

# Uncovering structure in biological networks

Jean-Jacques Daudin<sup>1</sup>, Vincent Lacroix<sup>2</sup>, Franck Picard<sup>1</sup>, Stéphane Robin<sup>1</sup>, Marie-France Sagot<sup>2</sup>

<sup>1</sup> UMR INA P-G/ENGREF/INRA MIA 518, Groupe SSB

16, rue Claude Bernard, 75005, Paris FRANCE

[daudin][picard][robin]@inapg.inra.fr

<sup>2</sup> UMR 5558 Biométrie et Biologie Évolutive

Université Claude Bernard, Lyon I, 43, Bd du 11 novembre 1918,

69622 Villeurbanne cedex, FRANCE

[sagot][lacroix]@biomserv.univ-lyon1.fr

**Abstract:** *The Erdős-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 these networks are often structured in prior unknown groups (functionally related proteins or social communities) with different connectivity properties. We define a generalisation of the Erdős-Rényi model called ERMG for Erdős-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.*

**Keywords:** Random graphs, Mixture models, Interaction network.

## 1 Introduction

The Erdős-Rényi model of a network is one of the oldest and best studied models 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-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. Some generalisations of the Erdős-Rényi model have been recently made in order to correct these shortcomings. For a review of these works see Albert and Barabási (2002) or Newman (2003). Besides, a special attention has been paid recently to the study of biological networks (see Alm and Arkin (2002) or Arita (2004)). One of the limits for these studies is that no existing network model seems to be completely satisfying to capture their structure.

One research direction is to incorporate clustering in the model. Assortative mixing or mixing patterns (see Newman and Girvan (2003) and (2004)) postulate that the vertices may be classified into groups with different connectivity properties. The key element is the mixing matrix which specifies the probability of connection between two groups. Newman (2003) 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 groups can be defined using external information such as language, race or age. However the inference is more difficult when groups and mixing parameters have to be inferred when the network topology is the only available information. A first step is the greedy optimisation algorithm proposed by Newman (2004). 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. We then apply this method to a network representing the small molecule metabolism of *Escherichia coli*.

## 2 Mixture model for the degrees

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 to the graph:

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

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:

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

'SCALE-FREE' NETWORK. In many practical situations, the Erdős-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

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

where  $k$  is any positive integer,  $\rho$  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 degree distribution. Consequently it is often better 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 Erdős-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.

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

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

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

We suppose that the conditional distribution of the degree 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:

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

*Remark:* 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  $\lambda_q$ s 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 Erdős-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 Erdős-Rényi Mixture for Graphs.

The ERMG model supposes that vertices are spread into  $Q$  groups 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  $\ell$ . Because the graph is undirected, these probabilities must be symmetric such that:

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

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

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

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

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

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  $\mathbf{II} = (\pi_{q\ell})$ .

## 3.2 Examples

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

**RANDOM GRAPHS WITH ARBITRARY DEGREE DISTRIBUTIONS.** The Erdős-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: 1.  $n$  partial edges (with only one starting vertex and no final vertex) are randomly chosen from the empirical degree distribution and 2. 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 ERMG model presented in Section 4.4.

**SCALE FREE NETWORK.** 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 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 to the elements of the oldest groups:  $\pi_{q,1} \geq \pi_{q,2} \geq \dots \geq \pi_{q,q-1}$ , we get the same kind of structure as the scale-free model.

**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 Newman *et al.* (2002). This type of network may be modelled by an ERMG with ones in the diagonal of  $\mathbf{II}$ .

**STAR PATTERN.** Many biological networks contain star patterns, *i.e.* many vertices connected to the same vertex and only to it, see the interaction network of *S. cerevisiae* in Zhang *et al.* (2005) for instance. This type of pattern may be modelled by an ERMG with extra-diagonal ones in  $\mathbf{II}$ .

## 4 Some properties of the ERMG model

### 4.1 Distribution of the degrees

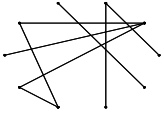
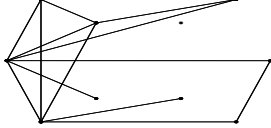
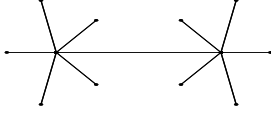
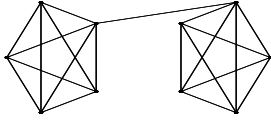
**Proposition 1.** *Given the label of a vertex, the conditional distribution of the degree of this vertex is Binomial (approximately Poisson):  $K_i \mid \{i \in q\} \sim \mathcal{B}(n-1, \bar{\pi}_q) \approx \mathcal{P}(\lambda_q)$ , where  $\bar{\pi}_q = \sum_{\ell} \alpha_{\ell} \pi_{q\ell}$  and  $\lambda_q = (n-1)\bar{\pi}_q$ .*

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

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

The result follows. ■

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

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

## 4.2 Between-group connectivity

**Definition 1.** The connectivity between group  $q$  and  $\ell$  is the number of edges connecting a vertex from group  $q$  to a vertex from group  $\ell$ .  $A_{q\ell} = \sum_i \sum_{j>i} Z_{iq} Z_{j\ell} X_{ij}$ .  $A_{qq}$  is actually the within-connectivity of group  $q$ .

**Proposition 2.** The expected connectivity between group  $q$  and  $\ell$  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 modelling 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  $i$ :  $C_i = \nabla_i / [K_i(K_i - 1)/2]$ , where  $\nabla_i$  is the number of edges between the neighbours 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:  $\hat{c} = \sum_i C_i/n$ .

Denoting  $\nabla$  the 'triangle' configuration ( $i \leftrightarrow j \leftrightarrow k \leftrightarrow i$ ) and  $\mathbb{V}$  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  $c = \Pr\{\nabla \mid \mathbb{V}\}$ . Because  $\nabla$  is a particular case of  $\mathbb{V}$ , we have:

$$c = \Pr\{\nabla \cap \mathbb{V}\} / \Pr\{\mathbb{V}\} = \Pr\{\nabla\} / \Pr\{\mathbb{V}\}. \quad (4)$$

This property suggests another estimate of  $c$  proposed by Newman *et al.* (2002):  $\hat{c} = 3 \sum_i \nabla_i / \sum_i V_i$ , 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 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

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

The same reasoning can be applied to  $\Pr\{V\}$  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

$$\pi_{q\ell} = \eta_q \eta_\ell. \quad (5)$$

The properties of the independent model are as follows.

**DISTRIBUTION OF DEGREES.** The conditional distribution of the degrees is Poisson with parameter  $\lambda_q$  such that:

$$\lambda_q = (n - 1) \eta_q \bar{\eta}, \quad (6)$$

where  $\bar{\eta} = \sum_\ell \alpha_\ell \eta_\ell$ , so  $\lambda_q$  is directly proportional to  $\eta_q$ .

**BETWEEN 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**

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

For the standard Erdős-Rényi model ( $Q = 1$ ,  $\alpha_1 = 1$ ,  $\bar{\eta} = \eta_1 = \sqrt{p}$ ), 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 Erdős-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 ERMG model, we use the complete-data framework defined by Dempster *et al.* (1977). 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.** *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.* This is a direct consequence of the decomposition  $\log \mathcal{L}(\mathcal{X}, \mathcal{Z}) = \log \mathcal{L}(\mathcal{Z}) + \log \mathcal{L}(\mathcal{X} | \mathcal{Z})$  where  $b(x; \pi) = \pi^x (1 - \pi)^{1-x}$ . ■

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:

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

where

$$\begin{aligned} \tau_{iq} &= \Pr\{Z_{iq} = 1 | \mathcal{X}\} = \mathbb{E}(Z_{iq} | \mathcal{X}), \\ \theta_{ijq\ell} &= \Pr\{Z_{iq} Z_{j\ell} = 1 | \mathcal{X}\} = \mathbb{E}(Z_{iq} Z_{j\ell} | \mathcal{X}). \end{aligned} \quad (8)$$

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

$$\sum_q \tau_{iq} = 1, \quad \theta_{ijq\ell} = \theta_{jilq}, \quad \sum_q \sum_\ell \theta_{ijq\ell} = 1. \quad (9)$$

## 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 purposes, it is important to know if  $\Pr\{Z_{iq} = 1 | \mathcal{X}\}$  is equal to  $\Pr\{Z_{iq} = 1 | \mathcal{X}_i\}$ , where  $\mathcal{X}_i$  is the set of all possible edges connecting  $i$ .  $\mathcal{X}_i$  is often called the set of neighbours of vertex  $i$ . In the following, we give a counter example to show that the notion of neighbourhood 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 neighbourhood 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

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

These approximate distributions can be calculated thanks to the following proposition.

**Proposition 5.** 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  $\tau_{iq}$  must therefore satisfy the 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 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 maximise the function  $\mathcal{Q}(\mathcal{X})$  given in (7) subject to  $\sum_q \alpha_q = 1$ . We get

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



## 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, Biernacki *et al.* (2000)). The ICL criterion uses the same penalty as BIC, but applies it to 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  $\alpha_{qs}$  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  $\alpha_{qs}$ , while it is proportional to  $n(n-1)/2$  for the  $\pi_{q\ell s}$ .

We therefore propose the following heuristic criterion:

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

## 6 Application to biological networks

The motivation for applying this methodology to biological networks is twofold: (1) obtain a more realistic random graph model for further work on the over-representation of network motifs (Shen-Orr *et al.* (2002)) and reaction motifs (Lacroix *et al.* (2005)); (2) study the properties of such graphs *per se* to get insight on the modular structure of biological networks.

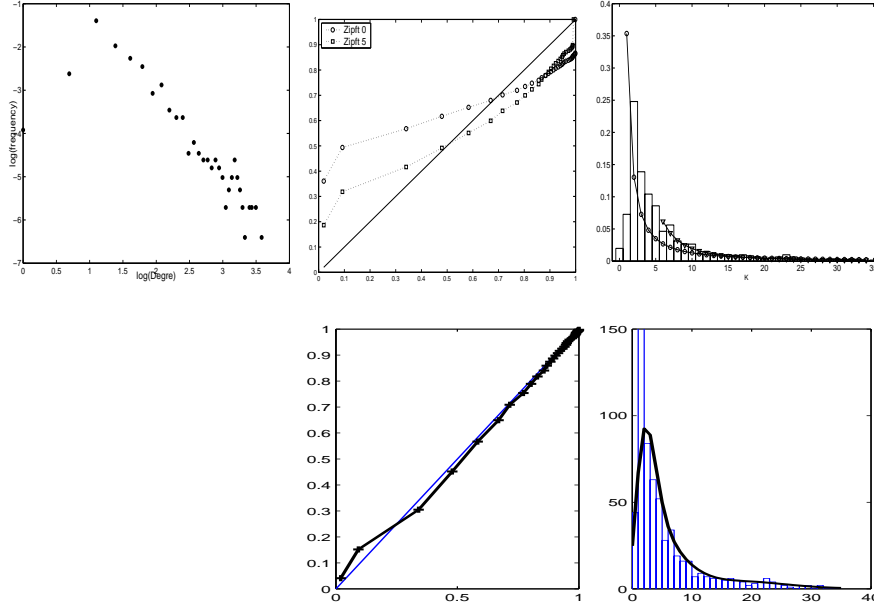
In this section, we will show that the ERMG model is more realistic than other models for describing the degree distribution and the clustering coefficient of a metabolic network. We also show that the groups identified by the method can be given a biological meaning. We apply the methodology developed in this paper to the metabolic network of the bacterium *Escherichia coli*. Although the method is generic and could be applied to other types of biological networks (such as protein interaction networks or transcriptional networks), we chose to first focus on metabolic networks because the data is more complete and reliable. In this network, vertices are chemical reactions. Two reactions are connected if they share a primary compound. For each reaction, a distinction is made between its primary compounds (main substrate and product) and its secondary compounds (cofactors). Only primary compounds are responsible for edges. Importantly, the same compound may be considered as primary with respect to one reaction and secondary with respect to another reaction. This method is an alternative way to deal with the known bias introduced by ubiquitous compounds (such as water) which artefactually connect a large number of reactions (Arita (2004)). Finally, since the information on the reversibility of reactions does not seem to be established (contradictions may be found within a same database), we chose to consider the general case where all reactions are reversible. The data we used was downloaded from <http://biocyc.org/>. The resulting graph is made up of  $n = 605$  vertices and the total number of edges is 1782.

### 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. 1), 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  $\rho$  is 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 down-sized. We would like to have a model which is well suited for the whole distribution of degrees.



**Figure 1.** Fit of the Zipf (top) and Poisson mixture with  $Q = 21$  groups (bottom) pdf to *E. coli* data. Left: log-log plot. Center: PP plots (top: threshold 1 –  $\circ$  – and 6 –  $\nabla$  –). Right: histogram of degrees with adjusted distributions (top: same thresholds).

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

| Threshold | $n$ | Power law  |                |    |                   | Poisson mixture |    |                   |  |
|-----------|-----|------------|----------------|----|-------------------|-----------------|----|-------------------|--|
|           |     | $\rho + 1$ | $\chi^2$ stat. | df | $p$ -value        | $\chi^2$ stat.  | df | $p$ -value        |  |
| 0         | 593 | -          | -              | -  | -                 | 67.25           | 29 | $7 \cdot 10^{-5}$ |  |
| 1         | 549 | 1.79       | 96.22          | 32 | $2 \cdot 10^{-9}$ | 58.5            | 28 | $6 \cdot 10^{-4}$ |  |
| 2         | 399 | 1.93       | 75.83          | 31 | $1 \cdot 10^{-6}$ | 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              |  |

POISSON MIXTURE. Using a mixture of Poisson distributions, we obtain the fit presented in the bottom of Fig. 1. The BIC criterion selects three groups with respective proportions  $\alpha_q = 8.9\%$ ,  $19.7\%$  and  $71.3\%$  and mean degrees  $\lambda_q = 21.5, 9.1, 3.0$ . Chi-square fit statistics are given in Table 2. Observe 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 Erdős-Rényi mixture modelling

NUMBER OF GROUPS AND PARAMETER ESTIMATES. Using the heuristic criterion defined in (11), we select  $Q = 21$  groups. Table 3 gives the estimates of proportions  $\alpha_q$  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. We also see that the clique structure strongly increases the mean degree  $\lambda_q$  of its elements. More generally, in this example, it turns out that the within connection probabilities  $\pi_{qq}$  are always maximal, although the modelling does not require this. Simulation studies (not shown) prove that it is not an artefact of the method, which can detect a group with no within connection.

The interpretation for the cliques (and pseudo-cliques) is straightforward, each of them corresponds to a single compound involved in all the reactions of the group. Examples of compounds responsible for cliques include chorismate, pyruvate, L-aspartate, L-glutamate, D-glyceraldehyde-3-phosphate and ATP. This illustrates an already established result: the structure of a metabolic network is mainly due to the presence of a few metabolites, called hubs (Jeong *et al.* (2000)). These metabolites constitute branching points around which metabolic pathways are organised. The originality in our case comes from the initial removal of secondary metabolites from our dataset which ensures that we identify meaningful hubs, that is, metabolites that really form the backbone of the network. Interestingly, a single hub may be “split” into two groups by our method. Indeed, the connection probability between groups 1 and 16 is 1, so these 2 groups actually constitute a clique together which again corresponds to a single compound (pyruvate). However, they are separated in two sub-cliques because of their very different connectivities with reactions of groups 7 and 10. This distinction is due to the use of two other compounds involved in most reactions of groups 1 and 7 (C02) and 1 and 10 (acetylCoA) but not of group 16. The identification of group 1 inside the pyruvate clique outlines the particular role played by this molecule in metabolism. It is indeed known to be a branching point between central metabolic pathways (glycolysis, TCA cycle, fermentation) and is found here to be a connector of connectors.

Complementary analysis of the groups show that they gather reactions that participate in the same class of metabolic pathways. For instance, group 1 corresponds to the generation of precursor metabolites, group 2 and 3 correspond to amino-acid biosynthesis, and group 4 to cofactor biosynthesis. This indicates that the groups found by our method are coherent in terms of biological processes.

**Table 3.** Parameter estimates of the ERMG model with  $Q = 21$  groups (values smaller than .5 % are masked for readability).

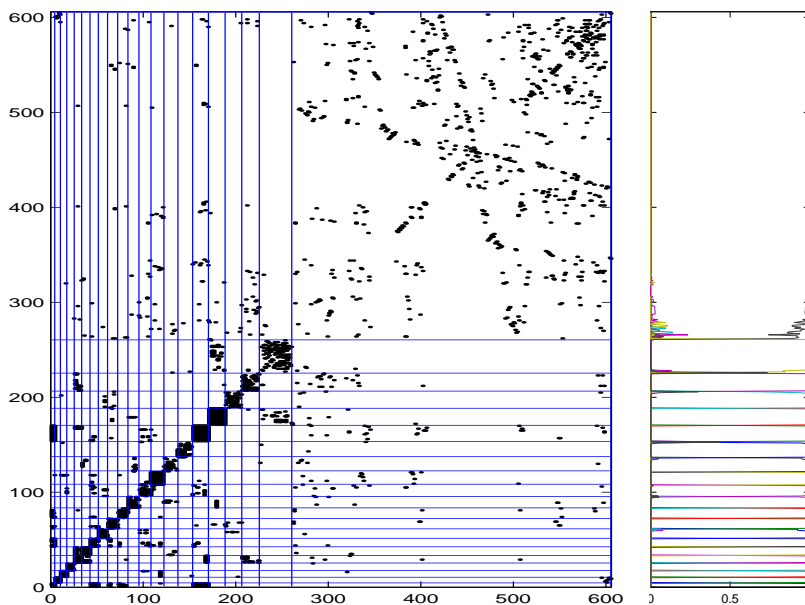
| $\alpha$ (%) | 0.7 | 1.0 | 1.2 | 1.3 | 1.3 | 1.5 | 1.5 | 1.6 | 1.8 | 1.8 | 2.0 | 2.1 | 2.3 | 2.6 | 2.7 | 2.8 | 3.0 | 3.0 | 3.3 | 5.8 | 56.8 |     |    |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|----|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

BETWEEN GROUP CONNECTIVITY AND CLUSTERING COEFFICIENT. The graph showing 1782 edges connecting 605 vertices is of course unreadable. Figure 2 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 emphasise 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  $\hat{\tau}_{iq}$  is maximal.

The bottom plot in Figure 2 gives the estimated *posterior* probabilities  $\hat{\tau}_{iq}$ . We see that the first groups are quite well defined. The last one (21) has more fuzzy limits: it is actually made of isolated reactions having not much in common.

Finally, we also compare the expected clustering coefficient  $c$  given in Proposition 3 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 Erdős-Rényi model would give  $\hat{c} = \hat{\pi} = 0.0098$ .

We conclude that the ERMG model provides a random graph model which seems to be well adapted to capture the structure of a biological network. This first application of our model to a biological network is promising in the sense that the groups we find are relevant (coherent sets of reactions gathered around central compounds). Future research directions include the study of probabilistic properties of the ERMG model (diameter, probability for a subgraph to be connected) which would give a strong statistical basis to the study of local structural properties.



**Figure 2.** Left: Dot plot representation of the graph after classification of the vertices into the 21 groups. Right: Posterior probabilities  $\tau_{iq}$ .

## Acknowledgements

The authors 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. They also thanks F. Forbes (INRIA Grnoble) for her advices regarding the estimation algorithm.

## Références

- ALBERT, R. and BARABÁSI, A. L. (2002). Statistical mechanics of complex networks. *R. Modern Physics*. **74** (1) 47–97.
- ALM, E. and ARKIN, A. P. (2002). Biological networks. *Cur. Op. Struct. Biol.* **13** 193–202.
- ARITA, M. (2004). The metabolic world of *Escherichia coli* is not small. *PNAS*. **101** (6) 1543–1547.
- BARABÁSI, A. L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science*. **286** 509–512.
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel.* **22** (7) 719–725.
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*. **39** 1–38.
- JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. and BARABÁSI, A. L. (2000). The large-scale organization of metabolic networks. *Nature*. **407** 651–654.
- LACROIX, V., GOMES FERNANDES, C. and SAGOT, M. F. (2005). Reaction motifs in metabolic networks. *Proceedings of 5th Workshop on Algorithms for Bioinformatics (WABI'05), Lecture Notes in Bioinformatics, subseries Lecture Notes in Computer Science*. **3692** 178–191.
- MOLLOY, M. and REED, B. . (1995). A critical point for random graphs with a given degree sequence. *Rand. Struct. and Algo.* 161–179.
- NEWMAN, M. E. J. (2003). *Handbook of Graphs and Networks*. (S. Bornholdt and H. G. Schuster, ed.), chapter Random graphs as models of networks. Wiley-VCH: Berlin.
- NEWMAN, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E* (69) 066133.
- NEWMAN, M. E. J., WATTS, D. J. and STROGATZ, S. H. (2002). Random graph models of social networks. *PNAS*. **99** 2566–2572.
- NEWMAN, M. E. J. and GIRVAN, M. (2003). *Statistical Mechanics of Complex Networks*. (R. Pastor-Satorras, J. Rubi, and A. Diaz-Guilera, ed.), chapter Mixing patterns and community structure in networks. Springer: Berlin.
- SHEN-ORR, S. S., MILO, R., MANGAN, S. and ALON, U. (2002). Networks motifs in the transcriptional regulation network of *escherichia coli*. *Nat. Genetics*. **31** 64–68.
- ZHANG, V. L., KING, O. D., WONG, S. L., GOLDBERG, D. S., TONG, A. H. Y., G., L., ANDREWS, B., BUSSEY, H., BOONE, C. and ROTH, F. P. (2005). Motifs, themes and thematic maps of an integrated *saccharomyces cerevisiae* interaction network. *Journal of Biology*. **4** (2) 1–13.