A mixture model for random graphs

Jean-Jacques Daudin, Franck Picard, Stéphane Robin

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A mixture model for random graphs

Jean-Jacques Daudin, Franck Picard, Stéphane Robin

N° 5840

22nd February 2006

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--5840--FR+ENG

A mixture model for random graphs

Jean-Jacques Daudin, Franck Picard, Stéphane Robin *

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet SELECT

Rapport de recherche n° 5840 — 22nd February 2006 — 19 pages

Abstract: The Erdos-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-word networks. The vertices of these networks are often structured in *prior* unknown clusters (functionally related proteins or social communities) with different connectivity properties. We define a generalization of the Erdos-Rényi model called ERMG for Erdos-Rényi Mixtures for Graphs. This new model is based on mixture distributions. We give some of its properties, an algorithm to estimate its parameters and apply this method to uncover the modular structure of a network of enzymatic reactions.

Key-words: Random Graphs, Mixture Models, Interaction Networks

* **UMR INA P-G/ENGREF/INRA MIA 518**, Paris, France. [daudin][picard][robin]@inapg.inra.fr

Unité de recherche INRIA Futurs
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Un modèle de mélange pour graphes aléatoires

Résumé : Le modèle de graphe d'Erdos-Rényi est simple et des expressions explicites de ses caractéristiques moyennes ou asymptotiques sont disponibles, mais il s'ajuste mal aux réseaux observés dans la réalité. Les sommets de ces réseaux sont fréquemment structurés en groupes *a priori* inconnus (protéines liées fonctionnellement, ou communautés sociales) et ayant des propriétés de connectivité différentes. Nous définissons une généralisation du modèle d'Erdos-Rényi appelée ERMG pour "Erdos-Rényi Mixtures for Graphs". Ce nouveau modèle est fondé sur des lois de mélange. Nous étudions certaines de ses propriétés, proposons un algorithme d'estimation de ses paramètres et appliquons cette méthode pour révéler la structure modulaire d'un réseau de réactions enzymatiques.

Mots-clés : Graphes aléatoires, modèles de mélange, réseaux d'interaction

1 Introduction

The Erdos-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 subgraphs, degree distribution, connectedness and clustering coefficient. However this theoretical model does not fit well to real-word, 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. Some generalizations of the Erdos-Rényi model have been recently made in order to correct these shortcomings. For a review of these works see **?** or **?**. A special attention has been paid recently to the study of biological networks (see **?** or **?**).

One research direction is to incorporate clustering in the model. Assortative mixing or mixing patterns (see **?** and (2004)) postulate that the vertices may be classified into clusters with different connectivity properties. The key element is the mixing matrix which specifies the probability of connection between two clusters. **?** gives some theoretical properties of such networks and an algorithm similar to Metropolis-Hasting for simulating networks for a given mixing matrix. The inference of the mixing parameters is quite easy if clusters can be defined using external information such as language, race or age. However the inference is more difficult when clusters and mixing parameters have to be inferred when the network topology is the only available information. A first step is the greedy optimization algorithm proposed by **?**. In this article we propose a new statistical method to infer the clustering of vertices and the parameters of the mixing model using a maximum-likelihood approach based only on the network topology.

Notations. 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 $X_{ii} = 0$. 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:

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

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:

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

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 (1). The scale-free (or Zipf) distribution has been intensively used as an alternative. The Zipf probability distribution function (pdf) is

$$\Pr\{K_i = k\} = c(\rho)k^{-(\rho+1)}, \tag{2}$$

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 Section 6 that this distribution may have a poor fit on real datasets as well.

First of all, it is important to notice that the Zipf distribution is used to model the tail of the degrees' distribution. Consequently 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 degrees' distribution is a mixture of Poisson distributions.

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

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

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 cluster q. We suppose that the conditional distribution of the degrees is a Poisson distribution :

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

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

$$\Pr\{K_i = k\} = \sum_{q=1}^{Q} \alpha_q \frac{e^{-\lambda_q}\lambda_q^k}{k!}. \tag{3}$$

Remark 2. Because vertices are connected between them, degrees are not independent from each other. However, in the standard situation where n is large and where the λ_qs are small with respect to n, the dependency between the degrees is weak.

In Section 6 we will show that this model fits well to several data sets. Nevertheless, we claim that modelling 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 groups may reveal some interesting underlying structure, such as preferential connections between groups. 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 propose a mixture model which explicitly describes the way edges connect vertices, accounting for some heterogeneity among vertices. In the following, we denote this model ERMG for Erdös-Rényi Mixture for Graphs.

The ERMG model supposes that vertices are spread into Q clusters with *prior* probabilities $\{\alpha_1, \dots \alpha_Q\}$. In the following, we use the same indicator variables $\{Z_{iq}\}$ defined in section 2. Then we denote $\pi_{q\ell}$ the probability for a vertex from group q to be connected with a vertex from group ℓ. Because the graph is undirected, these probabilities must be symetric such that:

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

Remark 3. In the following, we will use two equivalent notations: $\{Z_{iq} = 1\}$ or $\{i \in q\}$ to indicate that vertex i belongs to cluster q.

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

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

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

$$X_{ij} \mid \{i \in q, j \in \ell\} \sim \mathcal{B}(\pi_{q\ell}).$$

The main difference with Model (3) is that the ERMG model directly deals with edges. More than describing the clustered structure of vertices, our model describes the topology of the network using the connectivity matrix $\mathbf{\Pi} = (\pi_{q\ell})$.

3.2 Examples

In this section we aim at showing that the ERMG model can be used to generalize many particular structures of random graphs. Figure 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-word 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 **?**). A permutation algorithm is also proposed in **?**. 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 ERMG model presented in Section 4.4.
The scale-free network proposed by **?** is a particular case of random graphs with arbitrary distribution. To this extent, we can propose an analogous model in the ERMG framework. Suppose that the incoming vertices join the network in groups of respective size $n\alpha_q$ ($q = 1..Q$, $n\alpha_1$ being the number of original vertices). Assuming that the elements of a new group connect preferentially the elements of the oldest groups:

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

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

Example 2. Affiliation network.
An affiliation network or bipartite graph, 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 group are connected. This model has been studied by **?**. This type of network may be modelled by an ERMG with ones in the diagonal of $\mathbf{\Pi}$.

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 **?** for instance. This type of pattern may be modelled by an ERMG with extradiagonal ones in $\mathbf{\Pi}$.

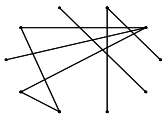
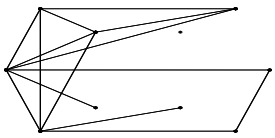
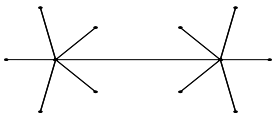
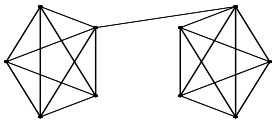
4 Some properties of the ERMG model

4.1 Distribution of the degrees

Proposition 4.1 *Given the label of a vertex, the conditional distribution of the degree of this vertex is Binomial (approximately Poisson):*

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

Table 1: Some typical network configurations and their formulation in the framework of the ERMG model

Description	Network	Q	$\mathbf{\Pi}$	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	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$	$\frac{1+3\varepsilon^2}{(1+\varepsilon)^2}$

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

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

$$\Pr\{i \leftrightarrow j \mid i \in q\} = \sum_\ell \Pr\{i \leftrightarrow j \mid i \in q, j \in \ell\} \Pr\{j \in \ell\} = \sum \alpha_\ell \pi_{q\ell} = \overline{\pi}_q.$$

The result follows. ■

4.2 Between-group connectivity

Definition 4.2 *The connectivity between group* q *and* ℓ *is the number of edges connecting a vertex from group* q *to a vertex from group* ℓ.

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

A_{qq} *is actually the within-connectivity of group* q.

Proposition 4.3 *The expected connectivity between group* q *and* ℓ *is:*

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

Proof According to Definition 4.2, $A_{q\ell}$ 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 modelling is usually available, this coefficient is empirically defined in most cases. **?** 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_is:

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

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 $\{1,\ldots,n\}$, the definition of C can be rephrased as $c = \Pr\{\nabla \mid \mathsf{V}\}$. Because ∇ is a particular case of V, we have:

$$c = \Pr\{\nabla \cap \mathsf{V}\} / \Pr\{\mathsf{V}\} = \Pr\{\nabla\} / \Pr\{\mathsf{V}\}. \tag{4}$$

This property suggests another estimate of c proposed by **?**:

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

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 4.4 *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 $\{1,\ldots,n\}$*

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

Proposition 4.5 *In the ERMG model, 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

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

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 (4). ■

4.4 Independent model

The model presented in Section 2 can be rephrased as an independent version of the ERMG model. Indeed the absence of preferential connection between groups corresponds to the case where

$$\pi_{q\ell} = \eta_q \eta_\ell. \tag{5}$$

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:

$$\lambda_q = (n-1)\eta_q\overline{\eta}, \tag{6}$$

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

Beween group 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 Section 6 that the observed number of connections between groups may be quite far from expected values.

Clustering coefficient

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

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 Figure 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 ERGM model, we use the complete-data framework defined by **?**. Let us denote $\mathcal{X}$ the set of all edges: $\mathcal{X} = \{X_{ij}\}_{i,j=1..n}$, and $\mathcal{Z}$ the set of all indicator variables for vertices: $\mathcal{Z} = \{Z_{iq}\}_{i=1,n}^{q=1,Q}$.

Proposition 4.6 *The complete-data log-likelihood is*

$$\log \mathcal{L}(\mathcal{X}, \mathcal{Z}) = \sum_i \sum_q Z_{iq} \log \alpha_q + \sum_i \sum_q \sum_{j>i} \sum_\ell Z_{iq} Z_{j\ell} \log b(X_{ij}; \pi_{q\ell}).$$

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

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

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

The log-likelihood of the observed data is obtained by summing the complete-data log-likelihood over all the possible values of the unobserved variables $\mathcal{Z}$. Unfortunately, it seems that no simple form of this function can be derived. Then we define the conditional expectation of the complete-data log-likelihood such that:

$$\begin{aligned} \mathcal{Q}(\mathcal{X}) &= \mathbb{E}\{\log \mathcal{L}(\mathcal{X},\mathcal{Z})|\mathcal{X}\} \\ &= \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_i \sum_q \sum_{j>i} \sum_\ell \theta_{ijq\ell} \log b(X_{ij};\pi_{q\ell}), \end{aligned} \tag{7}$$

where

$$\begin{array}{rcccl} \tau_{iq} & = & \Pr\{Z_{iq}=1 \mid \mathcal{X}\} & = & \mathbb{E}(Z_{iq} \mid \mathcal{X}), \\ \theta_{ijq\ell} & = & \Pr\{Z_{iq}Z_{j\ell}=1 \mid \mathcal{X}\} & = & \mathbb{E}(Z_{iq}Z_{j\ell} \mid \mathcal{X}). \end{array} \tag{8}$$

This log-likelihood involves the joint *posterior* probability for vertices i and j to belong to groups q and ℓ. Clearly, we have for i and j:

$$\sum_q \tau_{iq} = 1, \qquad \theta_{ijq\ell} = \theta_{ji\ell q}, \qquad \sum_q \sum_\ell \theta_{ijq\ell} = 1. \tag{9}$$

5 Estimation

In this section we propose an (approximate) E-M algorithm to estimate the parameters of the ERMG model by maximum likelihood. Since the EM algorithm uses the hidden structure of the data, it is crucial to determine the dependency among observed and hidden variables.

Since the data under study are represented as a graph, the ERMG model may look like a hidden Markov Field model. However, it is important to note that it is not. The main reason for this is that when using a hidden Markov model the topology of the graph needs to be known, whereas it is precisely the random object under study in the ERMG framework.

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 ERMG framework.

Assume that the vertices are divided in two groups, 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 group because no connection is possible

between vertices pertaining to two different groups. The same is true for vertices i and k. Therefore the three vertices are in the same group 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 ERMG framework since unconnected vertices provide as much information as connected vertices. This is why the likelihood can not be simplified for computation.

5.2 Approximate E step

The most difficult part of the estimation algorithm is the calculation of the $\tau_{i\ell}$s and $\theta_{ijq\ell}$s. Because of the strong dependency between edges, these *posterior* probabilities seem very difficult to derive. We propose a two step approximation.

Approximate joint distribution. In the first step, we approximate the joint distribution of the Z_{iq}s by the product of their respective conditional distributions given the other coordinates. Denoting $\mathcal{Z}_i = \{Z_{i1}, \dots Z_{iQ}\}$ and $\mathcal{Z}^i = \mathcal{Z} \setminus \mathcal{Z}_i$, we set

$$\Pr\{\mathcal{Z} \mid \mathcal{X}\} \simeq \prod_i \Pr\{\mathcal{Z}_i \mid \mathcal{X}, \mathcal{Z}^i\}. \tag{10}$$

These approximate distribution can be caculated thanks to the folowing proposition.

Proposition 5.1 *Denoting* $N_m^i = \sum_{j \neq i} Z_{jm}$ *and* $C_{im} = \sum_k Z_{km} X_{ik}$, *we have*

$$\Pr\{Z_{iq} = 1 \mid \mathcal{X}, \mathcal{Z}^i\} \propto \alpha_q \prod_m b(C_{im}; N_m^i, \pi_{qm}).$$

Predicting label variables. Approximation (10) can not be used as such since $\mathcal{Z}^i$ is unknown and has to be predicted. The second step of the approximation is hence to fix all $Z_{j\ell}$s ($j \neq i$) to their conditional expectations: $\widehat{Z}_{j\ell} = \tau_{j\ell}$. The posterior probabilities τ_{iq} must therefore satisfy the following fix point relation:

$$\widehat{\tau}_{iq} = \Pr\{Z_{iq} = 1 \mid \mathcal{X}, \widehat{\mathcal{Z}}^i\}.$$

The $\widehat{\tau}_{iq}$ are obtained by iterating the equation given in Proposition 5.1 until convergence. According to approximation (10), we then get $\widehat{\theta}_{ijq\ell} = \widehat{\tau}_{iq}\widehat{\tau}_{j\ell}$.

5.3 M step

At this step, we maximize the function $\mathcal{Q}(\mathcal{X})$ given in (7) subject to $\sum_q \alpha_q = 1$. We get

$$\widehat{\alpha}_q = \sum_i \widehat{\tau}_{iq}/n, \qquad \widehat{\pi}_{q\ell} = \sum_i \sum_j \widehat{\theta}_{ijq\ell} X_{ij} \Big/ \sum_i \sum_j \widehat{\theta}_{ijq\ell}\ .$$

5.4 Choice of the number of groups

Our purpose here is not to derive a specific criterion to select the number of groups in the ERMG model. This problem seems difficult to tackle, especially because the log-likelihood of the observed data $\log \mathcal{L}(\mathcal{X})$ is not calculable.

We propose a heuristic criterion inspired from the Integrated Completed Likelihood (ICL, **?**). The ICL criterion uses the same penalty as BIC, but applies it the complete-data log-likelihood, which is the only likelihood we can calculate in this case. The first term of (7) deals with Q proportions α_qs and involves n data. The second term deals with $Q(Q+1)/2$ probabilities $\pi_{q\ell}$s and involves $n(n-1)/2$ terms. Hence the Fisher information matrix derived from $\mathcal{Q}(\mathcal{X})$ is proportional to n for the α_qs, while it is proportional to $n(n-1)/2$ for the $\pi_{q\ell}$s.

We therefore propose the following heuristic criterion:

$$-2\mathcal{Q}(\mathcal{X}) + (Q-1)\log n + [Q(Q+1)/2]\log[n(n-1)/2]. \tag{11}$$

6 Application to biological networks

We apply the methodology developed in this paper to an interaction 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). This network is available at $\theta^\star biocyc.org$; it is made up of $n = 605$ vertices and the total number of edges is 1782. We first analyse the distribution of the degrees of edges and then apply the ERMG model.

6.1 Fit of the empirical distribution of the degrees

Zipf distribution. Many papers claim that the Zipf pdf (2) fits well the degrees of graphs, but these claims are rarely based on statistical criteria. Generally only a log-log plot is given. If we consider the log-log plot on our data (Fig. 2 (a)) we can see that a linear fit does not work for low degrees (e.g. < 4). In order to see how the Zipf pdf fits 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 (see Table 2).

We can see that the fit is not good even for the tail distribution with a high value of the threshold. One can say that the Zipf distribution is only a rough approximation of the true one. It is often better suited for the tail than for the whole distribution. Note that the fit seems better for the tail because we have less data when the threshold increases, so that the power of the chi-square test is downsized. We would like to have a model which is well suited for the whole distribution of degrees.

Poisson mixture. Using a mixture of Poisson distributions we obtain the following results. The BIC criterion selects three groups. Parameter estimates are given in Table 3 and chi-square statistics in Table 2.

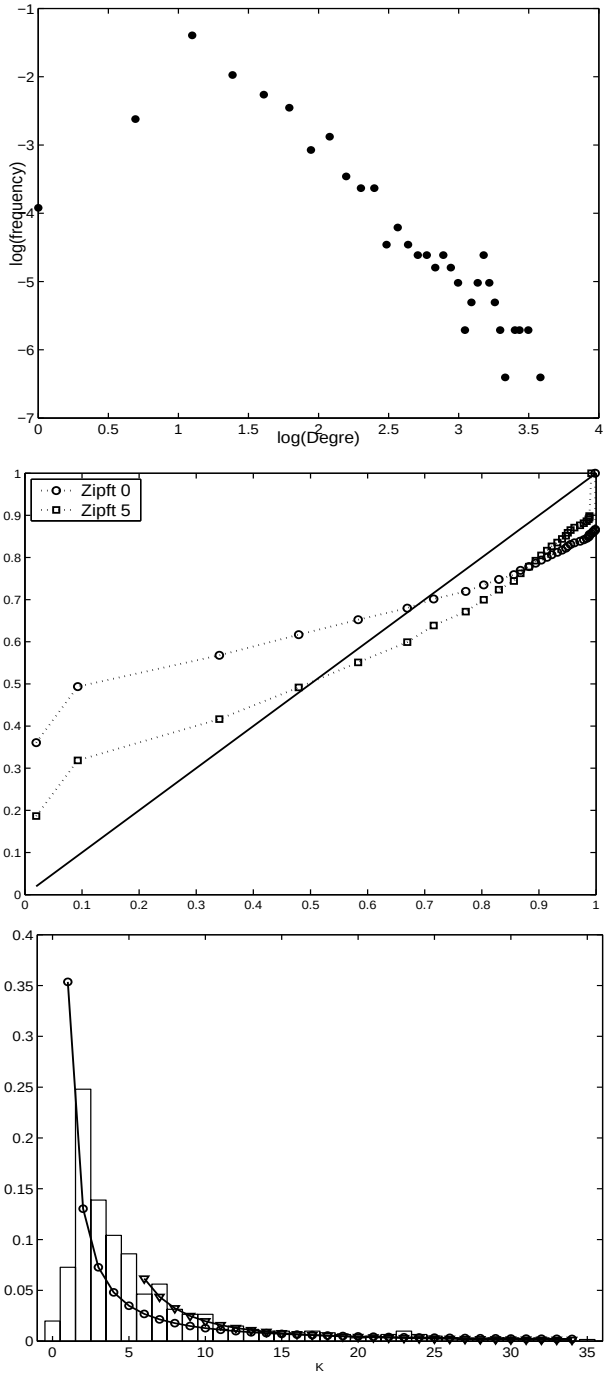


Figure 1: Fit of the Zipf pdf on the *E. Coli* data. Top: log-log plot. Center: PP plots with threshold values equal to 1 and 6. Bottom: histogram of degrees with adjusted power distributions (threshold $1 -\circ-$ and $6 -\triangledown-$).

Table 2: Fit of the power law and Poisson mixture: Chi-square statistic, degree of freedom and p-value for several thresholds.

			Power law			Poisson mixture		
Threshold	n	$\rho+1$	χ^2 stat.	df	p-value	χ^2 stat.	df	p-value
0	593	-	-	-	-	67.25	29	$7\ 10^{-5}$
1	549	1.79	96.22	32	$2\ 10^{-9}$	58.5	28	$6\ 10^{-4}$
2	399	1.93	75.83	31	$1\ 10^{-6}$	32.3	27	0.22
3	315	2.08	59.70	30	0.001	30.6	26	0.24
4	252	2.19	53.07	29	0.004	27.0	25	0.36
5	200	2.24	52.37	28	0.003	27.0	24	0.30
6	172	2.37	45.44	27	0.014	25.0	23	0.35

Table 3: Parameter estimates for the Poisson mixture model on degrees with 3 groups.

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

Note that the same values of the parameters of the mixture distribution have been used for all threshold values. One can see that the fit is better than the fit of the power law. 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.

6.2 Erdös-Rényi mixture modeling

Number of groups and parameter estimates. Using the heuristic criterion defined in (11), we select $Q = 21$ groups.

Table 4 gives the estimates of proportions α_q and connection probabilities $\pi_{q\ell}$. Among the first 20 groups, 8 are actually cliques ($\pi_{qq} = 1$) and 6 have within probability connectivity greater than 0.5. It turns out that all this cliques or pseudo-cliques gather reactions involving a same compound (as an substrate or a product). We also see that the clique structure strongly increases the mean degree λ_q of its elements.

The connection probability between groups 1 and 16 is 1, so these 2 groups actually constitute a clique together which again correspond to a single compound. However, they are separated in two sub-cliques because of their very different connectivities with reactions of groups 10 and 11. This distinction is due to the use of two secondary compounds involved in reactions of group 1 (and 10 and 11) but not of group 18.

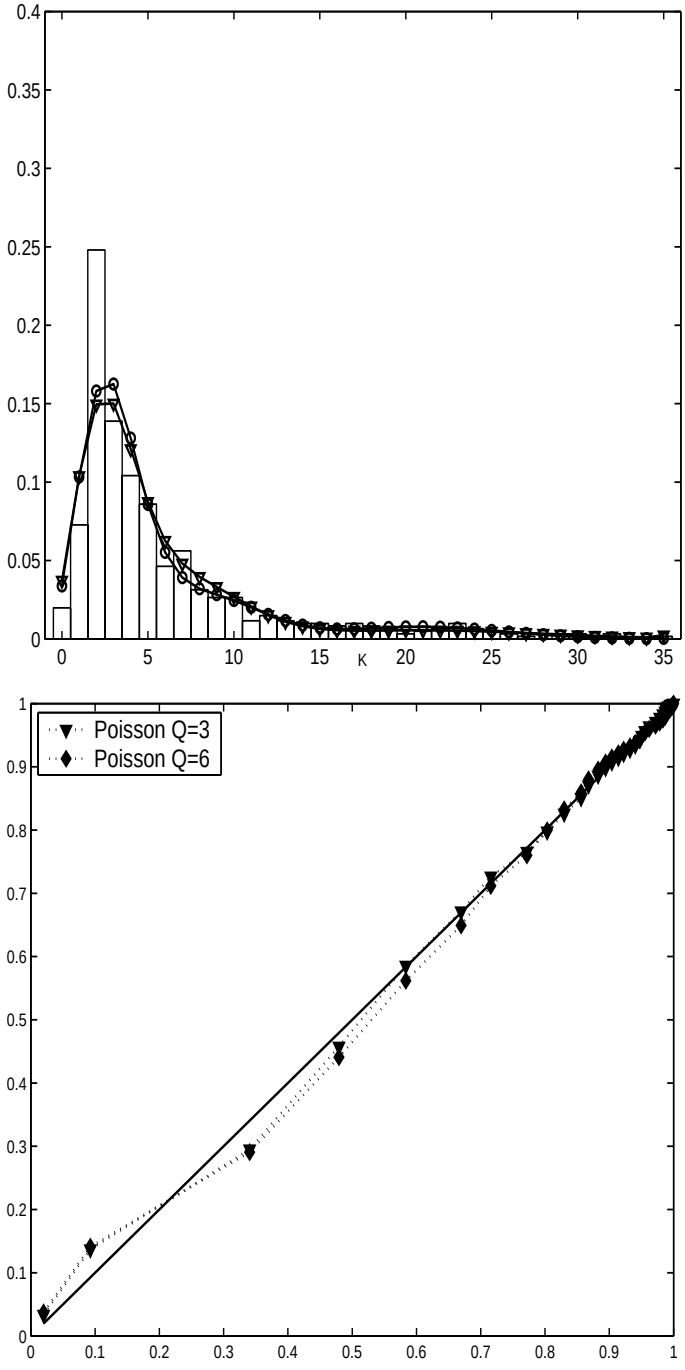


Figure 2: Fit of the Poisson mixture pdf on the *E. Coli* data. Top: empirical and fitted distributions with $Q = 3$ groups $(-\nabla-)$ and $Q = 21$ groups $(-\circ-)$. Bottom: PP plots with no threshold.

Table 4: Parameter estimates of the ERMG model with $Q = 21$ classes (values smaller than .5 % are masked 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

In this example, it turns out that the within connection probabilities π_{qq} are always maximal, although the modeling does not require this. Simulation studies (not shown) prove that it is not an artifact of the method, which can detect a group with no within connection.

Between group connectivity and clustering coefficient. The graph showing 1782 edges connecting 605 vertices is of course unreadable. Figure 3 presents the graph as a dot-plot where a dot at row i and column j indicates that the edge $i \leftrightarrow j$ is present. To emphasize the connections between the different groups, we reordered the vertices within groups. The limits between groups are obtained using a maximum a posteriori (MAP) classification of vertices: the vertex i is classified into group q for which $\widehat{\tau}_{iq}$ is maximal.

The bottom plot in Figure 3 gives the estimated *posterior* probabilities $\widehat{\tau}_{iq}$. We see that the first groups are quite well defined. The last one (21) has more fuzzy limits: is actually made of very isolated reactions having not much in common.

Figure 4 compares the expected mean degrees λ_q and connectivities $A_{q\ell}$ between the groups with the 'observed' ones. Because the true group to which each vertex belongs is not known, the observed values are not available. We propose to classify vertices using the MAP rule to get an estimate of

- the mean connectivity: $\overline{K}_q = \sum_i \widehat{\tau}_{iq} K_i \,/ \sum_i \widehat{\tau}_{iq}$
- and of the between group connectivity: $\widehat{A}_{q\ell} = \sum_i \sum_{j>i} \widehat{\tau}_{iq} \widehat{\tau}_{j\ell} X_{ij}$.

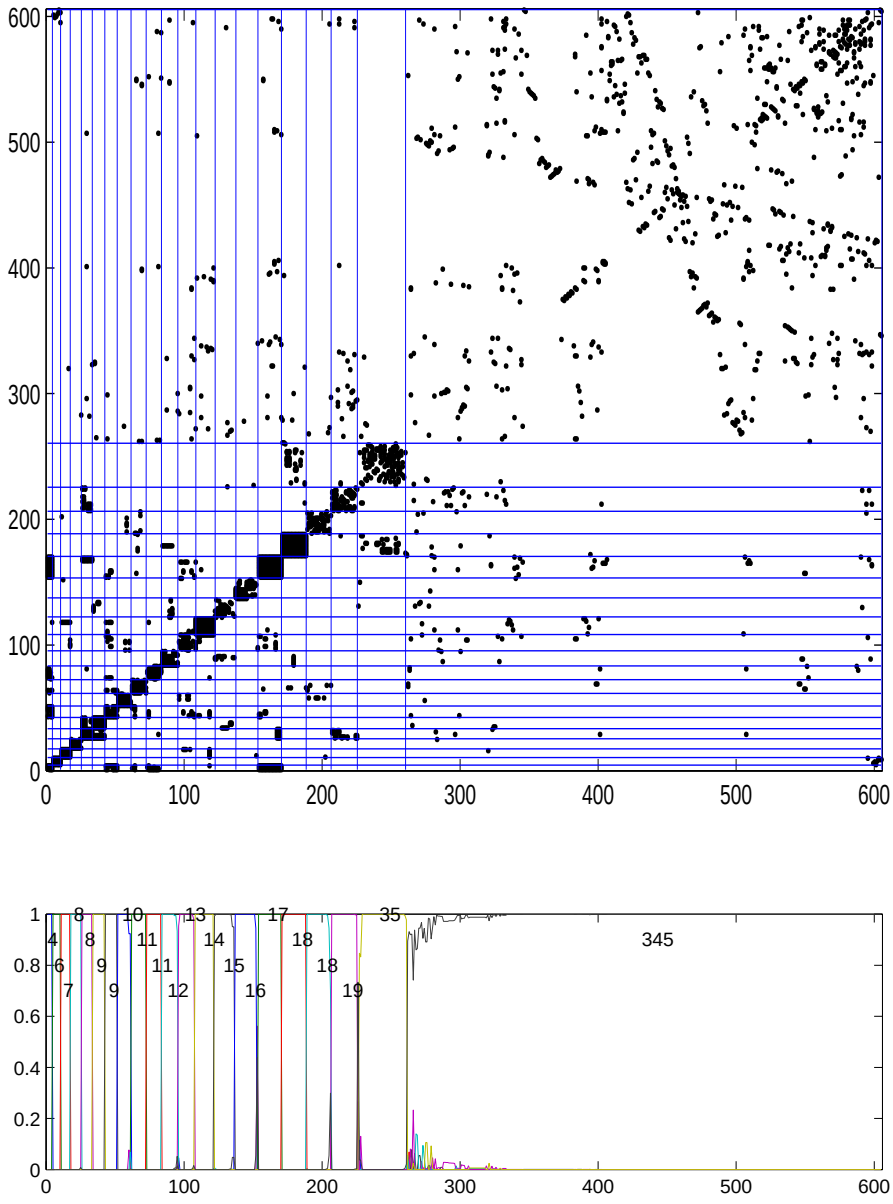


Figure 3: Top: Dot plot representation of the graph after classification of the vertices into the 6 groups. Bottom: Posterior probabilities τ_{iq}.

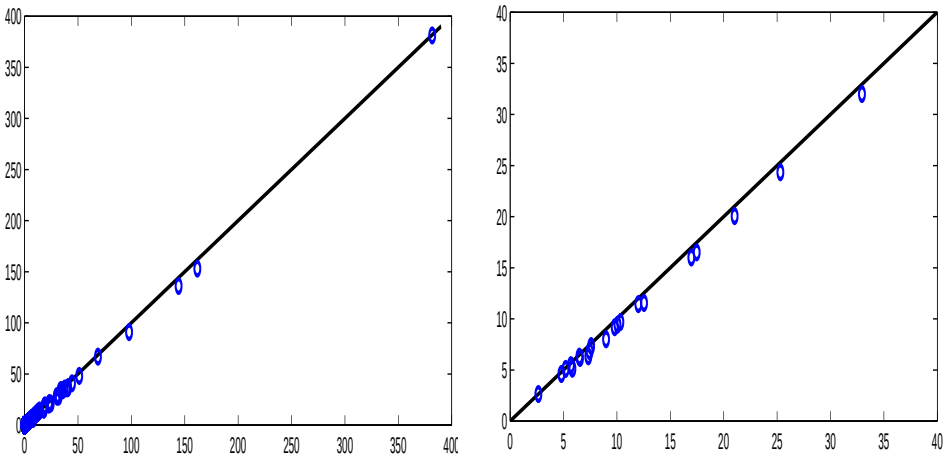


Figure 4: Connectivities between groups ($A_{q\ell}$: left) and mean degree (λ_q: right). x-axis = expected, y-axis = 'observed'

Indeed this comparison is not completely satisfying since both the expected and the 'observed' values are derived from the modeling. Therefore, the very good fit obtained in Figure 4 is partly artificial.

To end, we also compare the expected clustering coefficient c given in Proposition 4.5 with the observed one. The expected value for $Q = 21$ groups is 0.544, while the observed one is 0.626. The ERMG model therefore slightly underestimates this coefficient. On the same dataset, the Erdös-Rényi model would give $\widehat{c} = \widehat{\pi} = 0.0098$.

Acknowledgments. 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.

Unité de recherche INRIA Futurs
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399