Stat Comput
DOI 10.1007/s11222-007-9046-7

A mixture model for random graphs

J.-J. Daudin · F. Picard · S. Robin

Received: 18 December 2006 / Accepted: 14 November 2007

Abstract The Erdös–Rényi model of a network is simple and possesses many explicit expressions for average and asymptotic properties, but it does not fit well to real-world networks. The vertices of those networks are often structured in unknown classes (functionally related proteins or social communities) with different connectivity properties. The stochastic block structures model was proposed for this purpose in the context of social sciences, using a Bayesian approach. We consider the same model in a frequentest statistical framework. We give the degree distribution and the clustering coefficient associated with this model, a variational method to estimate its parameters and a model selection criterion to select the number of classes. This estimation procedure allows us to deal with large networks containing thousands of vertices. The method is used to uncover the modular structure of a network of enzymatic reactions.

Keywords Random graphs · Mixture models · Variational method

1 Introduction

The Erdös–Rényi model of a network is one of the oldest and best studied model and possesses many explicit expressions for average and asymptotic properties such as degree distribution, connectedness and clustering coefficient. However this theoretical model does not fit well to real-world, social, biological or Internet networks. For example the empirical degree distribution may be very different from the Poisson distribution which is implied by this model. Moreover empirical clustering coefficients of real networks are generally higher than the value given by this model. Other models have been proposed to correct these shortcomings (see Nowicki and Snijders 2001; Albert and Barabási 2002; or Molloy and Reed 1995). A good review of random graph models is given in Pattison and Robins (2007).

It appears that the available methods in the literature can be divided into two categories: model-based versus algorithmic methods. In the context of social sciences and using a Bayesian approach, the stochastic block structures model (Nowicki and Snijders 2001) assumes that vertices pertain to classes with different connectivity characteristics. This model provides a proper probabilistic framework but the proposed estimation method can not deal with networks made of more than 200 vertices. However, a special attention has been recently paid to the study of biological networks which are generally much larger (see Alm and Arkin 2002; or Arita 2004). Other algorithms have been proposed: assortative mixing or mixing patterns (Newman and Girvan 2003; Newman 2004). These methods are efficient on large networks but the absence of model makes the interpretation of the results more difficult.

The key element of those methods is the mixing matrix which specifies the probability of connection between two classes. The inference of the mixing parameters is quite easy if the classes can be defined using external information such as language, race or age. However the inference is more difficult when classes and mixing parameters have to be inferred when the network topology is the only available information.

In this article we use the model-based framework proposed by Nowicki and Snijders (2001) in a frequentist setting. We derive some new theoretical properties of this

J.-J. Daudin · F. Picard · S. Robin (✉)
Mathématiques et Informatique Appliquées, AgroParisTech and
INRA UMR518, 16, rue Clause Bernard, 75005 Paris, France
e-mail: robin@inapg.fr

model. We provide an estimation algorithm using a variational approach as well as a model selection criterion to choose the number of classes. This framework allows us to deal with thousands of vertices. Our method is illustrated on a biological network.

Notation In this article, we consider an undirected graph with n vertices and define the variable X_{ij} which indicates that vertices i and j are connected:

$$X_{ij} = X_{ji} = \mathbb{I}\{i \leftrightarrow j\},$$

where $\mathbb{I}\{A\}$ equals to one if A is true, and to zero otherwise. Furthermore, we assume that no vertex is connected to itself, meaning that $X_{ii} = 0$. However, the method we present below can be generalized to directed graphs ($X_{ij} \neq X_{ji}$) with self loops ($X_{ii} \neq 0$). In the following we note K_i the degree of vertex i, i.e. the number of edges connecting it:

$$K_i = \sum_{j \neq i} X_{ij}.$$

Erdös–Rényi model This model assumes that edges are independent and occur with the same probability p:

$$\{X_{ij}\} \text{ i.i.d.,} \quad X_{ij} \sim \mathcal{B}(p).$$

In this model, the degree of each vertex has a Binomial distribution, which is approximately Poisson for large n and small p. Noting $\lambda = (n-1)p$ we have

$$K_i \sim \mathcal{B}(n-1, p) \approx \mathcal{P}(\lambda).$$

2 Mixture model for the degrees

In many practical situations, the Erdös–Rényi model turns out to fit the data poorly, mainly because the distribution of the degrees is far from the Poisson distribution. The scale-free (or Zipf) distribution has been intensively used as an alternative. The Zipf probability distribution function (pdf) is

$$\Pr\{K_i = k\} = c(\rho) k^{-(\rho+1)}, \tag{1}$$

where k is any positive integer, ρ is positive, $c(\rho) = \sum_{k \geq 1} k^{-(\rho+1)} = 1/\zeta(\rho+1)$ and $\zeta(\rho+1)$ is Riemann's zeta function. Nevertheless, we will show in Sect. 6 that this distribution may have a poor fit on real datasets. This lack of fit has been already analyzed by Stumpf et al. (2005) and Tanaka and Doyle (2005).

First of all, it is important to notice that the Zipf distribution is used to model the tail of the degree distribution. Therefore it is often best suited for the tail than for the whole distribution. In particular this distribution has a null probability for $k = 0$ whereas some vertices may be unconnected in practice. Moreover the lack-of-fit of the Erdös–Rényi model may be simply due to some heterogeneities between vertices, some being more connected than others. A simple way to model this phenomenon is to consider that the degree distribution is a mixture of Poisson distributions.

In the mixture framework we suppose that vertices are structured into Q classes, and that there exists a sequence of independent hidden variables $\{Z_{iq}\}$ (with $\sum_q Z_{iq} = 1$) which indicate the label of vertices to classes. We note α_q the *prior* probability for vertex i to belong to class q, such that:

$$\alpha_q = \Pr\{Z_{iq} = 1\} = \Pr\{i \in q\}, \quad \text{with } \sum_q \alpha_q = 1.$$

Remark 1 In the following, we will use two equivalent notations: $\{Z_{iq} = 1\}$ or $\{i \in q\}$ to indicate that vertex i belongs to class q.

We suppose that the conditional distribution of the degrees is a Poisson distribution

$$K_i \mid \{i \in q\} \sim \mathcal{P}(\lambda_q).$$

Then the distribution of the degrees is a mixture of Poisson distributions such that

$$\Pr\{K_i = k\} = \sum_q \alpha_q \frac{e^{-\lambda_q} \lambda_q^k}{k!}. \tag{2}$$

For a complete review and a careful statistical analysis of the modeling of the degree distribution in networks, see (Jones and Handcock 2004).

Remark 2 Because vertices are connected, degrees are not independent. However, in the standard situation where n is large and $\lambda_q \ll n$, the dependency between degrees is weak.

In Sect. 6 we show that this model fits well to real data. Nevertheless, we claim that modeling the distribution of the degrees provides little information about the topology of the graph. Indeed, this model only deals with the degrees of vertices, but not explicitly with the probability for two given vertices to be connected. However, the observed number of connections between vertices from different classes may reveal some interesting underlying structure, such as preferential connections between classes. The mixture model for degrees is not precise enough to describe such a phenomenon. This motivates the definition of an explicit mixture model for edges.

3 Erdös–Rényi mixture for graphs

3.1 General model

We now describe the stochastic block structures model (Nowicki and Snijders 2001), a mixture model which explicitly describes the way edges connect vertices, accounting for some heterogeneity among vertices. In the following this model is called "mixture model for graphs".

The mixture model for graphs supposes that vertices are spread into Q classes with prior probabilities $\{\alpha_1, \dots, \alpha_Q\}$. In the following, we use the indicator variables $\{Z_{iq}\}$ (with $\sum_q Z_{iq} = 1$) defined in Sect. 2.

$$\alpha_q = \Pr\{Z_{iq} = 1\} = \Pr\{i \in q\}, \quad \text{with } \sum_q \alpha_q = 1.$$

Then we denote $\pi_{q\ell}$ the probability for a vertex from class q to be connected with a vertex from class ℓ. Because the graph is undirected, these probabilities must be symmetric such that

$$\pi_{q\ell} = \pi_{\ell q}.$$

We finally suppose that edges $\{X_{ij}\}$ are conditionally independent given the classes of vertices i and j:

$$\begin{cases} X_{ij} \mid \{i \in q, j \in \ell\} \sim \mathcal{B}(\pi_{q\ell}) & \text{for } i \neq j, \\ X_{ii} = 0. \end{cases}$$

The main difference with Model (2) is that the mixture model for graphs directly deals with edges. More than describing the clustered structure of vertices, our model describes the topology of the network using the connectivity matrix $\boldsymbol{\pi} = (\pi_{q\ell})$.

3.2 Examples

In this section we aim at showing that the mixture model for graphs can be used to generalize many particular structures of random graphs. Table 1 presents some typical network configurations. The first one is the Erdös–Rényi model. We present here some more sophisticated ones.

Example 1 Random graphs with arbitrary degree distributions. The Erdös–Rényi random graph model is a poor approximation of real-world networks whose degree distribution is highly skewed. A random network having the same degree distribution as the empirical one can be built as follows: n partial edges (with only one starting vertex and no final vertex) are randomly chosen from the empirical degree distribution. These partial edges are randomly joined by pairs to form complete edges (see Molloy and Reed 1995). A permutation algorithm is also proposed in Shen-Orr et al. (2002). This model assumes that the connectivity between two vertices is proportional to the degree of each vertex so it coincides with the independent case of the mixture model for graphs presented in Sect. 4.4.

The scale-free network proposed by Barabási and Albert (1999) is a particular case of random graphs with arbitrary distribution. To this extent, we can propose an analogous model in the mixture model for graphs framework. Suppose that the incoming vertices join the network in classes of respective size $n\alpha_q$ ($q = 1, \dots, Q$, $n\alpha_1$ being the number of original vertices). Assuming that the elements of a new class connect preferentially the elements of the oldest classes:

$$\pi_{q,1} \geq \pi_{q,2} \geq \cdots \geq \pi_{q,q-1},$$

we get the same kind of structure as the scale-free model.

Example 2 Affiliation network. An affiliation network is a social network in which actors are joined by a common participation in social events, companies boards or scientists' coauthorship of papers. All the vertices participating to the same class are connected. This model has been studied by Newman et al. (2002). This type of network may be modeled by a mixture model for graphs with ones in the diagonal of $\boldsymbol{\pi}$.

Example 3 Star pattern. Many biological networks contain star patterns, i.e. many vertices connected to the same vertex and only to it, see interaction networks of *S. Cerevisiae* in Zhang et al. (2005) for instance. This type of pattern may be modeled by a mixture model for graphs with extra-diagonal ones in $\boldsymbol{\pi}$.

4 Some properties of the mixture model for graphs

4.1 Distribution of the degrees

Proposition 1 *Given the label of a vertex, the conditional distribution of the degree of this vertex is Binomial (approximately Poisson):*

$$K_i \mid \{i \in q\} \sim \mathcal{B}(n-1, \overline{\pi}_q) \approx \mathcal{P}(\lambda_q),$$

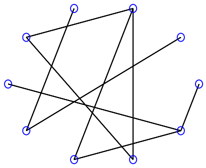
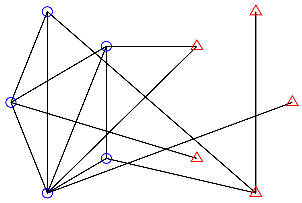
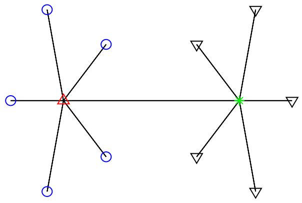
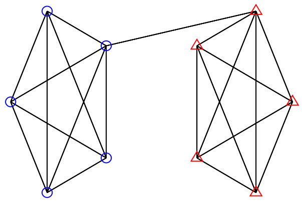
where $\overline{\pi}_q = \sum_\ell \alpha_\ell \pi_{q\ell}$ *and* $\lambda_q = (n-1)\overline{\pi}_q$.

Proof Conditionally to the belonging of vertices to classes, edges connecting vertex i belonging to class q are independent. The conditional connection probability is:

$$\begin{aligned} \Pr\{X_{ij} = 1 \mid i \in q\} &= \sum_\ell \Pr\{X_{ij} = 1 \mid i \in q, j \in \ell\} \Pr\{j \in \ell\} \\ &= \sum_\ell \alpha_\ell \pi_{q\ell} = \overline{\pi}_q. \end{aligned}$$

The result follows. □

Table 1 Some typical network configurations and their formulation in the framework of the mixture model for graphs. The node marks ($\circ$, $\triangle$, $\triangledown$, $\star$) refer to their class

Description	Network	Q	π	Clustering coef.
Random		1	p	p
Product connectivity (arbitrary degree distribution)		2	$\begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix}$	$\frac{(a^2+b^2)^2}{(a+b)^2}$
Stars		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	0
Clusters (affiliation networks)		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$	$\frac{1+3\varepsilon^2}{(1+\varepsilon)^2}$

4.2 Between-class connectivity

Definition 1 The connectivity between class q and ℓ is the number of edges connecting a vertex from class q to a vertex from class ℓ:

$$A_{q\ell} = \sum_{i<j} Z_{iq} Z_{j\ell} X_{ij}.$$

A_{qq} is the within-connectivity of class q.

Proposition 2 *The expected connectivity between class q and ℓ is*:

$$\mathbb{E}(A_{q\ell}) = n(n-1)\alpha_q \alpha_\ell \pi_{q\ell}/2.$$

Proof According to Definition 1, $A_{q\ell}$ is the sum over $n(n-1)/2$ terms. Conditionally to $\{Z_{iq}Z_{j\ell} = 1\}$, X_{ij} is a Bernoulli variable with parameter $\pi_{q\ell}$. Thus $\mathbb{E}(Z_{iq}Z_{j\ell}X_{ij}) = \mathbb{E}(Z_{iq}Z_{j\ell})\pi_{q\ell}$. The Z_{iq}s are independent, so we have $\mathbb{E}(Z_{iq}Z_{j\ell}) = \alpha_q\alpha_\ell$. The result follows. □

4.3 Clustering coefficient

This coefficient is supposed to measure the aggregative trend of a graph. Since no probabilistic modeling is usually available, this coefficient is empirically defined in most cases. Albert and Barabási (2002) propose the following definition of the empirical clustering coefficient for vertex i:

$$C_i = \nabla_i \Big/ \frac{K_i(K_i-1)}{2},$$

where ∇_i is the number of edges between the neighbors of vertex i: $\nabla_i = \sum_{j,k} X_{ij}X_{jk}X_{ik}/2$, whose minimum value is 0 and maximum value equals $K_i(K_i-1)/2$ for a clique. A first estimator of this empirical clustering coefficient is usually defined as the mean of the C_i's:

$$\widehat{c} = \sum_i C_i / n.$$

Denoting ∇ the "triangle" configuration ($i \leftrightarrow j \leftrightarrow k \leftrightarrow i$) and V the 'V' configuration ($j \leftrightarrow i \leftrightarrow k$) for any (i, j, k) uniformly chosen in $\{1, \dots, n\}$, the definition of C can be rephrased as $c = \Pr\{\nabla \mid \mathsf{V}\}$. Because ∇ is a particular case of V, we have

$$c = \Pr\{\nabla \cap \mathsf{V}\} / \Pr\{\mathsf{V}\} = \Pr\{\nabla\} / \Pr\{\mathsf{V}\}. \tag{3}$$

This property suggests another estimate of c proposed by Newman et al. (2002):

$$\widehat{c'} = 3 \sum_i \nabla_i \Big/ \sum_i V_i,$$

where V_i is the number of Vs in i: $V_i = \sum_{j>k,(j,k)\neq i} X_{ij} X_{ik}$. In the following we propose a probabilistic definition of this coefficient.

Definition 2 The clustering coefficient is the probability for two vertices j and k connected to a third vertex i, to be connected, with (i, j, k) uniformly chosen in $\{1, \dots, n\}$

$$c = \Pr\{X_{ij} X_{jk} X_{ki} = 1 \mid X_{ij} X_{ik} = 1\}.$$

Proposition 3 *In the mixture model for graphs, the clustering coefficient is*

$$c = \sum_{q,\ell,m} \alpha_q \alpha_\ell \alpha_m \pi_{q\ell} \pi_{qm} \pi_{\ell m} \Big/ \sum_{q,\ell,m} \alpha_q \alpha_\ell \alpha_m \pi_{q\ell} \pi_{qm}$$

Proof For any triplet (i, j, k), we have

$$\begin{aligned}\Pr\{\nabla\} = &\sum_{q,l,m} \alpha_q \alpha_\ell \alpha_m \\ &\times \Pr\{X_{ij} X_{jk} X_{ki} = 1 \mid i \in q, j \in \ell, k \in m\}, \\ = &\sum_{q,l,m} \alpha_q \alpha_\ell \alpha_m \pi_{q\ell} \pi_{qm} \pi_{\ell m}.\end{aligned}$$

The same reasoning can be applied to $\Pr\{\mathsf{V}\}$ recalling that the event V in (i, j, k) means that the top of V is i. The result is then an application of (3). □

4.4 Independent model

The model presented in Sect. 2 can be rephrased as an independent version of the mixture model for graphs. Indeed the absence of preferential connection between classes corresponds to the case where

$$\pi_{q\ell} = \eta_q \eta_\ell. \tag{4}$$

η_q is then the connection propensity of a vertex from class q, regardless the class of the other vertex. The properties of the independent model are as follows.

Distribution of degrees The conditional distribution of the degrees is Poisson with parameter λ_q such that

$$\lambda_q = (n-1)\eta_q \overline{\eta}, \tag{5}$$

where $\overline{\eta} = \sum_\ell \alpha_\ell \eta_\ell$, so λ_q is directly proportional to η_q.

Between class connectivity We get

$$\mathbb{E}(A_{q\ell}) = n(n-1)(\alpha_q \eta_q)(\alpha_\ell \eta_\ell)/2,$$

so the rows and columns of matrix $\mathbf{A} = (A_{q\ell})_{q,\ell}$ must all have the same profile. We will see in Sect. 6 that the observed number of connections between classes may be quite far from expected values.

Clustering coefficient

$$c = \frac{(\sum_q \alpha_q \eta_q^2)^2}{\overline{\eta}^2}.$$

For the standard Erdös–Rényi model ($Q = 1, \alpha_1 = 1, \overline{\eta} = \eta_1 = \sqrt{p}$), we get the known result: $c = \eta_1^4 / \eta_1^2 = p$.

Considering the independent case presented in Table 1 with $\alpha_1 = \alpha_2 = 1/2$ and $a = 0.9$, $b = 0.1$, we get $c = (0.9^2 + 0.1^2)^2 \simeq 0.67$. The corresponding Erdös–Rényi model with $p = (\alpha_1 a + \alpha_2 b)^2 = 1/4$ would lead to a strong underestimation of c since $c = p = 0.25$.

4.5 Likelihoods

In order to define the likelihood of the model, we use the incomplete-data framework defined by Dempster et al. (1977). Let $\mathcal{X}$ denote the set of all edges: $\mathcal{X} = \{X_{ij}\}_{i,j=1,\dots,n}$, and $\mathcal{Z}$ the set of all indicator variables for vertices: $\mathcal{Z} = \{Z_{iq}\}_{i=1,n}^{q=1,Q}$.

Proposition 4 *The complete-data log-likelihood is*

$$\begin{aligned}\log \mathcal{L}(\mathcal{X}, \mathcal{Z}) = &\sum_i \sum_q Z_{iq} \log \alpha_q \\ &+ \frac{1}{2} \sum_{i \neq j} \sum_{q,\ell} Z_{iq} Z_{j\ell} \log b(X_{ij}; \pi_{q\ell}).\end{aligned}$$

Proof We have $\log \mathcal{L}(\mathcal{X}, \mathcal{Z}) = \log \mathcal{L}(\mathcal{Z}) + \log \mathcal{L}(\mathcal{X} \mid \mathcal{Z})$ where

$$\log \mathcal{L}(\mathcal{Z}) = \sum_i \sum_q Z_{iq} \log \alpha_q,$$

$$\log \mathcal{L}(\mathcal{X} \mid \mathcal{Z}) = \frac{1}{2} \sum_{i \neq j} \sum_{q,\ell} Z_{iq} Z_{j\ell} \log b(X_{ij}; \pi_{q\ell}),$$

and $b(x; \pi) = \pi^x (1-\pi)^{1-x}$. □

The likelihood of the observed data $\mathcal{L}(\mathcal{X})$ is obtained by summing the complete-data likelihood over all the possible values of the unobserved variables $\mathcal{Z}$. Unfortunately, this sum is not tractable and it seems that no simpler form can be derived.

5 Estimation

In this section we propose a variational approach to perform an approximate maximum likelihood inference on the parameters. We follow the general strategy described in Jordan et al. (1999) or in the tutorial by Jaakkola (2000). A similar strategy is used in the bi-clustering framework by Govaert and Nadif (2005). The general statistical properties of the resulting estimators have not been investigated yet. However, this approximation allows us to deal with large networks (several thousands of nodes) whereas the Bayesian strategy adopted by Nowicki and Snijders (2001) restricts the estimation to 200 nodes.

5.1 Dependency graph

The X_{ij}s are independent conditionally to the Z_{iq}s, but are marginally dependent. For estimation purpose, it is important to know if $\Pr\{Z_{iq}=1 \mid \mathcal{X}\}$ is equal to $\Pr\{Z_{iq}=1 \mid \mathcal{X}_i\}$, where $\mathcal{X}_i$ is the set of all possible edges connecting i. $\mathcal{X}_i$ is often called the set of neighbors of vertex i. In the following, we give a counter example to show that the notion of neighborhood can not be used in the mixture model for graphs framework.

Assume that the vertices are divided in two classes, whose connectivity matrix is diagonal with $\pi_{11}=1$ and $\pi_{22}=a$ and $0<a<1$. Let us consider 3 vertices i, j, k with $X_{ij}=X_{ik}=1$. The vertices i and j are in the same class because no connection is possible between vertices pertaining to two different classes. The same is true for vertices i and k. Therefore the three vertices are in the same class and we have $\Pr\{Z_{i1}=1 \mid \mathcal{X}_i, X_{jk}\}>0$ if $X_{jk}=1$ and $\Pr\{Z_{i1}=1 \mid \mathcal{X}_i, X_{jk}\}=0$ if $X_{jk}=0$. Therefore $\Pr\{Z_{iq}=1 \mid \mathcal{X}\}$ depends on all the network and not only on edges connecting to the vertex i.

This counter example clearly shows that no neighborhood can be considered in the framework of mixture model for graphs, since unconnected vertices provide as much information as connected vertices. This is why the likelihood can not be simplified for computation.

5.2 Variational approach

As often for incomplete data models, the likelihood of the observed data $\mathcal{L}(\mathcal{X})$ is not tractable. EM (Dempster et al. 1977) is the most popular algorithm for this kind of problem. Unfortunately, EM requires the computation of the conditional distribution $\Pr(\mathcal{Z} \mid \mathcal{X})$ which is itself not tractable, as explained above. Therefore, we choose a variational approach that aims at optimizing a lower bound of $\log \mathcal{L}(\mathcal{X})$, denoted by

$$\mathcal{J}(R_{\mathcal{X}})=\log \mathcal{L}(\mathcal{X})-\mathrm{KL}[R_{\mathcal{X}}(\cdot), \Pr(\cdot \mid \mathcal{X})],$$

where KL denotes the Kullback–Leibler divergence, $\Pr(\mathcal{Z} \mid \mathcal{X})$ is the true conditional distribution of the indicator variables $\mathcal{Z}$ given the data $\mathcal{X}$, and $R_{\mathcal{X}}$ an approximation of this conditional distribution. $\mathcal{J}(R_{\mathcal{X}})$ equals $\log \mathcal{L}(\mathcal{X})$ iff $R_{\mathcal{X}}(\cdot)=\Pr(\cdot \mid \mathcal{X})$. We emphasize that $R_{\mathcal{X}}$ depends on the data $\mathcal{X}$.

As shown above, we are not able to calculate $\Pr(\cdot \mid \mathcal{X})$, so we will look for the "best" (in terms of Kullback–Leibler divergence) $R_{\mathcal{X}}$ in a certain class of distributions. The estimation algorithm we propose will alternate the maximization of $\mathcal{J}(R_{\mathcal{X}})$ (i) with respect to $R_{\mathcal{X}}$ and (ii) with respect to parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$. Propositions 5 and 6 give the solutions of the optimization problems (i) and (ii) respectively.

Approximate conditional distribution $R_{\mathcal{X}}$ Denoting $\mathcal{Z}_i=\{Z_{i1},\dots Z_{iQ}\}$, we constraint $R_{\mathcal{X}}$ to have the following form:

$$R_{\mathcal{X}}(\mathcal{Z})=\prod_i h(\mathcal{Z}_i;\boldsymbol{\tau}_i)$$

where $\boldsymbol{\tau}_i=(\tau_{i1},\dots,\tau_{iQ})$ and $h(\cdot;\boldsymbol{\tau})$ denotes the multinomial distribution with parameter $\boldsymbol{\tau}$. τ_{iq} can be interpreted as an approximation of $\Pr\{Z_{iq}=1 \mid \mathcal{X}\}$. This corresponds to the mean field approximation, as presented in Jaakkola (2000).

Proposition 5 *Given parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$, the optimal variational parameters $\{\widehat{\boldsymbol{\tau}}_i\}=\arg\max_{\{\boldsymbol{\tau}_i\}}\mathcal{J}(R_{\mathcal{X}})$ satisfy the following fixed point relation*:

$$\widehat{\tau}_{iq} \propto \alpha_q \prod_{j\neq i}\prod_{\ell} b(X_{ij};\pi_{q\ell})^{\widehat{\tau}_{j\ell}}.$$

Proof Based on the definition of the Kullback–Leibler divergence, we first rewrite $\mathcal{J}(R_{\mathcal{X}})$ as

$$\begin{aligned}\mathcal{J}(R_{\mathcal{X}}) &= \sum_{\mathcal{Z}} R_{\mathcal{X}}(\mathcal{Z})\log \Pr\{\mathcal{Z},\mathcal{X}\} \\ &\quad - \sum_{\mathcal{Z}} R_{\mathcal{X}}(\mathcal{Z})\log R_{\mathcal{X}}(\mathcal{Z}) \\ &= \sum_i\sum_q \tau_{iq}\log\alpha_q \\ &\quad + \frac{1}{2}\sum_{i\neq j}\sum_{q,\ell}\tau_{iq}\tau_{j\ell}\log b(X_{ij};\pi_{q\ell}) \\ &\quad - \sum_i\sum_q \tau_{iq}\log\tau_{iq}.\end{aligned}$$

We now have to maximize $\mathcal{J}(R_{\mathcal{X}})$ with respect to the τ_{iq}'s, subject to $\sum_q \tau_{iq} = 1$, for all i, i.e. to maximize $\mathcal{J}(R_{\mathcal{X}}) + \sum_i [\lambda_i(\sum_q \tau_{iq} - 1)]$ where λ_i is the Lagrange multiplier. The derivative with respect to τ_{iq} is

$$\log \alpha_q + \sum_{j \neq i} \sum_{\ell} \tau_{j\ell} \log b(X_{ij}; \pi_{q\ell}) - \log \tau_{iq} + 1 + \lambda_i.$$

This derivative is null iff $\widehat{\tau}_{iq}$'s satisfy the relation given in the proposition, $\exp(1 + \lambda_i)$ being the normalizing constant. □

From a practical point of view, the $\{\widehat{\boldsymbol{\tau}}_i\}$ are updated using a fixed point algorithm. At this time, we have no guaranty about the convergence toward a unique solution. In all situations we experienced, the algorithm converged rapidly.

Parameter estimates To complete the estimation procedure, we need to maximize $\mathcal{J}(R_{\mathcal{X}})$ with respect to parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$.

Proposition 6 *Given the variational parameters $\{\boldsymbol{\tau}_i\}$, the values of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ that maximize $\mathcal{J}(R_{\mathcal{X}})$ are*

$$\widehat{\alpha}_q = \frac{1}{n} \sum_i \widehat{\tau}_{iq}, \qquad \widehat{\pi}_{q\ell} = \sum_{i \neq j} \widehat{\tau}_{iq} \widehat{\tau}_{j\ell} X_{ij} \Big/ \sum_{i \neq j} \widehat{\tau}_{iq} \widehat{\tau}_{j\ell}.$$

Proof Due to the constraint on $\boldsymbol{\alpha}$, we have to maximize $\mathcal{J}(R_{\mathcal{X}}) + \lambda(\sum_q \alpha_q - 1)$. The calculation of the derivatives is straightforward and the result follows. □

Estimation algorithm The algorithm we propose is the following. Starting with some initial values $\{\boldsymbol{\tau}_i^{(0)}\}$ for the variational parameters, we iteratively update parameters $\boldsymbol{\tau}_i$, $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ as follows:

$$\left(\boldsymbol{\alpha}^{(h+1)}, \boldsymbol{\pi}^{(h+1)}\right) = \underset{(\boldsymbol{\alpha},\boldsymbol{\pi})}{\arg\max}\, \mathcal{J}\left(R_{\mathcal{X}}; \{\boldsymbol{\tau}_i^{(h)}\}, \boldsymbol{\alpha}, \boldsymbol{\pi}\right),$$

$$\left\{\boldsymbol{\tau}_i^{(h+1)}\right\} = \underset{\{\boldsymbol{\tau}_i\}}{\arg\max}\, \mathcal{J}\left(R_{\mathcal{X}}; \{\boldsymbol{\tau}_i\}, \boldsymbol{\alpha}^{(h+1)}, \boldsymbol{\pi}^{(h+1)}\right).$$

These updates are performed according to Propositions 5 and 6.

Proposition 7 *For a given number of classes Q, this algorithm generates a sequence $\{\{\boldsymbol{\tau}_i^{(h)}\}, \boldsymbol{\alpha}^{(h)}, \boldsymbol{\pi}^{(h)}\}_{h \geq 0}$ which increases $\mathcal{J}(R_{\mathcal{X}})$ such that*

$$\mathcal{J}\left(R_{\mathcal{X}}; \left\{\boldsymbol{\tau}_i^{(h+1)}\right\}, \boldsymbol{\alpha}^{(h+1)}, \boldsymbol{\pi}^{(h+1)}\right) \geq \mathcal{J}\left(R_{\mathcal{X}}; \left\{\boldsymbol{\tau}_i^{(h)}\right\}, \boldsymbol{\alpha}^{(h)}, \boldsymbol{\pi}^{(h)}\right).$$

Proof This is a direct consequence of Propositions 5 and 6, which both guaranty that $\mathcal{J}(R_{\mathcal{X}})$ increases. □

5.3 Choice of the number of classes

In practice the number of classes is unknown and should be estimated. We derive a Bayesian model selection criterion for this purpose which is based in the Integrated Classification Likelihood (ICL) criterion developed by Biernacki et al. (2000). We denote by $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$ the entire set of the mixture parameters which lies in $\Theta = A \times \Pi$, with A the Q-dimensional simplex and $\Pi =]0,1[^{Q(Q+1)/2}$. Then we denote by $g_1(\boldsymbol{\alpha} \mid m_Q)$ and $g_2(\boldsymbol{\pi} \mid m_Q)$ the prior distributions of the parameters for a model m_Q with Q classes. The ICL criterion is an approximation of the complete-data integrated likelihood defined such that:

$$\mathcal{L}(\mathcal{X}, \mathcal{Z} \mid m_Q) = \int_{\Theta} \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\theta}, m_Q) g(\boldsymbol{\theta} \mid m_Q) d\boldsymbol{\theta},$$

where $\mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \boldsymbol{\theta}, m_Q)$ is the complete-data likelihood of model m_Q with Q classes.

Proposition 8 *For a model m_Q with Q classes, the ICL criterion is*:

$$ICL(m_Q) = \max_{\boldsymbol{\theta}} \log \mathcal{L}(\mathcal{X}, \widetilde{\mathcal{Z}} \mid \boldsymbol{\theta}, m_Q) - \frac{1}{2} \times \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} - \frac{Q-1}{2} \log(n).$$

Proof The derivation of ICL is based on the following lemma by Biernacki et al. (2000) which can be applied to our case: if $g(\boldsymbol{\theta} \mid m_Q) = g_1(\boldsymbol{\alpha} \mid m_Q) \times g_2(\boldsymbol{\pi} \mid m_Q)$ then $\log \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid m_Q) = \log \mathcal{L}(\mathcal{Z} \mid m_Q) + \log \mathcal{L}(\mathcal{X} \mid \mathcal{Z}, m_Q)$. The derivation of the first term can be done directly, using a Dirichlet prior, $\mathcal{D}(\delta)$ on proportions, which gives:

$$\begin{aligned}
&\log \mathcal{L}(\mathcal{Z} \mid m_Q) \\
&\quad = \log \int \alpha_1^{n_1} \ldots \alpha_Q^{n_Q} \frac{\Gamma(Q\delta)}{\Gamma(\delta)^Q} \mathbb{I}\left(\sum_q \alpha_q = 1\right) d\alpha, \\
&\quad = \log \Gamma(Q\delta) + \sum_q \log \Gamma(n_q + \delta) - Q \log \Gamma(\delta) \\
&\qquad - \log \Gamma(n + Q\delta),
\end{aligned}$$

where n_q is the number of nodes in class q. Since n_qs are unknown, we replace the missing data $\mathcal{Z}$ by their prediction $\widetilde{\mathcal{Z}}$. Then we consider a non informative Jeffreys prior distribution which corresponds to $\delta = 1/2$. This gives:

$$\log \mathcal{L}(\widetilde{\mathcal{Z}} \mid m_Q) = \log \Gamma(Q/2) + \sum_q \log \Gamma(\tilde{n}_q + 1/2) - Q \log \Gamma(1/2) - \log \Gamma(n + Q/2),$$

with n the total number of nodes. Then we take the limit of this quantity for large n, and using the Stirling formula to approximate the Gamma function we obtain

$$\log \mathcal{L}(\widetilde{\mathcal{Z}} \mid m_Q) = \sum_q \tilde{n}_q \log(\tilde{n}_q) - n\log(n) - \frac{Q-1}{2}\log(n)$$
$$= \max_\alpha \log \mathcal{L}(\widetilde{\mathcal{Z}} \mid \boldsymbol{\alpha}, m_Q) - \frac{Q-1}{2}\log(n).$$

As for the second term, we have $n(n-1)/2$ Bernoulli random variables with fixed labels and $\log \mathcal{L}(\mathcal{X} \mid \mathcal{Z}, m_Q)$ can be calculated using a BIC approximation:

$$\log \mathcal{L}(\mathcal{X} \mid \mathcal{Z}, m_Q) \simeq \max_\pi \log \mathcal{L}(\mathcal{X} \mid \mathcal{Z}, \boldsymbol{\pi}, m_Q) - \frac{1}{2} \times \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2}.$$

Finally, the sum of these two separate terms completes the proof. □

6 Application to biological networks

We apply the methodology developed in this paper to an metabolic network of bacteria *Escherichia coli*: the small molecule interaction metabolism network. In this network, vertices are chemical reactions. Two reactions are connected if a compound produced by the first one is a part of the second one (or vice-versa). The original data are issued from http://biocyc.org/. They have been curated to remove some of the secondary compounds. The network we analyzed is available at http://pbil.univ-lyon1.fr/software/motus/; it is made up of $n = 605$ vertices and the total number of edges is 1782. We emphasize that the algorithm we propose is currently the only inferential method which can handle such a large network.

We first show that the Poisson mixture defined in Sect. 2 better fits the observed degree distribution than the scale free distribution. Then we apply the mixture model for graphs to uncover the structure of this metabolic network.

6.1 Fit of the empirical distribution of the degrees

Many papers claim that the Zipf pdf (defined in (1)) fits well the empirical degree distribution of real networks, but these claims are rarely based on statistical criteria. Moreover, the Zipf distribution is not defined for degree zero, so a threshold (minimal degree) must be defined arbitrarily. In order to assess the quality of fit of the Zipf pdf to the tail of the empirical distribution, we compute the usual chi-square statistics for different thresholds. The minimum chi-square estimate of ρ are computed for each threshold. Table 2 shows that the fit is not good even for the tail distribution with a high value of the threshold. Consequently, the Zipf pdf is only a rough approximation of the true one. It is often better suited for the tail than for the whole distribution.

The fit of the mixture of Poisson distributions is presented in Fig. 1. The BIC criterion selects three classes. Parameter estimates are given in Table 3, and Table 2 shows that the fit of the Poisson mixture is better than the fit of the Zipf distribution. The lack of fit for the two first lines is due to an unexpectedly high number of vertices with two connections: 12 vertices have no connection, 44 have one connection and 150 have two connections. This particular structure is due to a large number of chain reactions which constitute intermediates between two others.

6.2 Mixture modeling of the network

The ICL criterion selects a model with $Q = 21$ classes whose parameter estimates are given in Table 4. Figure 2 presents the graph as a dot-plot where a dot at row i and column j indicates that the edge $i \to j$ is present. To emphasize the connections between the different classes, vertices are reordered within classes. Limits between classes are obtained using a *maximum a posteriori* classification rule: vertex i is classified into the class for which $\widehat{\tau}_{iq}$ is maximal.

Among the first 20 classes, eight are cliques ($\pi_{qq} = 1$) and six have within probability connectivity greater than 0.5. It turns out that all those cliques or pseudo-cliques gather

Table 2 Fit of the power law and Poisson mixture to the degree distribution: Chi-square (χ^2) statistics, degree of freedom and ratio χ^2/df for several thresholds. The same values of the parameters of the Poisson mixture have been used for all thresholds

Threshold	n	$\rho+1$	Power law			Poisson mixture		
			χ^2 stat.	df	χ^2/df	χ^2 stat.	df	χ^2/df
0	593	–	–	–	–	67.25	29	2.32
1	549	1.79	96.22	32	3.01	58.5	28	2.09
2	399	1.93	75.83	31	2.45	32.3	27	1.20
3	315	2.08	59.7	30	1.99	30.6	26	1.18
4	252	2.19	53.07	29	1.83	27	25	1.08
5	200	2.24	52.37	28	1.87	27	24	1.13
6	172	2.37	45.44	27	1.68	25	23	1.09

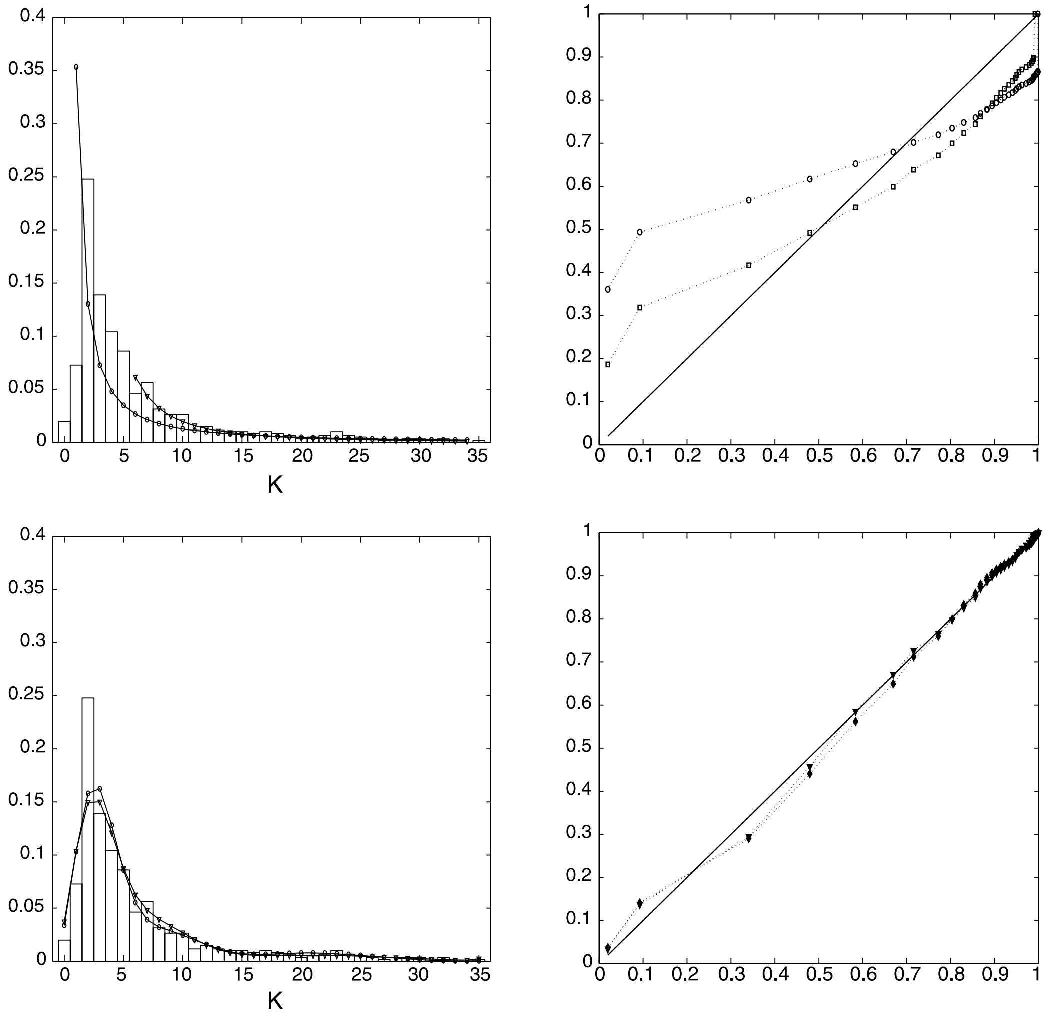


Fig. 1 Fit of the Zipf (*top*) and Poisson mixture (*bottom*) pdf on the *E. Coli* data. *Left*: histogram of degrees with adjusted distributions (Zipf: threshold 1 −◦− and 6 −▽−, Poisson mixture: 3 classes −▽− and 6 classes −◦−). *Right*: PP plots

Table 3 Parameter estimates for the Poisson mixture model on degrees with 3 classes

Class	1	2	3
α (%)	8.9	19.7	71.3
λ	21.5	9.1	3.0

reactions involving a same compound. Examples of compounds responsible for cliques include chorismate, pyruvate, L-aspartate, L-glutamate, D-glyceraldehyde-3-phosphate and ATP. That set of metabolites can be viewed as the backbone of the network.

Since the connection probability between classes 1 and 16 is 1, they constitute a clique which is associated with a single compound: pyruvate. However, that clique is split in two sub-cliques because of their different connectivities with reactions of classes 7 and 10. This distinction is due to the use of CO2 in class 7 and acetylCoA in class 10, which are secondary compounds involved in reactions of class 1 but not in those of class 16.

Remark 3 Table 4 also shows that the clique structure strongly increases the mean degree λ_q of its elements.

Remark 4 In this example, it turns out that the within connection probabilities π_{qq} are always maximal. A simulation study (not shown) prove that it is not an artifact of the method, which can detect classes without intra-connection ($\pi_{qq} = 0$).

To end, we also compare the expected clustering coefficient c given in Proposition 3 with the empirical one. The expected value for $Q = 21$ classes is 0.544, while the observed one is 0.626. The mixture model for graphs therefore slightly underestimates this coefficient. On the same dataset, the Erdös–Rényi model would give $\widehat{c} = \widehat{\pi} = 0.0098$.

Fig. 2 *Top*: Dot-plot representation of the adjacency matrix of the graph after classification of the vertices into the 21 classes

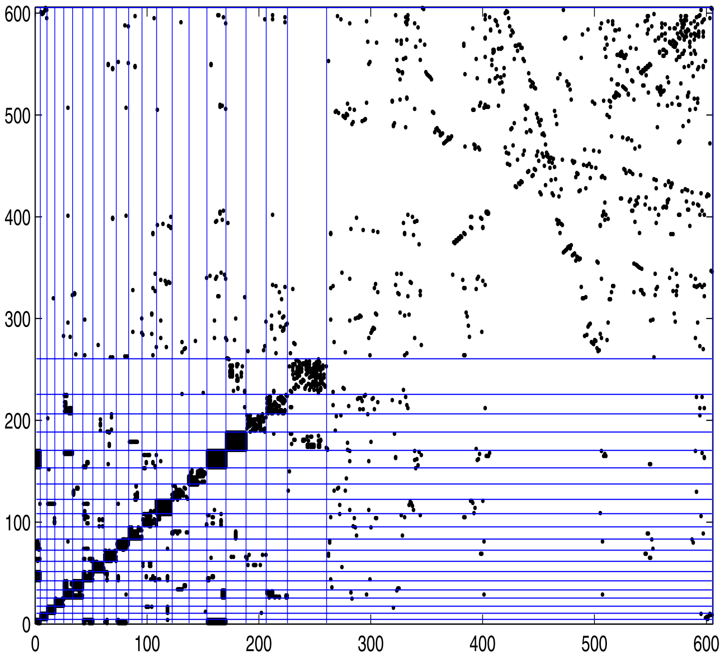


Table 4 Parameter estimates of the mixture model for graphs with $Q = 21$ classes: $\boldsymbol{\alpha}$, $\boldsymbol{\pi}$ and λ_qs. Values smaller than 0.5% are hidden for readability

α (%)	0.7	1.0	1.2	1.3	1.3	1.5	1.5	1.6	1.8	1.8	2.0	2.1	2.3	2.6	2.7	2.8	3.0	3.0	3.3	5.8	56.8
	100						64		11	43			2			100					
		100																			
			100									4	7		1			1			
				71																	
					100	28				1						18			16		
					28	100								6							
	64						58		10	4		7	5			5					
								63				5						3			
	11						10		65							1	2	2			
	43				1		4			67			1								
π											62			7			4				
(%)			4				7	5				28	5			5					
	2		7				5			1		5	100			1					
						6					7			25							
			1												40						
	100				18		5		1			5	1			100					
									2		4						100			6	
			1					3	2									21			
					16														19		
																	6			11	
																					1
λ_q	33	7	9	6	17	13	12	7	10	10	10	8	17	6	7	25	21	5	6	5	3

Acknowledgements The data and the original biological problem have been provided by V. Lacroix and M.-F. Sagot (INRIA-Hélix, INRIA, Lyon). The authors also thank C. Matias, E. Birmelé (CNRS-Statistic and Genome group, Evry univ.) and S. Schbath (INRA-MIG, Jouy-en-Josas) for all their helpful remarks and suggestions.

References

Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (2002)

Alm, E., Arkin, A.P.: Biological networks. Cur. Op. Struct. Biol. **13**, 193–202 (2002)

Arita, M.: The metabolic world of *Escherichia coli* is not small. PNAS **101**, 1543–1547 (2004)

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

Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719–725 (2000)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39**, 1–38 (1977)

Govaert, G., Nadif, M.: An EM algorithm for the block mixture model. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 643–647 (2005)

Jaakkola, T.: Advanced Mean Field Methods: Theory and Practice. MIT Press, Cambridge (2000). Chapter: Tutorial on variational approximation methods

Jones, J., Handcock, M.: Likelihood-based inference for stochastic models of sexual network formation. Theor. Pop. Biol. **65**, 413–422 (2004)

Jordan, M.I., Ghahramani, Z., Jaakkola, T., Saul, L.K.: An introduction to variational methods for graphical models. Mach. Learn. **37**, 183–233 (1999)

Molloy, M., Reed, B.: A critical point for random graphs with a given degree sequence. Rand. Struct. Algorithms **6**, 161–179 (1995)

Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**, 066133 (2004)

Newman, M.E.J., Girvan, M.: Statistical Mechanics of Complex Networks. Springer, Berlin (2003). Chapter: Mixing patterns and community structure in networks

Newman, M.E.J., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. PNAS **99**, 2566–2572 (2002)

Nowicki, K., Snijders, T.: Estimation and prediction for stochastic block-structures. J. Am. Stat. Assoc. **96**, 1077–1087 (2001)

Pattison, P.E., Robins, G.L.: Handbook of Probability Theory with Applications. Sage, Beverley Hills (2007). Chapter: Probabilistic network theory

Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Networks motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. **31**, 64–68 (2002)

Stumpf, M., Wiuf, C., May, R.: Subnets of scale-free networks are not scale-free: sampling properties of networks. Proc. Natl. Acad. Sci. USA **102**, 4221–4224 (2005)

Tanaka, R., Doyle, J.: Some protein interaction data do not exhibit power law statistics. FEBS Lett. **579**, 5140–5144 (2005)

Zhang, V.L., King, O.D., Wong, S.L., Goldberg, D.S., Tong, A.H.Y., Lesage, G., Andrews, B., Bussey, H., Boone, C., Roth, F.P.: Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. J. Biol. **4**, 1–13 (2005)