically when  $\mathbb{E}N(\mathbf{m}) \leq 10$ ). If we now concentrate on the distribution tails by looking at the empirical probabilities to exceed the 99% or 99.9% quantiles  $q_N$  and  $q_{\mathcal{P}\mathcal{A}}$ , we can also notice that they are closer to 1% or 0.1% for the Pólya-Aeppli distribution than for the normal distribution. For extremely rare motifs, quantiles  $q_{\mathcal{P}\mathcal{A}}$  for both 99% and 99.9% could not be correctly calculated because the corresponding Pólya-Aeppli distribution is both discrete and concentrated around 0. The values for the empirical tails provided in the table are therefore not meaningful in such cases, but thanks to the very small KS distances, we can check that the approximation is still good. Finally, observe that most of the time the normal distribution underestimates the quantile (the empirical right tail is overestimated) leading to false positives.

## 5 Discussion and conclusion

In this paper, we proposed a new way to assess the exceptionality of coloured motifs in networks which does not require to perform simulations. Indeed, we were able to establish analytical formulae for the mean and the variance of the count of a coloured motif in an Erdős-Rényi random graph model. Furthermore, using simulations, we showed that the motif count distribution can be quite accurately approximated with a Pólya-Aeppli distribution, and that neither the Gaussian nor the Poisson distributions are relevant. Altogether, these results now allow to derive a  $p$ -value for a coloured motif without performing simulations. Clearly, when several motifs have to be tested, which is the case in the context of motif discovery, one has to control for multiple testing. A conserva-

motif $\mathbf{m}$	$\mathbb{E}N(\mathbf{m})$	$\text{Var}N(\mathbf{m})$	$\widehat{\mathbb{E}}N(\mathbf{m})$	$\widehat{\text{Var}}N(\mathbf{m})$	$\widehat{a}$	$\widehat{\lambda}$	$KS_{\mathcal{N}}$ (%)	$KS_{\mathcal{P},\mathcal{A}}$ (%)	$\alpha = 1\%$				$\alpha = 0.1\%$			
									$q_{\mathcal{N}}$	$1 - \widehat{F}(q_{\mathcal{N}})$ (%)	$q_{\mathcal{P},\mathcal{A}}$	$1 - \widehat{F}(q_{\mathcal{P},\mathcal{A}})$ (%)	$q_{\mathcal{N}}$	$1 - \widehat{F}(q_{\mathcal{N}})$ (%)	$q_{\mathcal{P},\mathcal{A}}$	$1 - \widehat{F}(q_{\mathcal{P},\mathcal{A}})$ (%)
111	1023.65	27462.66	1021.97	27446.53	0.93	73.37	2.40	0.78	1407.4	1.6	1436	1.1	1533.9	0.23	1591	0.12
122	767.74	14941.43	766.05	14660.79	0.90	76.08	2.14	0.65	1047.7	1.5	1068	1.0	1140.2	0.25	1181	0.07
123	614.19	8546.68	615.26	8493.22	0.86	83.12	1.75	0.68	829.6	1.4	845	0.8	900.0	0.18	929	0.08
114	307.09	5729.89	307.77	5807.09	0.90	30.98	3.20	0.71	485.0	1.5	505	0.8	543.3	0.28	583	0.08
134	122.84	1305.02	123.06	1311.64	0.83	21.11	3.43	0.78	207.3	1.8	219	0.9	235.0	0.37	257	0.12
115	61.41	1180.68	61.72	1147.95	0.90	6.30	5.72	0.98	140.5	2.3	160	0.8	166.4	0.57	205	0.06
244	15.35	85.99	15.29	85.57	0.70	4.63	8.73	1.07	36.8	2.4	43	0.8	43.9	0.81	55	0.12
245	6.14	27.76	6.20	28.45	0.64	2.22	12.72	1.27	18.6	2.5	23	0.8	22.7	1.09	32	0.10
345	2.46	6.63	2.51	6.58	0.45	1.39	17.97	0.53	8.5	1.9	11	0.5	10.4	0.77	15	0.09
155	1.23	6.94	1.22	6.74	0.69	0.37	34.23	5.75	7.2	3.3	12	0.6	9.2	1.56	20	0.05
444	1.02	2.46	1.02	2.51	0.42	0.59	27.39	3.80	4.7	2.4	7	0.5	5.9	1.48	10	0.09
355	0.25	0.50	0.25	0.50	0.34	0.16	48.47	0.43	1.9	2.5	3	0.4	2.4	0.96	6	2e-05
455	0.12	0.20	0.13	0.20	0.23	0.09	51.63	0.16	1.2	0.6	2	0.1	1.5	0.65	4	0.03
555	0.008	0.01	0.007	0.008	0.035	0.007	52.61	2e-03	0.2	0.03	0	0.03	0.3	0.03	1	2e-05

Table 1: Quality of approximation of the count distribution for  $n = 500$  and  $p = 0.01$ . The empirical mean  $\widehat{\mathbb{E}}N(\mathbf{m})$ , variance  $\widehat{\text{Var}}N(\mathbf{m})$  and cumulative distribution function  $\widehat{F}$  have been obtained thanks to 10,000 random graphs.  $(\widehat{a}, \widehat{\lambda})$  are the parameters of the Pólya-Aeppli distribution.  $KS_{\mathcal{N}}$  and  $KS_{\mathcal{P},\mathcal{A}}$  are the Kolmogorov-Smirnov distances. For  $\alpha = 1\%$  then  $0.1\%$ ,  $q_{\mathcal{N}}$  is the  $1 - \alpha$  quantile of the normal distribution (idem for the Pólya-Aeppli distribution).

tive strategy that is classically used and that we would recommend is then to apply a Bonferroni correction.

In this work, we did not investigate the case of long motifs, but we can anticipate that motifs containing submotifs which are exceptional, will tend to be exceptional themselves. This type of phenomenon is also observed for patterns in sequences and a classical way to deal with it is to control for the number of sequence patterns of size  $k - 1$  (by using a Markov model of order  $k - 2$ ), when assessing the exceptionality of patterns of size  $k$ . However, in the case of networks, the problem is far from trivial and it is unclear, even for small values of  $k$  if the space of random graphs verifying these constraints will not be too small. In the worst case, this space may even be reduced to the observed graph itself.

Also in the case of very rare motifs, the expected distribution of the count is essentially concentrated around 0. Therefore, a single occurrence of such a motif will often be sufficient for it to be considered as exceptional. If we now consider the extreme case of a coloured graph where each vertex is assigned a different colour, then all possible motifs will be very rare and therefore, they may all be detected as exceptional. In practical cases, such as for the network representing the metabolic network of the bacterium *Escherichia coli*, the situation is less dramatic but indeed many colours are present only once. This issue may be partially addressed by considering a random graph model where the colours and the topology are not independent anymore. This would allow to discriminate between infrequent poorly connected colours and infrequent highly connected colours. Motifs containing the latter type of colours would be expected to have more occurrences and should therefore not be systematically considered as exceptional when they have a single occurrence.

More generally, we considered in this paper a very simple random graph model. Even though we think this work was necessary to establish a framework for accessing the exceptionality of coloured motifs, an important step is now to extend these results to other models of random graphs which better represent the structure of real networks. Different types of models have been proposed in the literature for this purpose: for instance, small-world networks, scale-free networks, preferential attachment models and fixed degree distribution models. However, these models do not provide the probabilistic distribution on edges which is required to compute the occurrence probability of a motif and the probability of two non disjoint occurrences. Moreover, it has been shown that subnetworks of scale-free networks loose the scale-free property (Stumpf *et al.* (2006)). This is a real drawback for modelling biological networks because they usually correspond to the partial knowledge we have of a system and are therefore far from complete. An interesting issue would be to generalise our work to a mixture of ER random graph models. These models seem indeed very flexible and are able to fit nicely biological networks (Daudin *et al.* (2008)).

Finally, we think there is still room for improvement on the approximation of the motif count distribution. Indeed, no theoretical evidence has been found so far supporting the use of a geometric distribution for the clump size. Analytically getting the third moment and eventually the fourth moment of the count could certainly allow to investigate other distributions.

## Acknowledgments

The authors would like to thank Etienne Birmelé, Jean-Jacques Daudin, Catherine Matias and Stéphane Robin for helpful discussions about the moment calculations. We particularly thank Jean-Jacques Daudin for providing a matlab program to automatically compute the term  $K(\alpha, \beta)$ . We also thank the anonymous reviewers for their helpful comments and suggestions for improving the manuscript. This work has been supported by the ANR (NEMO project BLAN08-1\_318829, REGLIS project NT05-3\_45205 and MIRI project BLAN08-1\_335497), and the ANR-BBSRC (MetNet4SysBio project ANR-07-BSYS 003 02).

## References

- ALM, E. and ARKIN, A. P. (Apr, 2003). Biological networks. *Curr Opin Struct Biol.* **13** (2) 193–202.
- CAMACHO, J., STOUFFER, D. and AMARAL, L. (2007). Quantitative analysis of the local structure of food webs. *Journal of Theoretical Biology.* **246** 260–268.
- DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18** 173–183.
- GILBERT, E. (1959). Random graphs. *The Annals of Mathematical Statistics.* **30** (4) 1141–1144.
- HERMELIN, D., FELLOWS, M., FERTIN, G. and VIALETTE, S. (2007). Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. In *Proc. 34th International Colloquium on Automata, Languages and Programming (ICALP), Wroclaw, Poland*, Lecture Notes In Computer Science.
- ITZKOVITZ, S., MILO, R., KASHTAN, N., ZIV, G. and ALON, U. (2003). Subgraphs in random networks. *Physical Review E.* **68** 026127.
- JANSON, S., RUCINSKI, A. and LUCZAK, T. (2000). *Random graphs*. Wiley.
- JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. and BARABASI, A. L. (Oct, 2000). The large-scale organization of metabolic networks. *Nature.* **407** (6804) 651–654.
- JOHNSON, N. L., KOTZ, S. and KEMP, A. W. (1992). *Univariate Discrete Distributions*. Wiley: New-York.
- LACROIX, V., FERNANDES, C. and SAGOT, M.-F. (2006). Motif search in graphs: application to metabolic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* **3** 360–368.

- LACROIX, V., COTTRET, C., ROGIER, O., FERNANDES, C., JOURDAN, F. and SAGOT, M.-F. (2008). Motus: a software and a webserver for the search and enumeration of node-labelled connected subgraphs in biological networks. submitted.
- MASLOV, S. and SNEPPEN, K. (May, 2002). Specificity and stability in topology of protein networks. *Science*. **296 (5569)** 910–913.
- MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (Oct, 2002). Network motifs: simple building blocks of complex networks. *Science*. **298 (5594)** 824–827.
- PICARD, F., DAUDIN, J.-J., KOSKAS, M., SCHBATH, S. and ROBIN, S. (2008). Assessing the exceptionality of network motifs. *J. Comp. Biol.* **15:1** 1–20.
- SCHBATH, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*. **1** 1–16.
- SHEN-ORR, S. S., MILO, R., MANGAN, S. and ALON, U. (May, 2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet.* **31 (1)** 64–68.
- STARK, D. (2001). Compound poisson approximations of subgraph counts in random graphs. *Random Structures & Algorithms*. **18** 39–60.
- STUMPF, M., WIUF, C. and MAY, R. (2006). Subnets of scale-free networks are not scale-free. *PNAS*. **103** 7566–7570.
- WAGNER, A. and FELL, D. A. (Sep, 2001). The small world inside large metabolic networks. *Proc Biol Sci.* **268 (1478)** 1803–1810.



# Supplementary material

## Occurrences of multiple motifs

We now give additional elements for the calculation of the first two moments of the count of a set of coloured motifs. These formulae appear to be useful when one allows some colours of the motif to be "blurred" as in Lacroix *et al.* (2006). For instance, we could look for the "composite" motif {red, blue} where blue means either light blue or dark blue. A simple way to address the problem would be to use the same formulae as before but changing the set of colours (in the above example, light blue and dark blue would be changed to blue). However, this method does not enable to address the case of a motif containing colours belonging to the two sets, for instance the motif {red, blue, light blue}. A general solution is to describe the motif as the union of several "single" motifs. For instance,  $\{\text{red, blue, light blue}\} := \{\text{red, dark blue, light blue}\} \cup \{\text{red, light blue, light blue}\}$ .

Formally, denote by  $\mathcal{M}$  a set of coloured motifs. The number of occurrences of  $\mathcal{M}$  in the graph  $G$  is the sum of the counts  $N(\mathbf{m})$  for  $\mathbf{m} \in \mathcal{M}$ . Consequently, the expected count  $\mathbb{E}N(\mathcal{M})$  is also the sum of the expected counts  $\mathbb{E}N(\mathbf{m})$ ,  $\mathbf{m} \in \mathcal{M}$ . The novelty only appears for the variance because it involves covariance terms:

$$\text{Var}N(\mathcal{M}) = \sum_{\mathbf{m} \in \mathcal{M}} \text{Var}(N(\mathbf{m})) + \sum_{\mathbf{m} \neq \mathbf{m}' \in \mathcal{M}} \text{Cov}(N(\mathbf{m}), N(\mathbf{m}')).$$

Getting the covariance  $\text{Cov}(N(\mathbf{m}), N(\mathbf{m}'))$  requires to calculate  $\mathbb{P}(Y_\alpha(\mathbf{m}) = 1 \text{ and } Y_\beta(\mathbf{m}') = 1)$  for  $\alpha, \beta \in I_k$  and  $\mathbf{m} \neq \mathbf{m}' \in \mathcal{M}$ . Similarly to equation (5), this probability is equal to  $K(\alpha, \beta) \times Q_{\mathbf{m}, \mathbf{m}'}(\alpha, \beta)$  where  $K(\alpha, \beta)$  has been previously introduced and  $Q_{\mathbf{m}, \mathbf{m}'}(\alpha, \beta) = \mathbb{P}(C(\alpha) = \mathbf{m}, C(\beta) = \mathbf{m}')$ . The latter quantity is null when  $\ell := |\alpha \cap \beta| > |\mathbf{m} \cap \mathbf{m}'|$ . When  $\ell \leq |\mathbf{m} \cap \mathbf{m}'|$ ,  $Q_{\mathbf{m}, \mathbf{m}'}(\alpha, \beta)$  is calculated in the same way as  $Q_{\mathbf{m}}(\alpha, \beta)$  (see Eq. (6)) and we get:

$$Q_{\mathbf{m}, \mathbf{m}'}(\alpha, \beta) = \sum_{\mathbf{m}^* \subset \{\mathbf{m} \cap \mathbf{m}'\}, |\mathbf{m}^*| = \ell} \frac{\gamma(\mathbf{m}^*)\gamma(\mathbf{m}^-)\gamma(\mathbf{m}'^-)}{s(\mathbf{m}^*)}$$

where  $\mathbf{m}^*$  is still the multiset of colours of the  $\ell$  common vertices and  $\mathbf{m}^-$  and  $\mathbf{m}'^-$  are the remaining colours from  $\mathbf{m}$  and  $\mathbf{m}'$ .

## Approximation quality

Tables 2, 3 and 4 give the results about the approximation quality for  $n = 100$  ( $p \in \{0.05, 0.01, 0.005\}$ ),  $n = 500$  ( $p \in \{0.05, 0.005\}$ ) and  $n = 1000$  ( $p \in \{0.05, 0.01, 0.005\}$ ).

$\mathbf{m}$	$\mathbb{E}N(\mathbf{m})$	$\text{Var}N(\mathbf{m})$	$\widehat{\mathbb{E}}N(\mathbf{m})$	$\widehat{\text{Var}}N(\mathbf{m})$	$\widehat{a}$	$\widehat{\lambda}$	$KS_{\mathcal{N}}$ (%)	$KS_{\mathcal{PA}}$ (%)	$\alpha = 0.1\%$			
									$q_{\mathcal{N}}$	$1 - \widehat{F}(q_{\mathcal{N}})$ (%)	$q_{\mathcal{PA}}$	$1 - \widehat{F}(q_{\mathcal{PA}})$ (%)
111	194.46	4868.66	194.42	4827.85	0.92	15.05	5.03	1.23	409.1	0.63	461	0.15
122	145.85	2646.09	145.81	2685.28	0.90	15.02	4.42	0.49	305.9	0.54	344	0.07
123	116.68	1520.30	117.19	1551.07	0.86	16.46	3.68	0.90	238.9	0.28	266	0.07
114	58.34	1023.94	58.29	1030.84	0.90	6.24	6.05	0.49	157.5	0.65	194	0.06
134	23.36	234.83	23.42	241.19	0.82	4.14	8.70	0.66	71.4	0.78	92	0.08
115	11.67	211.46	11.57	208.10	0.89	1.22	19.17	6.76	56.1	1.32	91	0.07
244	2.92	15.63	2.90	15.09	0.68	0.93	20.49	3.94	14.9	1.52	25	0.11
245	1.17	5.08	1.16	5.01	0.62	0.44	33.31	1.06	8.1	1.26	16	0.08
345	0.47	1.22	0.47	1.22	0.44	0.26	43.07	0.39	3.9	1.46	8	0.08
155	0.23	1.28	0.23	1.30	0.69	0.07	49.95	1.28	3.8	2.00	13	0.05
444	0.19	0.46	0.19	0.42	0.38	0.11	49.44	1.11	2.2	0.77	6	0.05
355	0.05	0.09	0.04	0.08	0.28	0.03	53.10	0.04	0.9	0.85	3	0.02
455	0.02	0.04	0.02	0.03	0.16	0.02	53.24	4e-03	0.6	0.33	2	0.01
555	0.002	0.002	0.002	0.002	0.05	0.002	51.40	1e-03	0.1	0.01	1	2e-05
111	8.00	29.43	8.00	29.64	0.57	3.40	11.33	2.02	24.8	0.90	32	0.09
122	6.00	16.99	5.94	16.74	0.48	3.11	11.83	0.53	18.6	0.65	24	0.04
123	4.80	11.77	4.75	11.90	0.43	2.71	13.46	0.78	15.4	0.62	20	0.06
114	2.40	6.87	2.36	6.73	0.48	1.23	19.06	1.62	10.4	0.97	16	0.04
134	0.96	2.02	0.95	1.96	0.35	0.62	28.54	0.44	5.3	0.74	9	0.02
115	0.48	1.36	0.48	1.33	0.47	0.25	43.23	0.40	4.0	0.76	9	0.05
244	0.12	0.19	0.11	0.17	0.20	0.09	51.97	0.30	1.4	0.53	4	0.01
245	0.05	0.07	0.05	0.07	0.20	0.04	53.33	0.02	0.9	0.87	3	0.01
345	0.02	0.02	0.02	0.02	0.10	0.01	53.17	3e-03	0.5	0.16	2	2e-05
155	0.01	0.02	0.008	0.01	0.17	0.006	52.35	0.05	0.3	0.16	2	NA
444	0.008	0.01	0.006	0.007	0.08	0.005	52.27	0.01	0.3	0.03	1	0.01
355	0.002	0.002	0.002	0.002	0.047	0.002	51.51	9e-04	0.1	0.01	1	2e-05
455	0.001	0.001	0.001	0.001	< 0	0.001	51.21	-	0.1	2e-05	-	-
555	6e-05	7e-05	0	0	-	-	-	-	-	-	-	-
111	2.00	4.42	2.01	4.48	0.38	1.243	20.04	2.67	8.5	0.88	13	0.03
122	1.50	2.73	1.49	2.69	0.28	1.07	22.97	0.93	6.6	0.70	10	0.01
123	1.20	2.00	1.17	1.93	0.25	0.89	23.93	0.38	5.4	0.64	8	0.02
114	0.60	1.10	0.59	1.11	0.30	0.41	36.44	1.05	3.9	1.23	7	0.02
134	0.24	0.36	0.24	0.36	0.20	0.19	47.88	0.14	2.1	0.39	4	0.03
115	0.12	0.22	0.13	0.24	0.29	0.09	51.52	0.17	1.6	0.85	4	0.05
244	0.03	0.04	0.027	0.03	0.12	0.02	53.45	0.02	0.6	0.28	2	2e-05
245	0.01	0.01	0.01	0.01	0.08	0.01	52.90	0.02	0.4	0.07	1	0.02
345	0.005	0.005	0.005	0.005	< 0	0.005	52.30	-	0.2	2e-05	-	-
155	0.002	0.003	0.004	0.008	0.30	0.003	51.63	0.01	0.3	0.08	1	0.04
444	0.002	0.002	0.002	0.002	< 0	0.002	51.58	-	0.1	2e-05	-	-
355	0.0005	0.0005	3e-04	0.0003	< 0	0.0003	50.66	-	0.05	2e-05	-	-
455	0.0002	0.0002	4e-04	0.0004	< 0	0.0004	50.76	-	0.06	2e-05	-	-
555	2e-05	2e-05	0	0	-	-	-	-	-	-	-	-

Table 2: Quality approximation of the count distribution for  $n = 100$  and  $p = 0.05$  (top),  $p = 0.01$  (middle),  $p = 0.005$  (bottom). The empirical mean  $\widehat{\mathbb{E}}N(\mathbf{m})$ , variance  $\widehat{\text{Var}}N(\mathbf{m})$  and cumulative distribution function  $\widehat{F}$  have been obtained thanks to 10,000 random graphs.  $(\widehat{a}, \widehat{\lambda})$  are the parameters of the Pólya-Aeppli distribution.  $KS$  are the Kolmogorov-Smirnov distances. For  $\alpha = 0.1\%$   $q_{\mathcal{N}}$  is the  $1 - \alpha$  quantile of the normal distribution (idem for the Pólya-Aeppli distribution). NA indicates numerical problems to compute the  $\mathcal{PA}$  distribution, whereas ‘-’ indicates quantities that have not been calculated ( $\widehat{a} < 0$  or null empirical mean and variance).

$\mathbf{m}$	$\mathbb{E}N(\mathbf{m})$	$\text{Var}N(\mathbf{m})$	$\widehat{\mathbb{E}}N(\mathbf{m})$	$\widehat{\text{Var}}N(\mathbf{m})$	$\widehat{a}$	$\widehat{\lambda}$	$KS_{\mathcal{N}}$ (%)	$KS_{\mathcal{P},\mathcal{A}}$ (%)	$\alpha = 0.1\%$			
									$q_{\mathcal{N}}$	$1 - \widehat{F}(q_{\mathcal{N}})$ (%)	$q_{\mathcal{P},\mathcal{A}}$	$1 - \widehat{F}(q_{\mathcal{P},\mathcal{A}})$ (%)
111	24904.19	10415630	24941.43	10463804	0.99	118.62	2.36	NA	34937.6	0.25	NA	NA
122	18678.15	5778818	18675.54	5816997	0.99	119.53	1.74	NA	26128.7	0.18	NA	NA
123	14942.52	2922292	14914.40	2934663	0.99	150.83	0.40	NA	20208.2	7e-02	NA	NA
114	7471.26	2112934	7479.15	2103159	0.99	53.00	1.75	NA	11960.7	0.25	NA	NA
134	2988.50	429279.6	2982.40	420630.6	0.99	42.00	2.53	NA	4986.6	0.32	NA	NA
115	1494.25	451540.6	1497.61	454673.2	0.99	9.83	3.68	1.68	3581.3	0.44	4208	0.04
244	373.56	23775.19	373.30	23459.00	0.97	11.69	5.22	1.32	846.6	0.63	977	0.16
245	149.42	6427.06	149.36	6412.56	0.95	6.80	5.66	0.56	396.8	0.77	485	0.11
345	59.77	1180.90	59.57	1159.17	0.90	5.82	6.24	0.47	164.8	0.73	205	0.11
155	29.88	1225.46	30.29	1242.37	0.95	1.44	18.73	16.78	139.2	1.63	220	0.10
444	24.90	350.77	24.80	345.33	0.87	3.32	11.04	4.33	82.2	1.25	110	0.22
355	5.98	57.78	5.98	56.56	0.81	1.14	17.66	7.55	29.2	1.40	48	0.10
455	2.99	17.18	3.00	16.90	0.70	0.91	20.63	4.78	15.7	1.58	27	0.13
555	0.20	0.48	0.20	0.49	0.42	0.12	49.01	1.27	2.3	0.87	6	0.10
111	256.77	2619.15	256.74	2596.12	0.82	46.21	3.30	0.99	414.2	0.35	435	0.13
122	192.58	1433.80	192.47	1412.41	0.76	46.17	3.17	0.85	308.6	0.29	324	0.10
123	154.06	890.78	154.35	893.83	0.70	45.46	2.84	0.45	246.7	0.24	259	0.04
114	77.031	564.99	76.36	558.69	0.76	18.36	5.14	1.25	149.4	0.52	165	0.07
134	30.81	141.53	30.57	140.58	0.64	10.92	5.84	0.51	67.2	0.51	77	0.03
115	15.41	114.06	15.36	111.20	0.76	3.73	9.10	0.33	47.9	0.77	63	0.02
244	3.85	10.74	3.80	10.73	0.48	1.98	15.18	1.72	13.9	1.05	19	0.07
245	1.54	3.77	1.55	3.66	0.40	0.92	23.09	0.46	7.5	0.77	12	0.04
345	0.62	1.06	0.61	1.04	0.26	0.45	35.66	0.51	3.8	0.86	7	0.03
155	0.31	0.96	0.30	0.91	0.50	0.15	47.67	1.27	3.2	1.42	8	0.04
444	0.26	0.42	0.25	0.40	0.24	0.19	47.39	0.95	2.2	0.60	5	0.02
355	0.06	0.09	0.06	0.08	0.16	0.05	53.37	0.03	0.9	0.80	3	2e-05
455	0.03	0.04	0.03	0.04	0.10	0.03	53.58	0.03	0.7	0.32	2	0.01
555	0.002	0.002	0.002	0.003	0.13	0.002	51.38	0.01	0.2	0.01	1	0.01

Table 3: Quality approximation of the count distribution for  $n = 500$  and  $p = 0.05$  (top),  $p = 0.005$  (bottom). The empirical mean  $\widehat{\mathbb{E}}N(\mathbf{m})$ , variance  $\widehat{\text{Var}}N(\mathbf{m})$  and cumulative distribution function  $\widehat{F}$  have been obtained thanks to 10,000 random graphs.  $(\widehat{a}, \widehat{\lambda})$  are the parameters of the Pólya-Aeppli distribution.  $KS$  are the Kolmogorov-Smirnov distances. For  $\alpha = 0.1\%$   $q_{\mathcal{N}}$  is the  $1 - \alpha$  quantile of the normal distribution (idem for the Pólya-Aeppli distribution). NA indicates difficulties to compute the  $\mathcal{P},\mathcal{A}$  distribution (empirical mean greater than 1500).

$\mathbf{m}$	$\mathbb{E}N(\mathbf{m})$	$\text{Var}N(\mathbf{m})$	$\widehat{\mathbb{E}}N(\mathbf{m})$	$\widehat{\text{Var}}N(\mathbf{m})$	$\widehat{a}$	$\widehat{\lambda}$	$KS_{\mathcal{N}}$ (%)	$KS_{\mathcal{PA}}$ (%)	$\alpha = 0.1\%$			
									$q_{\mathcal{N}}$	$1 - \widehat{F}(q_{\mathcal{N}})$ (%)	$q_{\mathcal{PA}}$	$1 - \widehat{F}(q_{\mathcal{PA}})$ (%)
111	199833.7	314809999	199562.5	317861578	0.999	250.42	1.17	NA	254657.2	0.31	NA	NA
122	149875.2	175755625	149798.2	179968751	0.998	249.16	1.21	NA	191254.5	0.21	NA	NA
123	119900.2	86883859	120091.4	85519373	0.997	336.81	0.64	NA	148668.9	0.16	NA	NA
114	59950.1	63565443	59924	64198229	0.998	111.76	1.27	NA	84684.1	0.21	NA	NA
134	23980.04	12670746	24019.74	12797890	0.996	89.99	1.62	NA	35074.8	0.23	NA	NA
115	11990.02	13673678	11987.86	13594556	0.998	21.12	2.31	NA	23381.8	0.27	NA	NA
244	2997.50	682659.7	3000.27	697930.4	0.991	25.68	4.18	NA	5581.9	0.51	NA	NA
245	1199.00	176888.0	1201.13	179141.7	0.987	16.00	3.85	0.56	2509.1	0.53	2819	0.05
345	479.60	31096.36	481.62	31496.74	0.970	14.51	3.96	0.46	1030.1	0.54	1166	0.08
155	239.80	29743.89	240.65	29747.06	0.984	3.86	8.55	2.36	773.6	1.19	1024	0.16
444	199.83	8536.03	199.96	8581.79	0.954	9.11	6.80	1.95	486.2	0.87	575	0.19
355	47.96	1300.61	48.33	1323.87	0.929	3.40	9.53	2.86	160.8	1.12	216	0.17
455	23.98	359.80	24.17	368.78	0.877	2.97	10.64	3.34	83.5	1.29	114	0.20
555	1.60	7.15	1.61	7.37	0.641	0.58	27.84	8.08	10.0	1.23	19	0.20
111	8213.85	682747.2	8197.27	683834.4	0.98	194.20	1.58	NA	10752.7	0.23	NA	NA
122	6160.39	374345.9	6164.75	374787.5	0.97	199.52	1.12	NA	8056.6	0.18	NA	NA
123	4928.31	200136	4927.37	204984.5	0.95	231.32	1.05	NA	6326.5	0.11	NA	NA
114	2464.16	139786.6	2460.80	138179.1	0.96	86.11	1.72	NA	3609.5	0.18	NA	NA
134	985.66	29788.37	984.84	29629.79	0.93	63.36	2.19	0.53	1516.8	0.22	1580	0.09
115	492.83	29351.43	494.34	29697.4	0.97	16.19	3.28	0.75	1026.9	0.31	1152	0.07
244	123.21	1772.00	123.54	1739.11	0.87	16.39	4.25	0.65	252.4	0.48	282	0.06
245	49.28	523.43	49.58	530.73	0.83	8.47	5.93	0.68	120.8	0.58	143	0.09
345	19.71	108.05	19.80	108.17	0.69	6.13	7.79	0.65	51.9	0.72	63	0.09
155	9.86	119.19	9.87	117.56	0.84	1.53	18.36	12.03	43.4	1.29	67	0.09
444	8.21	36.60	8.29	36.64	0.63	3.06	12.01	2.69	27.0	1.08	36	0.14
355	1.97	6.83	1.98	6.94	0.55	0.88	23.34	2.74	10.1	0.95	17	0.11
455	0.99	2.35	0.98	2.31	0.40	0.58	28.66	1.21	5.7	1.05	10	0.07
555	0.07	0.10	0.06	0.09	0.17	0.05	53.16	0.18	1.0	0.30	3	2e-05
111	2060.35	55750.96	2058.76	55507.45	0.93	147.26	2.06	NA	2786.8	0.27	NA	NA
122	1545.26	30335.99	1544.53	30678.85	0.90	148.06	1.34	NA	2085.8	0.17	NA	NA
123	1236.21	17343.32	1235.04	17059.00	0.86	166.76	1.36	0.59	1638.6	0.17	1668	0.09
114	618.11	11620.72	617.55	11686.33	0.90	61.99	2.67	0.79	951.6	0.25	991	0.10
134	247.24	2644.45	246.68	2612.29	0.83	42.57	2.99	0.71	404.6	0.33	427	0.11
115	123.62	2393.86	124.32	2378.23	0.90	12.35	4.33	0.49	275.0	0.45	315	0.08
244	30.90	174.02	30.94	174.27	0.70	9.33	6.19	0.52	71.7	0.61	83	0.13
245	12.36	56.14	12.49	55.82	0.63	4.57	9.63	1.02	35.6	0.66	45	0.11
345	4.94	13.38	5.00	13.45	0.46	2.71	12.90	1.18	16.3	0.62	22	0.08
155	2.47	14.03	2.43	13.40	0.69	0.74	27.60	7.84	13.7	1.56	24	0.06
444	2.06	4.97	2.07	4.99	0.41	1.21	20.01	3.06	9.0	1.19	13	0.08
355	0.49	1.02	0.49	0.99	0.34	0.32	40.73	0.63	3.6	0.94	7	0.04
455	0.25	0.40	0.24	0.39	0.22	0.19	47.91	0.11	2.1	0.42	5	2e-05
555	0.02	0.02	0.01	0.02	0.04	0.01	53.22	0.01	0.4	0.05	1	0.01

Table 4: Quality approximation of the count distribution for  $n = 1000$  and  $p = 0.05$  (top),  $p = 0.01$  (middle),  $p = 0.005$  (bottom). The empirical mean  $\widehat{\mathbb{E}}N(\mathbf{m})$ , variance  $\widehat{\text{Var}}N(\mathbf{m})$  and cumulative distribution function  $\widehat{F}$  have been obtained thanks to 10,000 random graphs.  $(\widehat{a}, \widehat{\lambda})$  are the parameters of the Pólya-Aeppli distribution.  $KS$  are the Kolmogorov-Smirnov distances. For  $\alpha = 0.1\%$   $q_{\mathcal{N}}$  is the  $1 - \alpha$  quantile of the normal distribution (idem for the Pólya-Aeppli distribution). NA indicates difficulties to compute the  $\mathcal{PA}$  distribution (empirical mean greater than 1500).