JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 15, Number 1, 2008
© Mary Ann Liebert, Inc.
Pp. 1–20
DOI: 10.1089/cmb.2007.0137

Assessing the Exceptionality of Network Motifs

F. PICARD,[1] J.-J. DAUDIN,[2] M. KOSKAS,[2] S. SCHBATH,[3] and S. ROBIN[2]

ABSTRACT

Getting and analyzing biological interaction networks is at the core of systems biology. To help understanding these complex networks, many recent works have suggested to focus on motifs which occur more frequently than expected in random. To identify such exceptional motifs in a given network, we propose a statistical and analytical method which does not require any simulation. For this, we first provide an analytical expression of the mean and variance of the count under any exchangeable random graph model. Then we approximate the motif count distribution by a compound Poisson distribution whose parameters are derived from the mean and variance of the count. Thanks to simulations, we show that the compound Poisson approximation outperforms the Gaussian approximation. The compound Poisson distribution can then be used to get an approximate *p*-value and to decide if an observed count is significantly high or not. Our methodology is applied on protein-protein interaction (PPI) networks, and statistical issues related to exceptional motif detection are discussed.

Key words: computational molecular biology, statistics, random graphs, network motifs, compound Poisson approximation.

1. INTRODUCTION

THE IMPORTANT PROGRESS of high-throughput biology allows us now to consider the cell as a whole system under study. This complex system is mainly represented by various networks of interacting components (e.g., transcriptional regulatory networks, protein-protein interaction networks, metabolic networks). To help understanding the organization and dynamics of cell functions, one usually tries to break down these complex networks into functional modules (Chen and Yuan, 2006) or into basic building blocks (Milo et al., 2002). These blocks are also called patterns of interconnection or motifs. Many definitions can be used to designate network motifs. In this paper a motif will refer to a subgraph with a fixed number of nodes and with a given topology. This type of motif is also called topological motif, which are different from dense subgraphs studied by Koyutürk et al. (2007) for instance. For transcriptional regulatory networks, some motifs such as the three-node feed-forward loop or the four-node bi-fan, may perform specific regulatory functions (Shen-Orr et al., 2002; Lee et al., 2002; Mangan and

[1] Laboratoire Statistique et Génome, UMR CNRS 8071, INRA 1152, Université d'Evry, Evry, France.
[2] UMR518 AgroParisTech/INRA, Paris, France.
[3] INRA, Unité Mathématique, Informatique et Génome UR1077, Jouy-en-Josas, France.

Alon, 2003; Ingram et al., 2006). Moreover, motifs seem to be conserved across species, which suggests a strong link between protein evolution and their belonging to particular topological structures (Wuchty et al., 2003; Batada et al., 2006; Chen and Dokholyan, 2006). Many recent works have suggested to focus on motifs which occur more frequently than expected in random (Milo et al., 2002, 2004; Shen-Orr et al., 2002; Prill et al., 2005). Such motifs seem indeed to reflect functional or computational units which combine to regulate the cellular behavior as a whole. Their possible function can be provided by common themes of the system in which they appear. Additional insight may be gained by mathematical analysis of their dynamics (Mangan and Alon, 2003; Prill et al., 2005; Ingram et al., 2006).

The common method that has been used for now to detect significantly over-represented motifs is based on simulations. Random graphs are first generated such that they preserve some characteristics of the biological network like the numbers of vertices and edges or the degree sequence (numbers of edges per vertex) (Milo et al., 2002, 2004). Then, either a z-score is calculated thanks to the empirical mean and variance of the count (Milo et al., 2002, 2004; Prill et al., 2005), or an estimation of the empirical p-value is derived from the empirical distribution of the count (Shen-Orr et al., 2002; Milo et al., 2002). Such methods are not totally satisfactory from a probabilistic point of view. Indeed, using a z-score means to assume that the motif count follows a Gaussian distribution which is only true asymptotically under some restrictive conditions. Moreover, to evaluate a p-value close to zero, a huge number of simulations have to be performed, which is usually not the case in previous studies because of high computational times. Getting theoretical properties on the motif count distribution would thus be very valuable to identify exceptional motifs.

Several approximations have been proposed under the so-called Erdös-Rényi model (Janson et al., 2000). This basic model originating in Erdös (1947) assumes that edges are independent and distributed according to a Bernoulli distribution with same parameter π. It means in particular that the probability to connect two nodes does not depend on the nodes. Under these assumptions, Poisson and compound Poisson approximations have been first proposed for rare motifs satisfying some conditions on their number of vertices and edges (Erdös and Rényi, 1960; Bollobas, 1981; Barbour, 1982; Karoński and Ruciński, 1983; Stark, 2001; Barbour et al., 1992). The asymptotic normality of the motif count has been also extensively studied and bounds on the approximation error have been derived (Barbour et al., 1987; Janson et al., 2000). However, except for the mean count which is simple to derive under the Erdös-Rényi model, no explicit formula of the parameters of these limiting distributions has never been provided. In particular, no general expression exists for the variance of the count.

Despite an important number of theoretical results on the motif count distribution, the Erdös-Rényi model can not be used as a reference model for biological networks, since it does not fit the connectivity heterogeneity which exists in these networks (Barabási and Albert, 1999). Finding an appropriate reference model is of major importance when searching for exceptional events, since a too simple reference model would consider any observation as being exceptional. Alternative models have been proposed to describe real networks (Barabási and Albert, 1999; Newman et al., 2001; Newman, 2003). Nevertheless, they are mainly based on summary statistics such as the degree distribution, whereas theoretical strategies to identify exceptional motifs need a reference model for the edge distribution, and not only for the degree distribution. Similarly, Middendorf et al. (2005) explored different models to understand the design of complex networks. However, their contribution is based on algorithms which allow the construction of networks, and no theoretical probabilistic model can be associated with the proposed algorithms. For those reasons, Matias et al. (2006) proposed exact formulas for the mean and the variance of the count under a general model assuming that edges are still independent but depend on both connected vertices. However, despite a nice theoretical framework, the non-exchangeability of this model hampers any calculus from the practical point of view.

As a matter of fact, there exists no consensus regarding the choice of an appropriate reference model for biological networks, and our point is not to solve this issue in the present work. However, the assessment of motif exceptionality requires the use of a reference model, and we consider four different models in the following. The first one is the Erdös-Rényi (ER) model, which is used for illustration purposes. Then we consider the popular fixed degree distribution (FDD) model (Newman et al., 2001), which is used by Milo et al. (2002) in the `Mfinder` software. We also study a random graph model with expected degree distribution (EDD) which is an exchangeable version of the FDD model. Some of its general properties have been studied in Park and Newman (2003) and Chung and Lu (2002). The last model is based on mixture distributions which are used to model heterogenous connectivity often observed in real networks. It is called the Erdös-Rényi Mixture for Graphs (ERMG) model. It has been studied in Daudin et al. (2008) and is analogous to the stochastic block model of Nowicki and Snijders (2001).

The first question we address in this paper is how to calculate in a unified way the exact mean and variance of a motif count under any exchangeable random graph model. These two quantities are indeed crucial to identify unexpected motifs. Provided that the occurrence probability of a given motif does not depend on the occurrence position (exchangeability assumption) and that disjoint occurrences are independent, we derive the expression of the first two moments of the count. When calculating the variance, we introduce the new concept of super-motifs, which are formed by two overlapping occurrences of a given motif. Then we calculate the first two moments for all three- and four-node undirected motifs on three protein-protein interaction (PPI) networks from the DIP database (Salwinski et al., 2004), and we discuss the influence of the reference models on those moments. We use the ER, EDD, and ERMG models as exchangeable reference models for which our methodology can be applied. Theoretical moments under those models will be compared with estimated moments (simulation based) for the FDD model.

The second question we focus on is which approximation of the motif count distribution to use to get accurate p-values. Note that no result exists yet on the *exact* distribution of this count. As regard to existing theoretical results under the Erdös-Rényi model, we compare the approximation quality of the following distributions: the Gaussian distribution, and the Pólya-Aeppli distribution. The later is a special compound Poisson distribution whose two parameters can be set from the exact mean and variance of the count we provide. Using simulations, we show that the Compound Poisson approximation is more appropriate than the Gaussian approximation to assess the exceptionality of network motifs. In a last step, we apply our method to identify exceptional undirected motifs of size 3 and 4 in three PPI networks (Salwinski et al., 2004). For the sake of simplicity, we consider undirected graphs and motifs. However, our methodology can be easily generalized to a directed framework as it is discussed in the conclusion.

2. DEFINITIONS AND NOTATIONS

2.1. Random graph with exchangeable distribution

Let us define a random graph G, where $\mathcal{V}$ denotes the set of fixed vertices with $\mathcal{V} = \{1, \ldots, n\}$. Random edges are described by a set of random variables $\mathbf{X} = \{X_{ij}, (i, j) \in \mathcal{V}^2\}$ such that X_{ij} equals 1 if nodes i and j are connected, and 0 otherwise ($X_{ij} = X_{ji}$ for undirected graphs). In the following, we consider random graphs with exchangeable probability distributions. Taking $\{i_1, \ldots, i_\ell\} \in \mathcal{V}^\ell$, it means that:

$$\forall \ell \in \{1, \ldots, n\},\ \left(X_{i_1,i_2}, \ldots, X_{i_{\ell-1},i_\ell}\right) \sim \left(X_{\sigma(i_1),\sigma(i_2)}, \ldots, X_{\sigma(i_{\ell-1}),\sigma(i_\ell)}\right),$$

for every permutation σ defined on the set $\{i_1, \ldots, i_\ell\}$. This exchangeability property is analogous to the stationarity property for random processes.

2.2. Network motif

We denote by $\mathbf{m}$ a network motif of size k, which is a connected subgraph with k vertices. It is defined by a fixed topology through its adjacency matrix also denoted by $\mathbf{m}$, with general term $m_{uv} = 1$ if nodes u and v are connected, and 0 otherwise. A typical example is the motif, which can be defined by three adjacency matrices depending on the position of the central edge, as shown in Table 1.

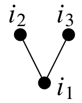
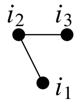
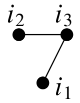
2.3. Position and occurrence of a motif

To define an occurrence of motif $\mathbf{m}$ we introduce notation I_k which is the set of all k-tuples of $\mathcal{V}$, namely

$$I_k = \left\{\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}^k \mid i_j \neq i_\ell, \forall j \neq \ell\right\}.$$

We consider $\alpha \in I_k$, a potential position of $\mathbf{m}$ in G. The number of such positions is $\binom{n}{k}$. In order to match a position with an adjacency matrix, we consider that $\alpha = (i_1, \ldots, i_k)$ with $i_1 < \cdots < i_k$. Then we

TABLE 1. NONREDUNDANT PERMUTATIONS OF THE ∨ MOTIF AT POSITION $\alpha = (i_1, i_2, i_3)$

$\mathbf{m}$	$\mathbf{m}'$	$\mathbf{m}''$
$\begin{bmatrix} 0 & 1 & 1 \\ . & 0 & 0 \\ . & . & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ . & 0 & 1 \\ . & . & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & 1 \\ . & 0 & 1 \\ . & . & 0 \end{bmatrix}$
		

introduce the random indicator variable $Y_\alpha(\mathbf{m})$ which equals one if motif $\mathbf{m}$ occurs at position α and 0 otherwise:

$$Y_\alpha(\mathbf{m}) = \prod_{1 \le u < v \le k} X_{i_u i_v}^{m_{uv}}.$$

Since the distribution of $\mathbf{X}$ is exchangeable, the distribution of $Y_\alpha(\mathbf{m})$ does not depend on α, and $Y_\alpha(\mathbf{m})$ is distributed according to a Bernoulli distribution $\mathcal{B}(\mu(\mathbf{m}))$, where $\mu(\mathbf{m})$ is the probability of occurrence of motif $\mathbf{m}$ at any position.

2.4. *Motif permutation*

Considering the occurrence of the ∨ motif at position $\alpha = (i_1, i_2, i_3)$ (Table 1), one can see that ∨ occurs at α with a given permutation on indices. This is why we need to define $\mathcal{R}(\mathbf{m})$, the set of non redundant permutations of $\mathbf{m}$, and we denote $\rho(\mathbf{m}) = |\mathcal{R}(\mathbf{m})|$, which equals 3 in the case of the ∨ motif, and 1 for the triangle. Note that $\rho(\mathbf{m}) = k!/|\text{aut}(\mathbf{m})|$, where aut($\mathbf{m}$) is the set of automorphisms of motif $\mathbf{m}$: $\text{aut}(\mathbf{m}) = \{\sigma \in \mathfrak{S},\ \sigma(\mathbf{m}) = \mathbf{m}\}$, with $\mathfrak{S}$ the set of permutations on the vertices of $\mathbf{m}$. We consider permutations of the motif rather than permutations of positions.

From a practical point of view, we propose to avoid the calculation of $|\text{aut}(\mathbf{m})|$, and to focus on $\rho(\mathbf{m})$. This calculation can be done by considering the $k!$ simultaneous permutations of the rows and columns of $\mathbf{m}$, each new element being compared with the previous ones to check for redundancy. The complexity of this method is then in $\mathcal{O}(k!^2)$ and does not depend on the size of the complete graph. Moreover, since we are searching for small-size motifs ($k = 3, 4$ typically), the computational time of this procedure is moderate.

2.5. *Number of occurrences of* m

Finally we define $N(\mathbf{m})$ the count of motif $\mathbf{m}$ such that:

$$N(\mathbf{m}) = \sum_{\alpha \in I_k} \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}').$$

Considering example in Figure 1, there is one occurrence of the triangle at position $(1, 2, 3)$, six occurrences of the ∨ motif: one occurrence of $\mathbf{m}$ and $\mathbf{m}''$ at $(1, 2, 3)$, and four occurrences of $\mathbf{m}'$ at positions $(1, 2, 3)$, $(1, 3, 4)$, $(2, 3, 4)$, $(3, 4, 5)$ (with $\mathbf{m}, \mathbf{m}', \mathbf{m}''$ defined in Table 1).

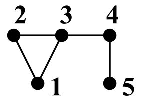


FIG. 1. A graph with 1 ▽ and 6 ∨.

3. CALCULATING MOMENTS UNDER AN EXCHANGEABLE MODEL

In this section, we aim at providing an automatic method to calculate the first and second moments of $N(\mathbf{m})$. This method requires the knowledge of $\mu(\mathbf{m})$, the probability of occurrence of motif $\mathbf{m}$. In a first step, we develop our method with $\mu(\mathbf{m})$ as a general term. This probability depends on the distribution of $\mathbf{X}$ and its derivation under different models will be given in the next section.

3.1. *Calculating the mean*

This calculation can be done directly since the distribution of Y_α does not depend on α. Indeed, the exchangeability assumption implies that permutations of motif $\mathbf{m}$ have the same probability of occurrence ($\mu(\mathbf{m}) = \mu(\mathbf{m}'),\ \forall \mathbf{m}' \in \mathcal{R}(\mathbf{m})$). It follows that:

$$\mathbb{E}N(\mathbf{m}) = |I_k| \times \sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} \mathbb{E}Y_\alpha(\mathbf{m}') = \binom{n}{k}\rho(\mathbf{m})\mu(\mathbf{m}). \qquad (1)$$

3.2. *Calculating the variance*

This calculation is based on the expectation of the squared count:

$$N^2(\mathbf{m}) = \sum_{\alpha,\beta \in I_k} \sum_{\mathbf{m}',\mathbf{m}'' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}')Y_\beta(\mathbf{m}''), \qquad (2)$$

and each term of this sum depends on the cardinality of the intersection $\alpha \cap \beta$ denoted by s. When $s = 0$, meaning that positions α and β are disjoint, the independence assumption between Y_α and Y_β leads to $\mathbb{E}\left[Y_\alpha(\mathbf{m})Y_\beta(\mathbf{m})\right] = \mathbb{E}Y_\alpha(\mathbf{m})\mathbb{E}Y_\beta(\mathbf{m})$. For $s \geq 1$, $\mathbf{m}'$ at α and $\mathbf{m}''$ at β share s vertices. Then we consider all possible overlaps between the two versions of $\mathbf{m}$ occurring at each position. We define the overlapping operation with s common vertices (denoted by $\underset{s}{\Omega}$) between motifs $\mathbf{m}'$ and $\mathbf{m}''$. Consequently,

$$\forall s = |\alpha \cap \beta| \geq 1,\ Y_\alpha(\mathbf{m}')Y_\beta(\mathbf{m}'') = Y_{\alpha\cup\beta}(\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''),$$

where $\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''$ represents what we call a "super-motif," which is a motif with $(2k - s)$ edges made of two overlapping occurrences of $\mathbf{m}'$ and $\mathbf{m}''$, two versions of $\mathbf{m}$. An example of super-motif built from the ∵ motif is provided in Figure 2.

To define the adjacency matrix of the super-motif $\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''$, we break down $\mathbf{m}'$ and $\mathbf{m}''$ such that

$$\mathbf{m}' = \left(\begin{array}{c|c} \underset{(k-s)\times(k-s)}{\mathbf{m}'_{11}} & \underset{(k-s)\times s}{\mathbf{m}'_{12}} \\ \hline \underset{s\times(k-s)}{\mathbf{m}'_{21}} & \underset{s\times s}{\mathbf{m}'_{22}} \end{array}\right),$$

$$\mathbf{m}'' = \left(\begin{array}{c|c} \underset{s\times s}{\mathbf{m}''_{11}} & \underset{s\times(k-s)}{\mathbf{m}''_{12}} \\ \hline \underset{(k-s)\times s}{\mathbf{m}''_{21}} & \underset{(k-s)\times(k-s)}{\mathbf{m}''_{22}} \end{array}\right),$$

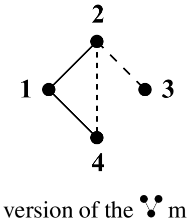


FIG. 2. Example of motif overlap. Let $\mathbf{m}$ be the version of the ∵ motif defined in Table 1; it is present at $\alpha = (1, 2, 4)$ and $\beta = (2, 3, 4)$. In this case $\alpha \cap \beta = (2, 4)$, and the corresponding super-motif $\mathbf{m}\underset{2}{\Omega}\mathbf{m}$ is the so-called whisk graph of size 4.

where $\mathbf{m}'_{22}$ and $\mathbf{m}''_{11}$ correspond to vertices in $\alpha \cap \beta$, and we set

$$\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'' = \left(\begin{array}{c|c|c} \mathbf{m}'_{11} & \mathbf{m}'_{12} & \mathbf{0} \\ \hline \mathbf{m}'_{21} & \max(\mathbf{m}'_{22}, \mathbf{m}''_{11}) & \mathbf{m}''_{12} \\ \hline \mathbf{0} & \mathbf{m}''_{21} & \mathbf{m}''_{22} \end{array} \right).$$

The max function in the central term indicates that for the s common vertices of α and β, all edges of $\mathbf{m}'_{22}$ and $\mathbf{m}''_{11}$ must be present; it is equivalent to the logical `OR`. Note that the operation $\underset{s}{\Omega}$ is not symmetric. Note that we also have to consider the number of possible super-motifs of type $\mathbf{m}'\underset{s}{\Omega}\mathbf{m}''$ which is $|\mathcal{R}(\mathbf{m})|^2$. The complexity of this enumeration is therefore smaller than $\mathcal{O}(k!^2)$.

The squared count can be rewritten as

$$N^2(\mathbf{m}) = \sum_{\substack{\alpha,\beta \in I_k: \\ |\alpha \cap \beta| = 0}} \sum_{\mathbf{m}',\mathbf{m}'' \in \mathcal{R}(\mathbf{m})} Y_\alpha(\mathbf{m}') Y_\beta(\mathbf{m}'')$$

$$+ \sum_{s=1}^{k} \sum_{\substack{\alpha,\beta \in I_k: \\ |\alpha \cap \beta| = s}} \sum_{\mathbf{m}',\mathbf{m}'' \in \mathcal{R}(\mathbf{m})} Y_{\alpha \cup \beta}(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}''),$$

and its expectation is:

$$\mathbb{E}N^2(\mathbf{m}) = \binom{n}{n-2k, k, k} \left[\sum_{\mathbf{m}' \in \mathcal{R}(\mathbf{m})} \mu(\mathbf{m}') \right]^2 \tag{3}$$

$$+ \sum_{s=1}^{k} \binom{n}{k-s, s, k-s, n-2k+s} \sum_{\mathbf{m}',\mathbf{m}'' \in \mathcal{R}(\mathbf{m})} \mu(\mathbf{m}' \underset{s}{\Omega} \mathbf{m}'').$$

The calculation of the first term follows from the independence of disjoint occurrences are independent. Put together, we can derive the formula for the variance of the count since $\mathbb{V}N(\mathbf{m}) = \mathbb{E}N^2(\mathbf{m}) - (\mathbb{E}N(\mathbf{m}))^2$.

3.3. *Calculating the occurrence probability*

Once the method to calculate the first and second moments of the count has been settled, we need to choose an appropriate model for the distribution of $\mathbf{X}$ in order to calculate $\mu(\mathbf{m})$, the probability of occurrence of motif $\mathbf{m}$. We derive this calculus in the case of three exchangeable random graph models.

Erdös-Rényi. The ER model assumes that all edges X_{ij} are independent and exist with probability π. Thanks to this independence property, the occurrence probability is a simple product:

$$\mu(\mathbf{m}) = \prod_{1 \le u < v \le k} \Pr\{X_{i_u i_v} = 1\}^{m_{uv}} = \prod_{1 \le u < v \le k} \pi^{m_{uv}} = \pi^{m_{++}/2},$$

where $m_{++} = \sum_{u,v} m_{uv}$ is twice the total number of edges in motif $\mathbf{m}$.

Expected degree distribution. This model generates graphs whose degrees follow a given distribution. It is defined as follows. Let D a random variable with a given distribution p, $p(d) = Pr\{D = d\}$, and $\{D_i\}$'s are i.i.d random variables with this distribution p. Conditionally to the $\{D_i\}$'s, edges $\{X_{ij}\}$'s are supposed to be independent and exist with a probability proportional to the product $D_i D_j$:

$$\begin{cases} \Pr\{X_{ij} = 1 \,|D_i, D_j\} = \gamma D_i D_j & \text{if } i = j; \\ \Pr\{X_{ij} = 1 \,|D_i, D_j\} = 0 & \text{otherwise.} \end{cases}$$

Denoting by $K_i = \sum_{j \neq i} X_{ij}$ the degree of node i, γ must be equal to $1/[(n-1)\mathbb{E}(D)]$ to insure that $\mathbb{E}(K_i | D_i) = D_i$.

In the EDD model, the conditional occurrence probability of the motif, given the expected degrees $\{D_i\}$, is

$$\Pr\{Y_\alpha(\mathbf{m}) = 1 \mid D_{i_1}, \dots, D_{i_k}\} = \gamma^{m_{++}/2} \prod_{u=1}^{k} D_{i_u}^{m_{u+}}.$$

The occurrence probability $\mu(\mathbf{m})$ is obtained by summation over the distribution of the D_is:

$$\mu(\mathbf{m}) = \sum_{d_{i_1},\dots,d_{i_k}} \prod_{u=1}^{k} p(d_{i_u}) \Pr\{Y_\alpha(\mathbf{m}) = 1 \mid D_{i_1} = d_{i_1}, \dots, D_{i_k} = d_{i_k}\}$$

$$= \gamma^{m_{++}/2} \prod_{u=1}^{k} \sum_{d_{i_u}} p(d_{i_u}) d_{i_u}^{m_{u+}} = \gamma^{m_{++}/2} \prod_{u=1}^{k} \mathbb{E}(D_u^{m_{u+}}).$$

Thus, the occurrence probability of $\mathbf{m}$ only depends on the product of some moments of the expected degree D.

Mixture model. The ERMG model is defined as follows: Nodes are spread among Q hidden classes with respective proportions $\alpha_1, \dots, \alpha_Q$. Edges $\{X_{ij}\}$ are independent conditionally to the class of the nodes. The connexion probability depends on the classes of both nodes such that:

$$\Pr\{X_{ij} = 1 \mid i \in q, j \in \ell\} = \pi_{q\ell}.$$

In the ERMG model, the conditional occurrence probability of the motif given the class of each node is:

$$\Pr\{Y_\alpha(\mathbf{m}) \mid i_1 \in c_1, \dots, i_k \in c_k\} = \prod_{1 \le u < v \le k} \pi_{c_u c_v}^{m_{uv}}.$$

The occurrence probability $\mu(\mathbf{m})$ of motif $\mathbf{m}$ is then

$$\mu(\mathbf{m}) = \sum_{c_1=1}^{Q} \cdots \sum_{c_k=1}^{Q} \alpha_{c_1} \dots \alpha_{c_k} \prod_{1 \le u < v \le k} \pi_{c_u c_v}^{m_{uv}}.$$

Calculus are illustrated in Table 2 for motifs and in the three models, and a computational trick is provided in the Appendix in the case of ERMG.

TABLE 2. OCCURRENCE PROBABILITIES OF THE AND MOTIF IN THREE MODELS

Motif	*ER*	*EDD*	*ERMG*
	π^2	$\gamma^2[\mathbb{E}(D)]^2\mathbb{E}(D^2)$	$\sum_{c_1=1}^{Q}\sum_{c_2=1}^{Q}\sum_{c_3=1}^{Q} \alpha_{c_1}\alpha_{c_2}\alpha_{c_3}\pi_{c_1c_2}\pi_{c_1c_3}$
	π^3	$\gamma^3[\mathbb{E}(D^2)]^3$	$\sum_{c_1=1}^{Q}\sum_{c_2=1}^{Q}\sum_{c_3=1}^{Q} \alpha_{c_1}\alpha_{c_2}\alpha_{c_3}\pi_{c_1c_2}\pi_{c_1c_3}\pi_{c_2c_3}$

3.4. Application to PPI networks

To illustrate the calculation of moments, we studied the PPI networks of *H. pylori*, *E. coli*, and *S. cerevisæ* from the DIP database (Salwinski et al., 2004). We consider three exchangeable models, whose parameters were estimated such that:

ER: The connexion probability π of the ER model is estimated by the proportion of observed edges in the network.
EDD-E: We calculated the empirical (E) distribution of the degrees in the network and used it as the distribution of the expected degrees $\{D_i\}$. This corresponds to the typical use of the EDD model.
ERMG: We fitted the mixture model using the variational method described in Daudin et al. (2008).

We also consider the FDD model for which the first two moments are estimated using simulations. For each model, we considered all undirected motifs of size $k = 3$ and 4. Table 3 gives the corresponding moments.

A first general remark is that the expectation and the variance are of different magnitude. This indicates that Poisson approximations for the count distribution (Barbour, 1982) would not be suitable, and will not be used in the following. We see that the choice of the model has a strong influence on the first two moments, which depends on the topology of the motifs.

Table 3. Expectations and Standard Deviations for the Count of all 3- and 4-Motifs in Three PPI Networks

		$\mathbb{E}N(\mathbf{m})$				$\sqrt{\mathbb{V}N(\mathbf{m})}$			
	N_{obs}	*ER*	*EDD-E*	*ERMG*	$\widehat{\mathbb{E}N(\mathbf{m})}$, *FDD*	*ER*	*EDD-E*	*ERMG*	$\sqrt{\widehat{\mathbb{V}N(\mathbf{m})}}$, *FDD*
Hpylo									
	14,113	5704.08	13,549.90	13,602.97	14,113	311.08	3611.74	2659.18	0
	75	10.85	196.63	66.91	52.82	3.40	102.06	20.41	7.80
	98,697	22,880.50	142,771.00	94,578.81	84,115.87	1919.66	62,567.50	27,039.88	3324.82
	112,490	7626.83	101,428.00	93,741.08	112,490	681.76	46,043.30	27,257.36	0
	1058	32.64	1553.90	516.66	284.88	6.89	1105.05	208.76	27.40
	3535	130.55	13,247.10	2897.13	2410.48	44.07	9617.37	1120.34	452.89
	79	0.37	614.58	34.80	22.09	0.66	666.63	20.00	9.88
	0	0.00	20.26	0.17	0.10	0.02	32.40	0.45	0.32
Ecoli									
	248,093	52,744.70	99,126.40	243,846.93	248,093	1281.87	20,851.70	51,676.68	0
	11,368	72.47	2197.38	10,221.17	3579.49	8.90	797.30	3041.98	68.58
	9,557,956	399,151.00	2,339,200.00	9,555,414.55	5,950,903.40	14,743.70	774,109.00	3,019,630.93	67,739.86
	6,425,495	133,050.00	1,537,740.00	5,772,005.15	6,425,495	5089.62	484,152.00	1,672,086.51	0
	487,408	411.31	38,890.60	417,190.55	76,467.39	29.14	19,122.60	170,502.21	1117.56
	2,154,048	1645.22	306,789.00	1,929,516.68	547,802.44	214.52	145,764.00	739,836.65	15,593.00
	273,621	3.39	20,117.90	204,093.45	18,422.25	2.04	12,876.60	94,018.80	891.99
	14,882	0.00	867.24	8904.75	317.27	0.05	707.94	4660.71	32.96
Scere									
	436,131	123,668.00	87,993.40	389,503.34	436,131	1900.49	14,409.10	43,699.24	0
	10,567	58.74	253.21	4499.68	596.01	7.78	97.79	1026.22	27.78
	7,530,597	873,206.00	1,011,160.00	6,453,832.37	5,643,320.62	20,415.70	292,154.00	984,085.02	83,158.13
	12,227,236	291,069.00	1,065,250.00	7,974,881.99	12,227,236	7065.36	427,908.00	1,653,822.72	0
	165,085	311.08	2182.33	86,658.32	7659.12	20.29	1133.64	35,938.74	196.22
	993,733	1244.32	27,588.80	442,611.59	76,434.54	170.94	17,622.20	144,261.26	5851.70
	116,667	0.89	376.37	40,118.23	276.40	0.97	407.40	18,259.56	51.11
	8601	0.00	5.41	1959.07	0.43	0.01	10.71	993.27	0.66

$\mathbb{E}N(\mathbf{m})$ and $\sqrt{\mathbb{V}N(\mathbf{m})}$ have been theoretically calculated under ER, EDD-E and ERMG models. $\widehat{\mathbb{E}N(\mathbf{m})}$ and $\sqrt{\widehat{\mathbb{V}N(\mathbf{m})}}$ are the empirical estimators calculated using simulations under the FDD model.

and **:** The expected count under the ER and EDD-E models are very different. This is due to nodes with high degree ($D \simeq 50$) which are observed in the empirical distribution and which generate lots of occurrences of these motifs. In the ER model, the probability for a node to have such a degree is about 10^{-35}. The larger standard deviation obtained under EDD-E is due to the random sampling among the degrees.

: For *H. pylori* and *E. coli*, the expected number of triangles under ERMG is close to the observed one, while the other models are quite far. This also holds for and motifs which reveal local clustering. This clustering trend is well captured by the ERMG model which detects communities of nodes.

Then Table 3 shows that moments are lower for the FDD models, especially for the variance, since the conservation of the empirical degree distribution consitutes a strong constraint when generating networks. The extreme case is given by the and the , whose counts are exactly specified by the empirical degree distribution, leading to a null variance (this is true for all star motifs with more than three branches). This comparative study of the first two moments of the counts reveals that the reference model will have a deep impact on the assessment of motifs exceptionality, and this point will be investigated in the last section.

4. COMPOUND POISSON APPROXIMATION

The knowledge of the moments of the count under some null model is not sufficient to assess its significance. To decide whether a motif **m** is over-represented in a given network, one needs to calculate the probability $\Pr\{N(\mathbf{m}) \geq N_{\text{obs}}(\mathbf{m})\}$, where $N_{\text{obs}}(\mathbf{m})$ is the observed number of occurrences of **m** and $N(\mathbf{m})$ the random number of occurrence under the reference model. To do so, we need to specify the distribution of $N(\mathbf{m})$ under the reference model. Unfortunately, even in the Erdös-Rényi model, the exact distribution seems very difficult to derive, so only an approximate distribution can be proposed at this time.

4.1. Motivations

One particularity of network motifs is that their occurrences naturally tend to overlap. Two occurrences of a motif **m** overlap if they share at least one vertex. Thus, the approximate distribution must account for the existence of clumps, i.e., sets of overlapping occurrences. Clumps result in numerous occurrences with a reduced number of vertices. For example, an occurrence of the four-branch star motif accounts for four overlapping occurrences of the three-branch star motif, i.e., for a clump of size 4 involving only five vertices. Another example is provided with the whisk motif in Figure 2, which shows five occurrences of the motif, leading to a clump of size 5 involving four vertices.

Compound Poisson distributions are particularly relevant to describe how the count of events occurring in clumps may vary. The number of clumps is supposed to have a Poisson distribution with mean λ, and the clump sizes are supposed to be independent with common distribution. The Pólya-Aeppli (denoted by $\mathcal{PA}$) distribution (or geometric Poisson, according to Johnson et al. [1992]) is obtained when the clump size has a geometric distribution $\mathcal{G}(1-a)$, so the mean size of a clump is $(1-a)^{-1}$. In this case, the number of observed events W has distribution $\mathcal{PA}(\lambda, a)$, and $\Pr\{W = w\}$ is equal to:

$$\begin{cases} e^{-\lambda} a^{w} \displaystyle\sum_{c=1..w} \frac{1}{c!} \binom{w-1}{c-1} \left[\frac{\lambda(1-a)}{a}\right]^{c} & \text{if } w > 0, \\ e^{-\lambda} & \text{if } w = 0. \end{cases}$$

We propose to use the Pólya-Aeppli distribution as an approximation of the distribution of the count $N(\mathbf{m})$ for two main reasons. (i) This distribution is an excellent approximation (from both a theoretical and a practical point of view) for word counts in random sequences (Schbath, 1995; Robin and Schbath, 2001). (ii) The Pólya-Aeppli distribution only involves two parameters that can be easily computed when the first two moments are known. Note that argument (i) is not sufficient because the topology of a network motif is quite different from the topology of a sequence motif. Still, parameter a could be interpreted as the overlapping probability of the motif **m**, i.e., the probability that an occurrence of **m** overlaps another one.

The first two moments of the $\mathcal{PA}(\lambda, a)$ distribution are $\lambda/(1-a)$ and $\lambda(1+a)/(1-a)^2$. Given these moments, parameters can be calculated as:

$$a = [\mathbb{V}N(\mathbf{m}) - \mathbb{E}N(\mathbf{m})]/[\mathbb{E}N(\mathbf{m}) + \mathbb{V}N(\mathbf{m})] \tag{4}$$

$$\lambda = (1-a)\mathbb{E}N(\mathbf{m}).$$

p-values can be calculated using the algorithms given in Barbour et al. (1992) or in Nuel (2007).

4.2. *Simulation-based comparison of distribution approximations*

Simulation design. The objective of the simulations is to compare the Gaussian and the Póly-Aeppli approximations for the distribution of the motif counts. To do so, we focus on all three- and four-node undirected motifs, and we use the PPI networks of three organisms provided by the DIP database (Salwinski et al., 2004). Our aim is to study the approximation quality whatever the underlying random graph model. For each PPI, 10,000 random graphs are generated according to the following models: FDD, EDD, and ERMG models. Since the FDD and the EDD models are based on the observed degree distributions, we also check that the ERMG model leads to degree distributions which are in accordance with the observed networks (Fig. 3). Technical details regarding this aspect of the ERMG model can be found in Daudin et al. (2008).

As previously, the FDD model can not be theoretically used in our framework, and we use empirical moments based on simulations to calculate parameters (a, λ), and we propose to compare the Gaussian and the compound Poisson approximations in this setting.

Quality of approximation. Comparisons are based on two criteria:

- The Kolmogorov-Smirnov distance $KS = \sup_i |\hat{F}(i) - F(i)|$ between the theoretical (F) and empirical $(\hat{F})$ cumulative distribution functions. This criterion allows a global comparison along the whole range of a random variable.

Since we are more concerned by the tails of the distribution for computing the p-values of the count, we use a second quality criterion:

- $1 - \hat{F}(Q_{\mathcal{N}})$ and $1 - \hat{F}(Q_{\mathcal{PA}})$ are the empirical probabilities of exceeding the 0.999 Gauss and Pólya-Aeppli quantiles respectively. This criterion should be close to 0.001 for a good approximation.

Note that these criteria will not be calculated in the case of the and the under the FDD model, as the empirical variance of their respective counts is nul as explained previously.

Results:

1. A first result is that the shape of the count distributions highly depends on the model (Fig. 4). Typically, the FDD model generates symetrical distributions, whereas the EDD model generates highly skewed and peaked distributions whatever the motifs and the network, as shown in Table 4. Distributions from the ERMG show intermediate skewness and kurtosis. The fact that the FDD model leads to count distributions which are symetrical reflects how constraint is the model. This symmetry, as well as the small variance of the count means that visited configurations among simulations are very similar. On the contrary, highly skewed distributions (EDD model) indicate that extreme configurations are explored (with high values for the count). This behavior is linked to the simulation procedure, in which the expected degree is sampled using the observed degree.
2. Then the Kolmogorov-Smirnov distance indicates that the Pólya-Aeppli approximation allows a better fit to the count distribution, compared with the Gaussian approximation. This result is consistent accross motifs and accross networks, with few exceptions (Tables 5–7).
3. Tables 5–7 show that the 0.999 quantile is systematically under-estimated by the Gaussian approximation. Consequently, the use of the Gaussian approximation can lead to false positive results: some motifs could be thought as being exceptional, whereas they are not.

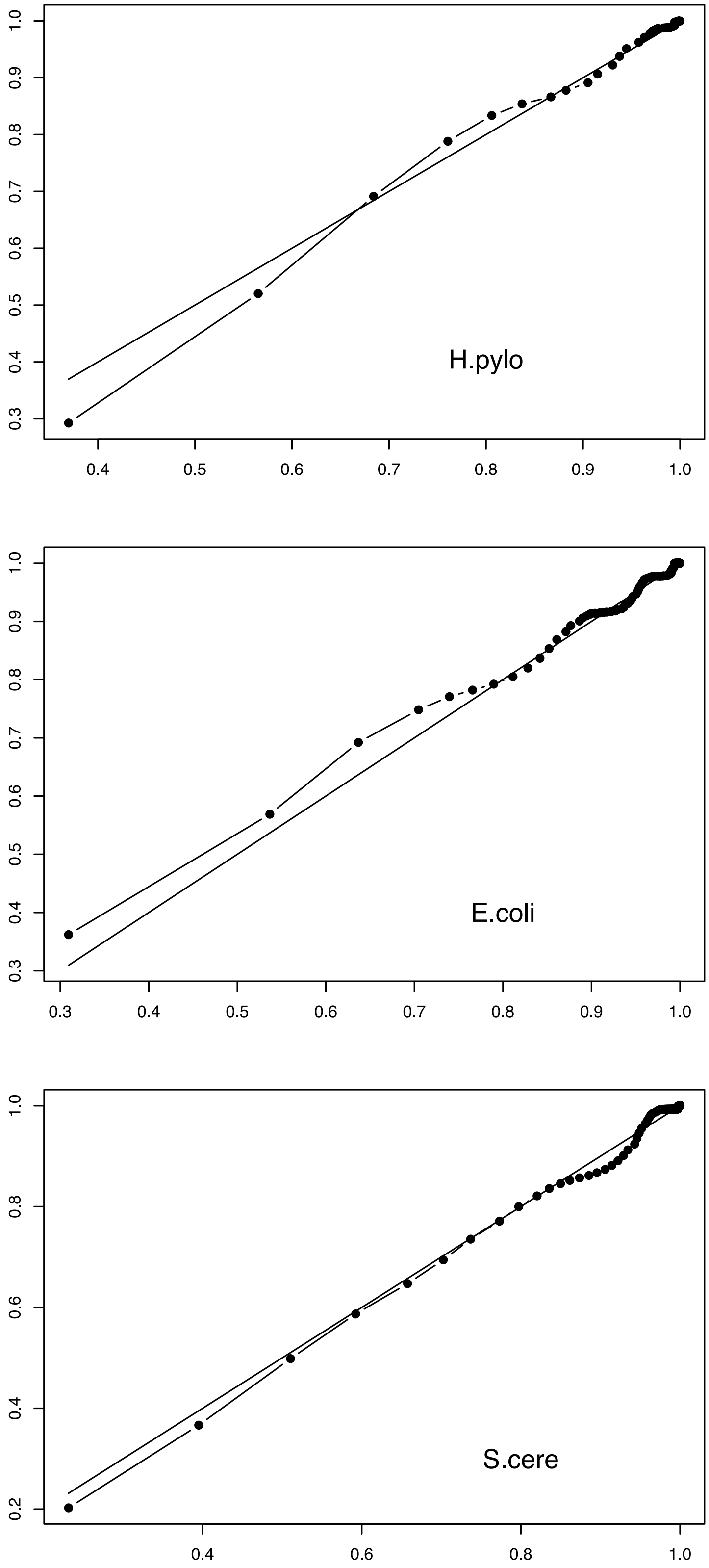
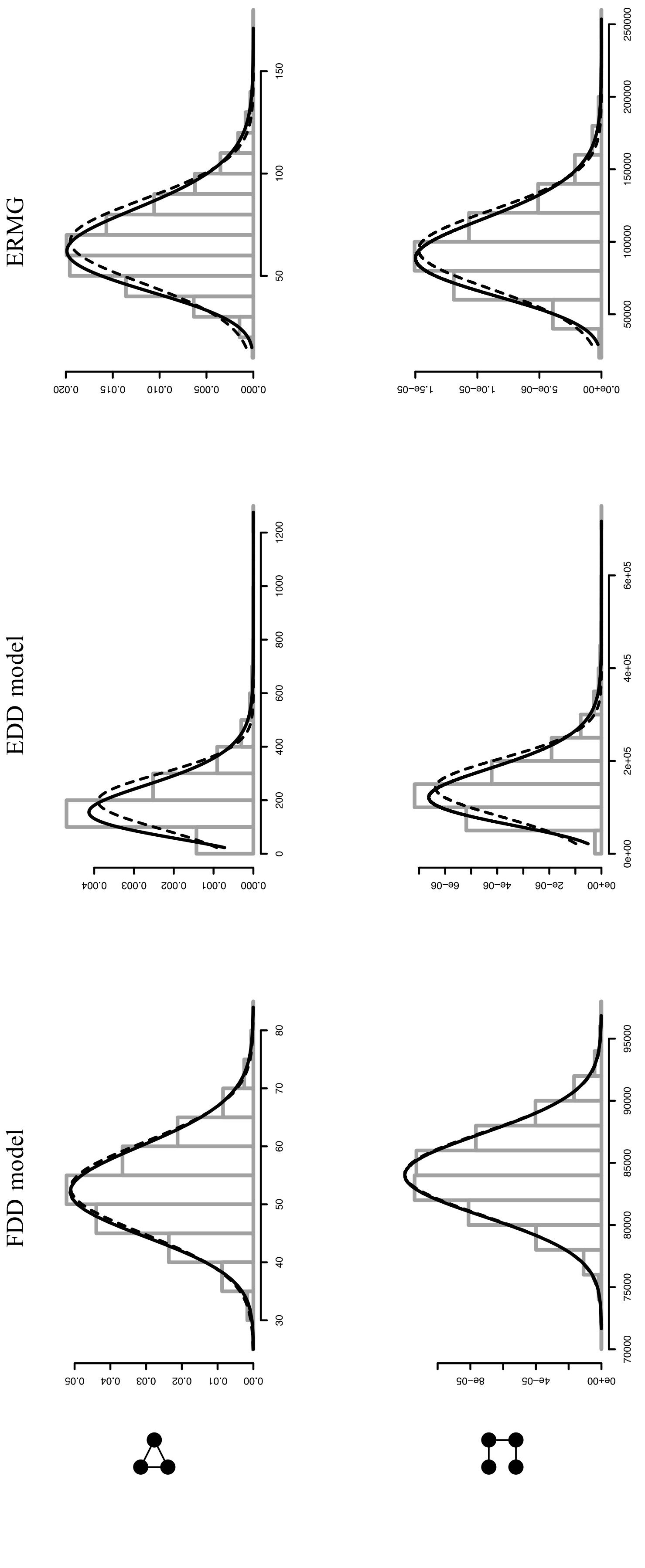


FIG. 3. PP-plots of the fitted degree distributions for three protein-protein interaction (PPI) networks using the Erdös-Rényi Mixture for Graphs (ERMG) model.


FDD model
EDD model
ERMG
0.00
0.01
0.02
0.03
0.04
0.05
30
40
50
60
70
80
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
1000
1200
0.000
0.005
0.010
0.015
0.020
50
100
150
0e+00
4e−05
8e−05
70000
75000
80000
85000
90000
95000
0e+00
2e−06
4e−06
6e−06
0e+00
2e+05
4e+05
6e+05
0.0e+00
5.0e−06
1.0e−05
1.5e−05
50000
100000
150000
200000
250000

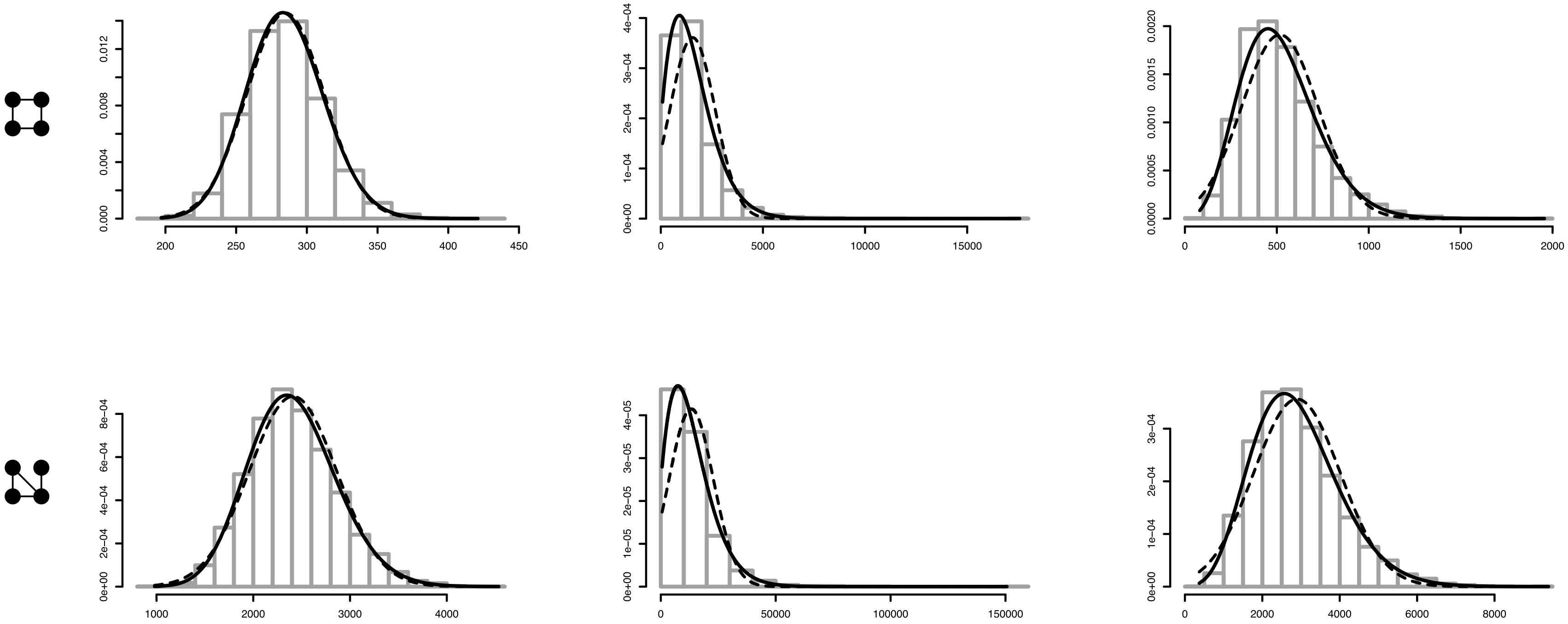


FIG. 4. Count distribution according to three models: fixed-degree distribution (FDD), expected-degree distribution (EDD), and Erdös-Rényi Mixture for Graphs (ERMG) models. Plain line, $\mathcal{PA}$; dotted line, $\mathcal{N}$.

Table 4. Empirical Skewness and Kurtosis of the Count Distributions

	Skewness			*Kurtosis*		
	FDD	*EDD*	*ERMG*	*FDD*	*EDD*	*ERMG*
Hpylo						
	—	0.79	0.35	—	4.10	3.13
	0.18	1.55	0.59	3.05	7.59	3.45
	0.17	1.34	0.64	2.97	6.34	3.59
	—	1.22	0.44	—	5.83	3.28
	0.34	2.36	0.97	3.26	14.40	4.56
	0.40	2.27	0.80	3.23	13.58	3.94
	1.08	3.89	1.21	4.95	37.78	5.30
	3.60	6.49	3.01	17.64	101.81	15.48
Ecoli						
	—	0.54	0.50	—	3.49	3.36
	0.03	0.93	0.73	2.94	4.48	3.88
	0.04	0.89	0.80	2.98	4.40	4.07
	—	0.81	0.70	—	4.20	3.80
	0.19	1.35	1.04	3.12	6.24	4.87
	0.10	1.27	0.98	3.01	5.90	4.66
	0.20	1.77	1.19	3.10	8.68	5.53
	0.29	2.32	1.37	3.19	13.05	6.46
Scere						
	—	0.61	0.23	—	3.55	3.05
	0.05	1.30	0.76	3.01	5.89	3.89
	0.12	1.01	0.41	3.04	4.62	3.27
	—	0.97	0.32	—	4.34	3.12
	0.18	1.78	1.27	3.01	8.43	5.55
	0.24	1.88	1.11	3.09	9.02	4.97
	0.73	3.34	1.25	4.10	22.21	5.40
	1.60	6.34	1.48	5.79	70.29	7.31

5. EXCEPTIONAL MOTIFS IN PPI NETWORKS

In this last section, we propose to assess the exceptionality of all three- and four-node undirected motifs in three PPI networks (Table 8). This assessment is done using the FDD, EDD, and ERMG models, and the Gaussian and the Pólya-Aeppli approximations.

The first result is that the magnitude of the p-values strongly depends on the reference model. In the case of the FDD model, observed counts are so far from the expectation that p-values are close to zero. On the contrary, for ERMG, p-values are moderate, and can always be calculated. This tendency is reinforced by the approximation used to calculate the p-values, as already mentioned in the simulation study: using the Gaussian approximation leads to the assessment of more exceptional motifs, and regarding the simulation study, some of those exceptional motifs are likely to be false positives. Nevertheless, Table 8 clearly shows that the first factor influencing the procedure is the reference model, and not the distribution approximation.

Even if our primary objective was not to provide an appropriate probabilistic reference model for biological networks, these results suggest some comments. Since the count of every motif is exceptional when using the FDD model, one explanation is that this model is too simple to account for the observed biological variability. As previously mentioned, the degree sequence is one of the characteristics of biological networks, and using this characteristic only does not seem to be sufficient to model the complexity of biological networks. Then using the EDD model could be an alternative, since the model considers the expected degree distribution, and shows more flexibility than the FDD model. However, considering the

TABLE 5. QUALITY APPROXIMATION OF THE COUNT DISTRIBUTION FOR **HPYLO** PPI NETWORK

	$\widehat{\mathbb{E}N(\mathbf{m})}$	$\sqrt{\widehat{\mathbb{V}(\mathbf{m})}}$	KS$_{\mathcal{N}}$	KS$_{\mathcal{PA}}$	$1 - \hat{F}(Q_{\mathcal{N}})$	$1 - \hat{F}(Q_{\mathcal{PA}})$
FDD						
	14,113	0	—	—	—	—
	52.82	7.80	4.26	0.92	1.70	0.20
	84,115.87	3324.82	1.34	0.96	1.30	1.00
	112,490	0	—	—	—	—
	284.88	27.40	3.00	1.37	2.50	1.50
	2410.48	452.89	3.43	1.65	3.90	0.80
	22.09	9.88	8.71	2.52	9.60	2.70
	0.10	0.32	52.88	0.04	0.80	0.00
EDD						
	13,549.90	3611.74	5.62	3.02	6.80	2.80
	196.63	102.06	9.93	4.57	12.50	4.70
	142,771.00	62,567.50	8.63	4.30	11.30	4.00
	101,428.00	46,043.30	7.85	3.31	10.40	3.30
	1553.90	1105.05	12.86	8.45	15.60	5.20
	13,247.10	9617.37	12.35	7.78	15.40	4.80
	614.58	666.63	18.22	19.94	18.10	4.50
	20.26	32.40	21.63	39.40	17.50	3.50
ERMG						
	13,602.97	2659.18	2.48	0.56	2.90	0.70
	66.91	20.41	5.41	1.42	4.80	1.00
	94,578.81	27,039.88	4.40	1.78	5.90	0.90
	93,741.08	27,257.36	2.97	0.81	4.60	0.70
	516.66	208.76	6.32	2.30	8.90	2.40
	2897.13	1120.34	5.60	1.82	6.90	1.60
	34.80	20.00	9.18	2.60	10.10	2.10
	0.17	0.45	50.06	0.15	2.40	0.50

KS, Kolmogorov-Smirnov distance (×100); $1 - \hat{F}(Q)$, empirical probability of exceeding the 0.999 quantile (×1000); FDD, fixed-degree distribution; EDD, expected-degree distribution; ERMG, Erdös-Rényi mixture for graphs.

expected degree distribution instead of the fixed degree distribution does not lead to convicing results regarding Table 8: motifs are either all exceptional or all non exceptional. Such drastic behavior, may be linked to a variable quality of fit of the model to the data. Finally, the ERMG model leads to moderate results. While no motif is exceptional in the PPI network of *E. coli*, the and motifs are exceptional at the 5% level in *H. pylori* network, and 6 over 8 undirected motifs are exceptional in the PPI network of *S. cerevisiæ*, with moderate p-values. This could indicate that ERMG could be an appropriate reference model for PPI networks. However, this will have to be further explored.

6. CONCLUSION

We provide an exact method to calculate the mean and variance of the count of any network motif, whatever its topology. These formula hold for any exchangeable random graph model satisfying the independence property of disjoint motif occurrences. The generalization of our method to the directed case is straightforward. In this case adjacency matrices $\mathbf{X}$ and $\mathbf{m}$ are not symmetric anymore, and formulas to calculate the mean and variance still hold. Moreover, our method can be applied when the purpose is to count strict occurrences of a motif only: Instead of counting the number of subgraphs of G which are isomorphic to a given motif, for instance to a , one may be interested in counting the so-called *induced* subgraphs. For instance, one may wish to count no occurrence of the motif in a (rather than three

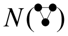
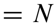
TABLE 6. QUALITY APPROXIMATION OF THE COUNT DISTRIBUTION FOR **ECOLI** PPI NETWORK

	$\widehat{\mathbb{E}N(\mathbf{m})}$	$\sqrt{\widehat{\mathbb{V}(\mathbf{m})}}$	$\mathrm{KS}_{\mathcal{N}}$	$\mathrm{KS}_{\mathcal{PA}}$	$1-\hat{F}(Q_{\mathcal{N}})$	$1-\hat{F}(Q_{\mathcal{PA}})$
FDD						
	248,093	0	—	—	—	—
	3579.49	68.58	0.91	0.52	0.80	0.60
	5,950,903.40	67,739.86	0.62	0.51	1.20	1.20
	6,425,495	0	—	—	—	—
	76,467.39	1117.56	2.04	1.87	2.30	2.00
	547,802.44	15,593.00	0.95	0.82	1.30	1.10
	18,422.25	891.99	1.63	1.16	1.90	1.40
	317.27	32.96	2.69	1.13	2.90	1.60
EDD-E						
	99,126.40	20,851.70	2.74	1.17	4.80	2.00
	2197.38	797.30	5.55	2.34	8.90	3.20
	2,339,200.00	774,109.00	5.00	2.18	9.10	2.90
	1,537,740.00	484,152.00	4.52	1.94	7.90	2.40
	38,890.60	19,122.60	7.83	4.08	13.00	4.10
	306,789.00	145,764.00	7.50	3.58	12.10	3.90
	20,117.90	12,876.60	10.02	6.00	14.40	4.40
	867.24	707.94	12.61	9.76	16.80	4.50
ERMG						
	243,846.93	51,676.68	3.30	1.33	4.80	2.20
	10,221.17	3041.98	4.64	1.77	7.30	2.20
	9,555,414.55	3,019,630.93	5.13	2.03	7.90	2.70
	5,772,005.15	1,672,086.51	4.54	1.65	6.90	2.50
	417,190.55	170,502.21	6.50	2.56	10.30	2.50
	1,929,516.68	739,836.65	6.14	2.34	9.50	2.70
	204,093.45	94,018.80	7.30	3.06	11.00	3.00
	8904.75	4660.71	8.46	3.90	12.50	3.50

KS, Kolmogorov-Smirnov distance (×100); $1-\hat{F}(Q)$, empirical probability of exceeding the 0.999 quantile (×1000); FDD, fixed-degree distribution; EDD, expected-degree distribution; ERMG, Erdös-Rényi mixture for graphs.

occurrences with our current definition). It means to take care of absent edges in the motif, i.e., to consider the new random indicator

$$Y'_\alpha(\mathbf{m}) = \prod_{1\le u<v\le k} X_{i_u,i_v}^{m_{uv}}(1-X_{i_u,i_v})^{(1-m_{uv})}.$$

The number of induced occurrences of motif $\mathbf{m}$ denoted by $N'(\mathbf{m})$, is thus simply the sum of these indicators over all positions α and all versions of the motif $\mathbf{m}$. Explicit formulas for the mean and variance of $N'(\mathbf{m})$ can be deduced from our results. The key argument is that the count $N'(\mathbf{m})$ can be expressed like a linear combination of counts of the form $N(\mathbf{w})$ for some motifs $\mathbf{w}$ of size k (Kocay, 1981). For instance, $N'($ $) = N($$)$ and $N'($ $) = N($$) - 3N($ $)$. Getting the expectation $\mathbb{E}N'(\mathbf{m})$ is then trivial and getting the variance just requires the expression of the covariance between the counts $N(\mathbf{m_1})$ and $N(\mathbf{m_2})$ of two motifs $\mathbf{m_1}$ and $\mathbf{m_2}$ of size k. This covariance is equal to $\mathbb{E}N(\mathbf{m_1})N(\mathbf{m_2}) - \mathbb{E}N(\mathbf{m_1})\mathbb{E}N(\mathbf{m_2})$ and the first term is obtained like for Equation (3):

$$\mathbb{E}N(\mathbf{m_1})N(\mathbf{m_2}) = \binom{n}{n-2k,k,k} \sum_{\mathbf{m}'_1\in\mathcal{R}(\mathbf{m_1}),\mathbf{m}'_2\in\mathcal{R}(\mathbf{m_2})} \mu(\mathbf{m}'_1)\mu(\mathbf{m}'_2)$$

$$+\sum_{s=1}^{k}\binom{n}{k-s,s,k-s,n-2k-s} \sum_{\mathbf{m}'_1\in\mathcal{R}(\mathbf{m_1}),\mathbf{m}''_2\in\mathcal{R}(\mathbf{m_2})} \mu(\mathbf{m}'_1 \underset{s}{\Omega} \mathbf{m}''_2).$$

Table 7. Quality Approximation of the Count Distribution for **Scere** PPI Network

	$\widehat{\mathbb{E}N(\mathbf{m})}$	$\sqrt{\widehat{\mathbb{V}(\mathbf{m})}}$	KS$_{\mathcal{N}}$	KS$_{\mathcal{PA}}$	$1-\hat{F}(Q_{\mathcal{N}})$	$1-\hat{F}(Q_{\mathcal{PA}})$
FDD						
	436,131	0	—	—	—	—
	596.01	27.78	1.54	0.46	1.00	0.50
	5,643,320.62	83,158.13	1.08	0.94	1.70	1.60
	12,227,236	0	—	—	—	—
	7659.12	196.22	1.62	1.28	1.50	1.30
	76,434.54	5851.70	1.79	1.13	2.00	1.40
	276.40	51.11	5.05	2.87	6.10	3.80
	0.43	0.66	39.62	0.08	1.50	0.00
EDD-E						
	87,993.40	14,409.10	4.83	3.21	4.80	3.10
	253.21	97.79	8.93	5.06	11.80	3.50
	1,011,160.00	292,154.00	7.15	4.36	8.60	3.30
	1,065,250.00	427,908.00	6.87	2.97	7.80	2.10
	2182.33	1133.64	11.19	6.59	15.40	4.60
	27,588.80	17,622.20	11.73	7.18	15.60	4.10
	376.37	407.40	18.64	21.09	20.20	4.10
	5.41	10.71	25.44	35.90	17.70	2.70
ERMG						
	389,503.34	43,699.24	1.48	0.52	2.40	1.60
	4499.68	1026.22	5.41	3.18	7.00	3.20
	6,453,832.37	984,085.02	2.90	1.42	3.80	2.30
	7,974,881.99	1,653,822.72	2.38	0.42	3.30	1.00
	86,658.32	35,938.74	8.15	4.35	11.50	3.90
	442,611.59	144,261.26	7.35	4.29	10.40	4.60
	40,118.23	18,259.56	8.60	4.25	12.10	3.90
	1959.07	993.27	9.29	4.35	11.90	3.20

KS, Kolmogorov-Smirnov distance (×100); $1-\hat{F}(Q)$, empirical probability of exceeding the 0.999 quantile (×1000); FDD, fixed-degree distribution; EDD, expected-degree distribution; ERMG, Erdös-Rényi mixture for graphs.

The main difficulty when searching for exceptional network motifs is that the theoretical count distribution remains unknown in real networks. Consequently, getting exact results on the moments of the count is of primary interest for the future characterization of this distribution. Our approach focuses on the first two moments, but can be extended to moments of any order.

Since no theoretical result is yet available on the motif count distribution, we proposed an approximation, which is based on the Pólya-Aeppli distribution. We showed that this approximation is accurate in a large range of situations, and we demonstrated that the Gaussian approximation is not satisfactory. Consequently, strategies based on z-scores such as the method proposed by Shen-Orr et al. (2002) are not reliable. In addition, p-values can be easily computed thanks to the Pólya-Aeppli approximation we propose. However, let us recall that when using the Pólya-Aeppli approximation, the underlying hypothesis is that the distribution of the size of the motif clumps is geometric, which is unlikely to be true. Future developments will be needed to theoretically address the distribution of these clumps size.

Our approach is based on direct computations avoiding simulations that would be very numerous to be accurate in the case of small p-values. Typically, a p-value of about 10^{-5} would require at least 10^{7} simulations. From a historical point of view, the question of motif exceptionality has first arisen in the case of DNA sequences analysis. In this context, the first shuffling-based approaches were rapidly competed by Markov models which allowed the derivation of statistical tools without any simulation. In the case of network motifs, similar developments should be done and our results constitute one step towards this direction.

TABLE 8. EXCEPTIONAL MOTIFS IN THREE PPI NETWORKS USING THREE REFERENCE MODELS AND TWO DISTRIBUTION APPROXIMATIONS

	N_{obs}	FDD-Pv$_{\mathcal{N}}$	FDD-Pv$_{\mathcal{PA}}$	EDD-Pv$_{\mathcal{N}}$	EDD-Pv$_{\mathcal{PA}}$	ERMG-Pv$_{\mathcal{N}}$	ERMG-Pv$_{\mathcal{PA}}$
Hpylo							
	14,113	—	—	4.38e−01	4.13e−01	4.24e−01	4.06e−01
	75	**2.23e−03**	**4.36e−03**	8.83e−01	9.06e−01	3.46e−01	3.31e−01
	98,697	**5.78e−06**	**1.22e−05**	7.59e−01	7.42e−01	4.39e−01	4.12e−01
	112490	—	—	4.05e−01	3.65e−01	2.46e−01	2.34e−01
	1058	**2.15e−175**	**1.80e−52**	6.73e−01	6.15e−01	**4.76e−03**	**1.33e−02**
	3535	**6.51e−03**	**1.11e−02**	8.44e−01	8.58e−01	2.85e−01	2.63e−01
	79	**4.27e−09**	**2.54e−05**	7.89e−01	7.51e−01	**1.35e−02**	**3.11e−02**
	0	6.18e−01	1.00e−00	7.34e−01	1.00e−00	6.52e−01	8.50e−01
Ecoli							
	248,093	—	—	**4.53e−13**	**1.24e−08**	4.67e−01	4.46e−01
	11,368	**0.00e+00**	**0.00e+00**	**6.43e−31**	**7.02e−13**	3.53e−01	3.30e−01
	9,557,956	**0.00e+00**	**0.00e+00**	**5.54e−21**	**2.33e−10**	5.00e−01	4.68e−01
	6,425,495	—	—	**2.89e−24**	**1.14e−11**	3.48e−01	3.26e−01
	487,408	**0.00e+00**	**0.00e+00**	**5.90e−122**	**3.48e−23**	3.40e−01	3.10e−01
	2,154,048	**0.00e+00**	**1.03e−265**	**4.18e−37**	**1.15e−12**	3.81e−01	3.49e−01
	273,621	**0.00e+00**	**1.24e−115**	**1.39e−86**	**1.09e−17**	2.30e−01	2.14e−01
	14,882	**0.00e+00**	**2.61e−41**	**1.59e−87**	**3.30e−15**	9.98e−02	1.09e−01
Scere							
	436,131	—	—	**2.86e−129**	**6.21e−33**	1.43e−01	1.44e−01
	10,567	**0.00e+00**	**1.31e−128**	**0.00e+00**	**1.13e−22**	**1.69e−09**	**1.21e−06**
	7,530,597	**2.51e−114**	**8.44e−99**	**1.32e−110**	**8.61e−27**	1.37e−01	1.38e−01
	12,227,236	—	—	**2.70e−150**	**3.11e−22**	**5.07e−03**	**9.54e−03**
	165,085	**0.00e+00**	**1.09e−322**	**0.00e+00**	**3.19e−22**	**1.45e−02**	**2.73e−02**
	993,733	**0.00e+00**	**1.64e−65**	**0.00e+00**	**9.33e−22**	**6.66e−05**	**8.90e−04**
	116,667	**0.00e+00**	**1.71e−33**	**0.00e+00**	**1.28e−18**	**1.38e−05**	**7.22e−04**
	8601	**0.00e+00**	**1.54e−10**	**0.00e+00**	**3.19e−16**	**1.14e−11**	**5.25e−06**

The p-values of <5% are in boldface.

7. APPENDIX

Practical calculation of $\mu(\mathbf{m})$

We recall that when using ERMG as a reference model, the probability of occurrence of motif $\mathbf{m}$ is:

$$\mu(\mathbf{m}) = \sum_{c_1=1}^{Q}\cdots\sum_{c_k}^{Q}\alpha_{c_1},\ldots,\alpha_{c_k}\prod_{1\le u<v\le k}\pi_{c_u,c_v}^{m_{u,v}}. \qquad \text{(A.1)}$$

The computation of this formula might have a complexity as high as $O(Q^k)$ since the products might have to be computed for each $(\alpha_{c_1},\ldots,\alpha_{c_k}) \in [\![1,Q]\!]^k$ (depending on the values of the binary digits $m_{u,v}$). Since the computation of each sum indexed say by $u \in \{1,\ldots,k\}$ is of complexity $O(Q^{d_u+1})$, where d_u is the number of indexes v such that $m_{u,v} = 1$, we propose to reduce the computation time by computing the indexes of lowest degrees first.

Let $1 \le u \le k$ be an index and let us call $\mathbf{m}$-degree of u the cardinal of the set $\{1 \le v \le k,\ m_{u,v} = 1\}$. We define recursively the *reduced* $\mathbf{m}$*-degree* of any index the following way:

1. The indexes of lowest $\mathbf{m}$-degree have a reduced $\mathbf{m}$-degree equal to their $\mathbf{m}$-degree.
2. The indexes whose reduced $\mathbf{m}$-degree is computed are removed from the formula.
3. If there remains indexes go back to step 1, else end.

Lemma. The computation time of formula (A.1) is $O(kQ^{D+1})$ where D is the maximum of the reduced **m**-degrees of the indexes $u \in [\![1,k]\!]$.

Proof. Let us call $d_{\min}$ the lowest **m**-degree of the indexes. Let u be any index of **m**-degree $d_{\min}$. The reduced **m**-degree of u is $d_{\min}$. We compute all the terms involving u in formula (A.1), which costs $O(Q^{d_{\min}+1})$. Then formula (A.1) involves $k-1$ indexes. The **m**-degrees of the indexes previously involved in the computations of the terms containing the index u have decreased by 1 and the others did not change.

The first time any index v is of minimal **m**-degree is when its **m**-degree equals its reduced **m**-degree d. Consequently the computation of the terms involving v costs Q^{d+1} products at most. This proves that the computation time is at most Q^{D+1} for the removal of any index. This has to be done k times. Hence the total computation time is $O(kQ^{D+1})$. ■

Remark. Removing the indexes whose reduced **m**-degree has been computed decreases the degrees of the remaining indexes. Hence, the reduced **m**-degree of an index is at most its degree.

REFERENCES

Barabási, A.L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286, 509–512.

Barbour, A. 1982. Poisson convergence and random graphs. *Math. Proc. Camb. Phil. Soc.* 92, 349–359.

Barbour, A., Karoński, M., and Ruciński, A. 1987. A central limit theorem for decomposable random variables with applications to random graphs. *J. Combinat. Theor. Ser. B* 457, 125–145.

Barbour, A.D., Holst, L., and Janson, S. 1992. *Poisson Approximation*. Oxford University Press, New York.

Batada, N., Hurst, L., and Tyers, M. 2006. Evolutionary and physiological importance of hub proteins. *PLOS Comp. Biol.* 2, 748–756.

Bollobas, B. 1981. Random graphs. *Combinatorics. London Mathematics Society Lecture Note Series 52*. Cambridge University Press, New York.

Chen, J., and Yuan, B. 2006. Detecting functional modules in the yeast protein-protein intereaction network. *Bioinformatics* 22, 2283–2290.

Chen, Y., and Dokholyan, N. 2006. The coordinated evolution of yeast proteins constrained by functional modularity. *Trends Genet.* 22, 416–419.

Chung, F., and Lu, L. 2002. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. USA* 99, 15879–15882.

Daudin, J.-J., Picard, F., and Robin, S. 2008. A mixture model for random graphs. *Statistics and Computing* (in press). DOI: 1007/s11222-007-9046-7.

Erdös, P. 1947. Some remarks on the theory of graphs. *Bull. Am. Math. Soc.* 53, 292–294.

Erdös, P., and Rényi, A. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* 5, 17–61.

Ingram, P., Stumpf, M.P., and Stark, J. 2006. Network motifs: structure does not determine function. *BMC Genom.* 7, 108.

Janson, S., Rucinski, A., and Luczak, T. 2000. *Random Graphs*. Wiley, New York.

Johnson, N.L., Kotz, S., and Kemp, A.W. 1992. *Univariate Discrete Distributions*. Wiley, New York.

Karoński, M., and Ruciński, A. 1983. On the number of strictly balanced subgraphs of a random graph. *Lect. Notes Math.* 1018, 79–83.

Kocay, W. 1981. An extension of Kelly's lemma to spanning subgraphs. *Congr. Num.* 31, 109–120.

Koyutürk, M., Szpankowski, W., and Grama, A. 2007. Assessing significance of connectivity and conservation in protein interaction networks. *J. Comp. Biol.* 14, 747–764.

Lee, T., Rinaldi, N., Robert, F., et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804.

Mangan, S., and Alon, U. 2003. Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. USA* 100, 11980–11985.

Matias, C., Schbath, S., Birmelé, E., et al. 2006. Network motifs: mean and variance for the count. *REVSTAT* 4, 1–20.

Middendorf, M., Ziv, E., and Wiggins, C. 2005. Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. *Proc. Natl. Acad. Sci. USA* 102, 3192–3197.

Milo, R., Itzkovitz, S., Kashtan, N., et al. 2004. Superfamilies of evolved and designed networks. *Science* 303, 1538–1542.

Milo, R., Kashtan, N., Itzkovitz, S., et al. 2004. On the uniform generation of random graphs with prescibed degree sequences. *Cond-Mat* 0312028, pg. 1–4.

Milo, R., Shen-Orr, S., Itzkovitz, S., et al. 2002. Networks motifs: simple building blocks of complex networks. *Science* 298, 824–827.

Newman, M.E.J. 2003. *Handbook of Graphs and Networks*. Wiley-VCH, Berlin.

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.

Nowicki, K., and Snijders, T. 2001. Estimation and prediction for stochastic blockstructures. *J. Am. Statist. Assoc.* 96, 1077–1087.

Nuel, G. 2007. Cumulative distribution function of a geometric Poisson distribution. *J. Stat. Comp. Sim.* epub. DOI: 1080/629360600997371.

Park, J., and Newman, M. 2003. The origin of degree correlations in the internet and other networks. *Phys. Rev. E* 68, 026112.

Prill, R., Iglesias, P.A., and Levchenko, A. 2005. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol.* 3, 11.

Robin, S., and Schbath, S. 2001. Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comput. Biol.* 8, 349–359.

Salwinski, L., Miller, C., Smith, A., et al. 2004. The database of interacting proteins: 2004 update. *NAR* 449–451.

Schbath, S. 1995. Compound Poisson approximation of word counts in DNA sequences. *ESAIM Probab. Statist.* 1, 1–16 (*www.emath.fr/ps/*).

Shen-Orr, S.S., Milo, R., Mangan, S., et al. 2002. Networks motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68.

Stark, D. 2001. Compound Poisson approximation of subgraph counts in random graphs. *Random Struct. Algorithms* 18, 39–60.

Wuchty, S., Oltvai, Z., and Barabasi, A.-L. 2003. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.* 35, 176–179.

Address reprint requests to:
Dr. F. Picard
Laboratoire Statistique et Génome
UMR CNRS 8071–INRA 1152
Université d'Evry
523 place des Terrasses
F-91000 Evry, France

E-mail: picard@genopole.cnrs.fr