Biometrics 66, 1043–1051
December 2010

DOI: 10.1111/j.1541-0420.2009.01378.x

Model for Heterogeneous Random Networks Using Continuous Latent Variables and an Application to a Tree–Fungus Network

Jean-Jacques Daudin,[1,*] Laurent Pierre,[2] and Corinne Vacher[3]

[1]UMR AgroParisTech/INRA518, AgroParisTech, Paris, France
[2]University Paris X, Nanterre, France
[3]UMR1202 INRA/University Bordeaux I BioGeCo, Bordeaux, France
[*]*email:* jean-jacques.daudin@agroparistech.fr

Summary. The mixture model is a method of choice for modeling heterogeneous random graphs, because it contains most of the known structures of heterogeneity: hubs, hierarchical structures, or community structure. One of the weaknesses of mixture models on random graphs is that, at the present time, there is no computationally feasible estimation method that is completely satisfying from a theoretical point of view. Moreover, mixture models assume that each vertex pertains to one group, so there is no place for vertices being at intermediate positions. The model proposed in this article is a grade of membership model for heterogeneous random graphs, which assumes that each vertex is a mixture of extremal hypothetical vertices. The connectivity properties of each vertex are deduced from those of the extreme vertices. In this new model, the vector of weights of each vertex are fixed continuous parameters. A model with a vector of parameters for each vertex is tractable because the number of observations is proportional to the square of the number of vertices of the network. The estimation of the parameters is given by the maximum likelihood procedure. The model is used to elucidate some of the processes shaping the heterogeneous structure of a well-resolved network of host/parasite interactions.

Key words: Ecological network; Grade of membership model; Heterogeneous random graph; Maximum likelihood; Mixture model.

1. Introduction

Complex networks are extensively studied in different domains such as social sciences and biology. The network representation of the data is graphically attractive, but there is clearly a need for a synthetic model, giving an enlightening representation of complex networks. Statistical methods have been developed for analyzing complex data such as networks in a way that could reveal underlying data patterns through some form of classification.

There are two ways of producing a synthetic representation of such data: multidimensional scaling where a position in a metric space is assigned to each vertex, and unsupervised classification of the vertices using a mixture model.

The first approach is well described in Hoff, Raftery, and Handcock (2002). Hoff (2005) develops a more general model including additional information on vertices. The recent development of the random dot product graphs (RDPG) (Marchette and Priebe, 2008) follows the same approach, with a special focus put on the probabilistic properties of such models (degree distribution, clustering coefficient, giant component) (Young and Scheinerman, 2007).

Unsupervised classification of the vertices of networks is a rapidly developing area with many applications in social and biological sciences. The underlying idea is that common connectivity behavior shared by several vertices leads to their grouping in one *meta-vertex*, without losing too much information. Then, the initial complex network can be reduced to a simpler *meta-network*, with few *meta-vertices* connected by few *meta-edges*. Picard et al. (2009) show applications of this idea to biological networks and Nowicki and Snijders (2001) and Handcock, Raftery, and Tantrum (2007) to social networks.

The literature about classification of vertices can be divided into two classes:

1. Usual mixture model using discrete latent variables giving the assignment of each vertex to a group, where each vertex is supposed to pertain to only one group. Nowicki and Snijders (2001) were among the first to propose what they called a stochastic block structure model because their model was on the line of an older nonstochastic block structure model largely developed in social science. Their estimation method is made through Bayesian Markov chain Monte Carlo (MCMC) algorithms for networks with less than 200 vertices. Daudin, Picard, and Robin (2008) have given more insight on the same model, the degree distribution and the clustering coefficient, and used a variational method for estimating the parameters (`Mixnet`, 2009 package). The variational method allows us to deal with several thousand vertices. Using a different approach, Handcock et al. (2007) assigned a position in a metric space to each vertex and then used a Gaussian mixture model on these positions to cluster the vertices.
2. Individual mixture model, where each vertex pertains partially to several groups, so the mixture is at the individual level and not at the population level, as is the case

for usual mixture models. This class of model has been developed in social science for usual multivariate data, under the name of grade of membership (see Manton, Woodbury, and Tolley, 1994; Erosheva, 2005). The idea is that there are hypothetical extreme profiles and that each sample unit is a mixture of these extreme profiles and inherits their properties through a weighted mean. This idea has been developed under the name of mixed-membership model, see Erosheva, Fienberg, and Lafferty (2004) and Airoldi et al. (2008) for networks.

The proposed model is on the line of the second class. It is similar to the Airoldi et al. (2008) mixed-membership model, but it is expressed in a much simpler form, leaving aside a huge number of random latent variables. A model with a vector of parameters for each vertex is tractable because the number of observations is proportional to the square of the number of vertices of the network.

The estimation step for mixture models for networks is a difficult task. Maximum likelihood procedure is generally not possible, due to the huge dimension of the space where the latent discrete variables reside. Daudin et al. (2008) and Airoldi et al. (2008) use variational methods and Nowicki and Snijders (2001) and Handcock et al. (2007) use MCMC. The statistical properties of variational estimates are not well known. They maximize a pseudolikelihood and are, by definition, inferior to maximum likelihood estimates. MCMC is highly computationally intensive and their mixing properties for high-dimensional discrete variables are doubtful for large networks. Conversely, a great advantage of the proposed model is that it allows us to obtain standard maximum likelihood with a quick and robust algorithm. We restrict our interest to the case of pure relational information between vertices, putting aside any additional information on vertices. The intensity of relation between vertices may be continuous or binary. In this article we deal with binary variables. Extension to a more general case is possible, but this is not done in this article.

The most salient characteristic of the proposed model is that it is based on extremal hypothetical vertices. Therefore we will call it extremal vertices model for random graph (EVMRG). We define the EVMRG model in Section 2. In Section 3, we give a maximum-likelihood estimation algorithm.

In Section 4, we use the EVMRG model to synthesize the heterogeneity of an ecological network, i.e., a network having species as vertices and interspecific interactions as edges. Ecological networks have long fascinated biologists because of the diversity and the complexity of the interactions between species. In the *Origin of Species* (Darwin, 1869), Charles Darwin wrote: *It is interesting to contemplate a tangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent on each other in so complex a manner, have all been produced by laws acting around us.* Comparative analyses of ecological networks have highlighted some invariant topological properties, confirming that there are common laws governing the structure of apparently diverse species assemblages. Uncovering these laws is a crucial challenge for biologists because it would allow important advances in conservation and environmental management.

Modularity is a prevalent topological feature in large ecological networks (i.e., >150 species). Modules, also named compartments in the ecological literature, are recognizable subsets of interacting species, with species more likely to be linked within than across subsets (Lewinsohn et al., 2006). On the ecological timescale, modules may arise through spatial or temporal segregation of the species. Species occurring in the same place and at the same time are more likely to fall into the same module, because they have a higher probability of interacting with each other than with species occurring elsewhere or at another time. However, modularity may also reflect more ancient events, such as phylogenetic splits.

Various methods have been used for detecting modularity in ecological networks (e.g., correspondence analysis [Lewinsohn et al., 2006]; edge betweenness algorithm [Vacher, Piou, and Desprez-Loustau, 2008a]; simulated annealing algorithm [Olesen et al., 2007]. Most often, each species is assigned to one module (only) and the network is finally represented as a set of nonoverlapping modules. Such simplification of the network structure is an issue because some species may actually not belong to any module because they are loosely linked to all the other species of the network, or may belong to several modules. For instance, species with large spatial or temporal distributions, or species with complex life-cycles going through very different habitats during their lives, are likely to interact with species belonging to different modules. The misclassified species may obscure the common or complementary features of the species belonging to a module and therefore hinder our understanding of the processes shaping species webs.

Grade of membership models present the advantage of allowing the species to have intermediate positions in the simplified representation of the network. To our knowledge, they have never been used for synthesizing the heterogeneity of ecological networks. In this study, we used the EVMRG model to synthesize the heterogeneity of a well-resolved interaction network between forest tree species and parasitic fungal species, which was shown to be modular in a previous study (Vacher, Piou, et al., 2008). Then we searched for the factors governing the position of the species in the model. A wide range of potential factors was investigated, including the phylogenetic history of the species, their life-history strategy, their introduction status, and the intensity with which they were sampled.

2. Model EVMRG

2.1 *Model*

Vertices. Consider a graph with n vertices, labeled in $\{1,\ldots,n\}$. The model is based on Q hypothetical unobserved extreme vertices.

Each vertex i is the weighted mean of Q extreme hypothetical vertices (EHV), with weights given by $Z_i = (z_{i1},\ldots,z_{iQ})$, with $z_{iq} \geqslant 0$ and $\sum_q z_{iq} = 1$. Q is assumed to be a fixed constant with $Q << n$. Q will be determined as part of a model-selection problem in Section 3.3. The Q extreme vertices are put at the end of the canonical unit vectors $(1,0\ldots0)$, $(0,1,0\ldots0)\ldots(0\ldots0,1)$ in $\boldsymbol{R}^Q$ in an arbitrary order. The set of vertices $\{1,\ldots,n\}$ is contained in the simplex $S_Q = \{x, \in [0,1]^Q, \sum_{q=1}^{Q} x_q = 1\}$, so that the EHV are extreme points of S_Q. Each EHV is supposed to be typical of the group of vertices that are near it in S_Q, with more extremal connectivity properties than its neighboring real vertices.

Edges. Each edge from a vertex i to a vertex j is associated to a binary random variable X_{ij} following a Bernoulli distribution with probability P_{ij}. The probability that there is an edge from EHV q to EHV l is equal to a_{ql}. The connectivity properties of each vertex i are a mixture of the connectivity properties of the EHV so that P_{ij} can be expressed using the weights z_{iq} and z_{jl} and the connectivity matrix A between the EHVs:

$$P_{ij} = \sum_{q,l=1,Q} z_{iq} a_{ql} z_{jl}$$

which gives the matrix relation

$$P = ZAZ',$$

with

- P the (n,n) matrix containing the p_{ij},
- Z the (n,Q) matrix containing the z_{iq} and Z' the transpose of $Z, Z \in S_Q^n$,
- and $A \in [0,1]^{Q^2}$, the (Q,Q) matrix containing the a_{ql}, the connectivity matrix between the EHVs.

The random variables X_{ij} are assumed to be independent. Let X be the (n,n) matrix containing the random variables X_{ij}. Finally the model is summarized by

$$X \sim \boldsymbol{B}(ZAZ') \qquad (1)$$

where $\boldsymbol{B}$ denotes the Bernoulli distribution, $Z \in S_Q^n$, and $A \in [0,1]^{Q^2}$.

The parameters of the model are A and Z. This model may be classified in the set of the semiparametric statistical models, for each individual (vertex) has its own set of parameters $(z_{i1}, \ldots, z_{iQ})$. Using statistical models, it is generally impossible to estimate as many parameters as the number of individuals. Moreover there are $Q^2 + n(Q-1)$ parameters, so this number approaches infinity with n. However, the number of observations contained in X is not proportional to n but to n^2, so the ratio of the number of parameters with the number of observations approaches 0 when $n \to \infty$. In practice, for each vertex i, there are n data, $(x_{i1}, \ldots, x_{in})$, available to estimate the $Q-1$ linearly independent parameters contained in the vector $(z_{i1}, \ldots, z_{iQ})$.

We can choose whether the graph is directed or indirected by leaving the X_{ij} loose or setting $X_{ij} = X_{ji}$ for all i, j. If the graph is directed A contains Q^2 parameters. If the graph is indirected, A is symmetric and contains $\frac{Q(Q+1)}{2}$ parameters. Note that we assume in the following that there is no self-loop ($X_{ii} = 0$, for $i = 1, n$).

2.2 *Relation between EVMRG and Other Models*

Several models have been proposed with a functional form similar to $X \sim \boldsymbol{B}(ZAZ')$: the mixture model for random graphs, the RDPG, and the mixed membership stochastic blockmodel (MMB). Table 1 summarizes the functional definition of the different models. Note that the model proposed by Hoff et al. (2002) and Handcock et al. (2007) do not have this functional form and is not included in this comparison.

2.2.1 *Relation between EVMRG and Mixture Model.* In a mixture model for random graphs (Nowicki and Snijders, 2001; Daudin et al., 2008), the variables Z are random and are equal to 0 or 1. In the EVMRG model the variables Z are fixed parameters, and take their values in the simplex S_Q^n. In a mixture model, each vertex is assumed to pertain to only one group. The mixture model is a mixture of populations of pure vertices. In the EVMRG model, each vertex is a compound of EHV, so the mixture is at the individual level. However, there are two practical applications of the two models:

1. The clustering of the items, i.e., the classification of each item in a group. The key element is $E(Z/X = x)$ in the mixture model and directly Z for EVMRG. Note that $E(Z/X = x)$ in the mixture model, and Z in EVMRG, take their values in the same set S_Q^n.
2. The connectivity matrix A is the key element for the description and interpretation of groups in the two models, see Daudin et al. (2008). In the mixture model, A is the mean connectivity matrix in the sense that the probability of connection is the weighted mean of the connections between the vertices. In the EVMRG, however, A represents an extreme connectivity matrix. As a result A is more contrasted in EVMRG than in the mixture model.

2.2.2 *Relation between EVMRG and RDPG.* The multidimensional scaling method, applied to the similarity matrix P,

Table 1

Summary of the models $X \sim \boldsymbol{B}(P = f(ZAZ'))$ or $X \sim \boldsymbol{B}(P = f(UAV'))$, with VEM = variational EM, OLS = ordinary least squares, $S_Q = \{x, \in [0,1]^Q, \sum_{q=1}^{Q} x_q = 1\}$, $\boldsymbol{B}(.)$ is the Bernoulli probability distribution function, and $\boldsymbol{M}(.)$ is the multinomial probability distribution function with one trial and Q classes

Model	f	Z	A	P	Estimation method
EVMRG	Id	$Z \in S_Q^n$	$A \in [0,1]^{Q^2}$	$P = f(ZAZ')$	ML
Mixture model	Id	$Z \in S_Q^n$ and binary	$A \in [0,1]^{Q^2}$	$P = f(ZAZ')$	VEM /MCMC
RDPG	$f : \boldsymbol{R} \to [0,1]$, monotone	$U, V \in \boldsymbol{R}_Q^n$	$A = Id$	$P = f(UAV')$	OLS
DEDICOM	Id	$Z \in \boldsymbol{R}_Q^n,\ Z'Z = I$	$A \in \boldsymbol{R}^{Q^2}$	$P = f(ZAZ')$	OLS
MMB	$f(x) = \rho x,\ \rho \in]0,1]$	$Z \in S_Q^n,\ U_{ij} \sim \boldsymbol{M}(Z_i)$, $V_{ji} \sim \boldsymbol{M}(Z_j)$	$A \in [0,1]^{Q^2}$	$P_{ij} = f(U'_{ij} A V_{ji})$	VEM

consists in positioning each vertex in a metric space so that the similarity between vertices is approximatively kept. The underlying model is $P = TT'$, where the (n, k)-matrix T contains the coordinates of the vertices in a k-dimensional metric space. The naive multidimensional scaling method is not well suited for modeling P, with two major drawbacks: TT' does not lie in $[0,1]^{n^2}$ if $T \in \boldsymbol{R}^k$ and TT' is symmetric so it is not suited for the modeling of directed graphs.

The RDPG defined in Marchette and Priebe (2008) is

$$P_{ij} = f(t_i' t_j) \text{ with } t_i \in \boldsymbol{R}^k \text{ and } f(x) \in [0,1].$$

f is a simple threshold in Marchette and Priebe (2008): $f(x) = 0$ if $x < 0$, $f(x) = x$ if $0 \leqslant x \leqslant 1$ and $f(x) = 1$ if $x > 1$. Young and Scheinerman (2007) propose to constrain T to lie in $\frac{1}{\sqrt{k}}[0,1]^k$.

To get around the second drawback, the RDPG model is extended with two vectors for each vertex, an in-vector V and an out-vector U, so the model becomes $P_{ij} = f(u_i'.v_j)$.

Another way to get around the symmetry of P, is the DEcomposition into VIrectional COMponents, called DEDICOM, which was proposed by Harshman (1978) and well described in Trendafilov (2002). This model uses only one vector for each vertex but inserts a nonsymmetric (k, k)-matrix A in the dot product. The model is

$$X = TAT' + E$$

the matrix T is constrained by $T'T = I$ and T and A are obtained by minimizing $\|X - TAT'\|^2$. Several algorithms have been proposed to achieve this task (see Kiers et al., 2002).

2.2.3 *Relation between EVMRG and MMB.* The MMB (see Airoldi et al., 2008) is similar to EVMRG, with a more complex setting, which is not easy to understand:

- The lines of Z (i.e., the random vectors of weights $Z_i = (z_{i1} \dots z_{iQ})$) are assumed to be identically and independently distributed along a Dirichlet distribution with parameter α
- For each pair of vertices (i, j) in this order, two multinomial random variables $U_{i \to j}$ and $V_{i \leftarrow j}$ are generated with respective probabilities Z_i and Z_j
- A is a (Q, Q) matrix $\in [0,1]^{Q^2}$
- ρ is a sparsity parameter
- X_{ij} is a Bernoulli random variable with probability $\rho U'_{i \to j} A V_{i \leftarrow j}$

The EVMRG is essentially a marginalized version of the MMB model: the MMB model assumes a hierarchical structure: $X|U, V, A$ and $U, V|Z$, whereas the EVMRG integrates U, V from this structure to obtain $X|A, Z$. Moreover the EVMRG model does not need the ad hoc sparsity parameter ρ.

2.3 *Model Identifiability*

As defined so far, the model is not identifiable. Let P be a known matrix and assume that A and Z exist so that $P = ZAZ'$. It is generally possible to find other sets of parameters $\tilde{A}$ and $\tilde{Z}$ so that $P = \tilde{Z}\tilde{A}\tilde{Z}'$. Let H be a (Q, Q) matrix with the following properties (called $H-$properties):

(1) H^{-1} exists
(2) $H\mathbf{1}_Q = \mathbf{1}_Q$, with $\mathbf{1}_Q = (1 \dots 1)'$, made of Q ones
(3) $\tilde{Z} = ZH \geqslant 0$
(4) $\tilde{A} = H^{-1}AH'^{-1} \in [0,1]^{Q^2}$

Then we have:

- $\tilde{Z}\tilde{A}\tilde{Z}' = ZHH^{-1}AH'^{-1}H'Z' = P$
- $\tilde{Z}\mathbf{1}_Q = ZH\mathbf{1}_Q = Z\mathbf{1}_Q = \mathbf{1}_Q$ so $\tilde{Z} \in S_Q^n$ by condition 3
- $\tilde{A} \in [0,1]^{Q^2}$ by condition 4

so $(\tilde{A}, \tilde{Z})$ and (A, Z) are equivalent admissible sets of parameters.

The existence of such $H-$matrix is proved in Web Appendix A, with a toy example for illustration.

We propose to choose Z, which maximizes $Tr(ZZ')$ among the equivalent versions of (A, Z). The choice is motivated by two reasons: This constraint implies unicity of (Z, A) provided that $n \gg Q$ and the n vertices are different. Moreover the EHVs should not be too far from real vertices to confer upon them some reality. This closeness between EHV and some vertices is naturally provided by the maximization of $Tr(ZZ')$.

Finally the model is now:

$$X \sim \boldsymbol{B}(Z'AZ) \tag{2}$$

where $\boldsymbol{B}$ denotes the Bernoulli distribution, $Z \in S_Q^n$, $A \in [0,1]^{Q^2}$, and $Tr(Z'Z)$ is maximum.

3. Parameter Estimation

The log likelihood is

$$L = \sum_{i \neq j} x_{ij} \log\left(\sum_{q,l=1,Q} z_{iq} a_{ql} z_{jl}\right) + (1 - x_{ij}) \log\left(1 - \sum_{q,l=1,Q} z_{iq} a_{ql} z_{jl}\right) \tag{3}$$

and the constraints on the parameters are

$$\begin{aligned} A &\in [0,1]^{Q^2} \\ Z &\in S_Q^n. \end{aligned}$$

Note that the set of admissible solutions, $[0,1]^{Q^2} \times S$, is a convex polyhedron.

3.1 *Log-Likelihood Derivatives*

After some algebraic manipulations we obtain

$$\frac{\partial L}{\partial Z} = RZA' + R'ZA$$

with R a (n, n) matrix with $r_{ij} = \frac{x_{ij} - p_{ij}}{p_{ij}(1 - p_{ij})}$, and

$$\frac{\partial L}{\partial A} = Z'RZ.$$

3.2 *Estimation Algorithm*

3.2.1 *Algorithm.* The constraints on the parameters are linear, but the log likelihood is not linear.

Let $A^{(k)}$ and $Z^{(k)}$ be the parameter estimates at step k, $P^{(k)} = Z^{(k)}A^{(k)}Z^{(k')}$ and $R^{(k)}$ a (n, n) matrix with $r_{ij}^{(k)} = \frac{x_{ij} - p_{ij}^{(k)}}{p_{ij}^{(k)}(1 - p_{ij}^{(k)})}$.

The linear approximation of the log likelihood (3) at point $(A^{(k)}, Z^{(k)})$ is

$$L(A,Z) \approx L(A^{(k)}, Z^{(k)}) + Tr\left[(A - A^{(k)})' \frac{\partial L}{\partial A}(A^{(k)}, Z^{(k)})\right]$$
$$+ Tr\left[(Z - Z^{(k)})' \frac{\partial L}{\partial Z}(A^{(k)}, Z^{(k)})\right].$$

The algorithm is the following:

- Find initializing values $(A^{(0)}, Z^{(0)})$
- At step $(k+1)$ use a linear programming algorithm to maximize the function in (A, Z):

  $$f_k(A,Z) = Tr\left[A'Z^{(k)'}R^{(k)}Z^{(k)}\right]$$
  $$+ Tr\left[Z'(R^{(k)}Z^{(k)}A^{(k)'} + R^{(k)'}Z^{(k)}A^{(k)})\right]$$

  under the constraints: $A \in [O,1]^{n^2}$ and $Z \in S_Q^n$.
  Let Z^{LP_k} be the solution of the previous linear program. Compute $L(A, Z)$ on regularly spaced points along the line $(A^{(k)}, Z^{(k)}) \rightarrow (A^{LP_k}, Z^{LP_k})$ and keep the best one, $(A^{(k+1)}, Z^{(k+1)})$. A further improvement around this point and along the same line is obtained by dichotomy. Then, go to step $k+2$ if the following stopping rule is not true.

  $$|L(A^{(k+1)}, Z^{(k+1)}) - L(A^{(k)}, Z^{(k)})| < \alpha.$$

3.2.2 *Initialization.* The algorithm is convergent because the likelihood is increased at each step. However, it may converge to a local maximum depending on the initialization. We use several random initializations based on *k*-means and select the best starting point. Another possibility would be to use the algorithm described by Kiers et al. (2002) for DEDICOM.

3.2.3 *Assessment of the Identification of the Model.* The model is not identifiable as it stands (see Section 2.3). In practice we have not seen any problem coming from the lack of identifiability when using the above algorithm. After convergence, we obtain a unique instance of the equivalent class of parameters (A, Z) by maximizing $Tr(Z'Z)$ under the constraint that $ZAZ' = Z^{(k)}A^{(k)}Z^{(k)}$, with k the iteration number at convergence.

3.3 *Choice of the Number of Groups*

Several criteria have been proposed for choosing the number of groups in finite mixture models, such as Akaike information criteria (AIC), Bayesian information criteria (BIC), or integrated completed likelihood (ICL), see McLachlan and Peel (2000) and Biernacki, Celeux, and Govaert (2000). From a theoretical point of view, penalized likelihood criteria are asymptotically consistent under some conditions satisfied by BIC but not by AIC, see Gassiat (2002). From a practical point of view and for moderate sample sizes, AIC has a known tendency to overestimate the number of groups for Gaussian mixtures but gives correct results for latent class models. Conversely BIC give good results for Gaussian mixtures but underestimates the number of groups for latent class models. AIC is equal to minus two times the log likelihood plus two times the number of estimated parameters, and BIC has a similar definition with the logarithm of the number of observations in place of the coefficient 2.

- For directed networks:

  $$AIC(Q) = -2L(\hat{A}_Q, \hat{Z}_Q) + 2(Q^2 + n(Q-1))$$
  $$BIC(Q) = -2L(\hat{A}_Q, \hat{Z}_Q) + Q^2 \log(n(n-1)) + n(Q-1)\log(n).$$

- For indirected networks:

  $$AIC(Q) = -2L(\hat{A}_Q, \hat{Z}_Q) + 2(Q(Q+1)/2 + n(Q-1))$$
  $$BIC(Q) = -2L(\hat{A}_Q, \hat{Z}_Q) + Q(Q+1)/2 \log(n(n-1)/2) + n(Q-1)\log(n).$$

$(\hat{A}_Q, \hat{Z}_Q)$ are the maximum likelihood estimates of (A, Z) for Q groups. Some more theoretical work is needed to study the asymptotic properties of these criteria in the context of EVMRG, for the number of parameters tends to infinity with the number of vertices, which is a nonstandard framework. In this article, we use these criteria from a practical point of view and without any theoretical background. We have made some simulations to see if these criteria are able to recover the true number Q. For low and moderate sample sizes AIC has given good results, better than BIC which underestimated Q. Therefore we use AIC in the study of the following example.

4. Example

4.1 *Data*

The ecological network considered in this study consisted of 543 interactions among 51 forest tree taxa (all but 6 being true species or groups of cultivars belonging to the same genetic continuum) and 154 parasitic fungal species. The network is composed of 205 vertices and 543 edges. It is a bipartite graph because tree–fungus interactions are the only possible ones. All the observations of tree–fungus interactions originated from the database of the French governmental organization in charge of forest health monitoring (the *Département Santé des Forêts (DSF)*) for the 1972–2005 period. The methods used for data collection have been described in more detail in previous analyses of the DSF database (Vacher, Piou, et al., 2008; Vacher, Vile, et al., 2008). We have rechecked fungal species names in the Index Fungorum database (`www.indexfungorum.org`) since our initial analyses: 17 species names were updated, and three of the previously used species names were found to be synonymous. The fusion of synonymous species accounts for the smaller number of fungal species in this study than in our previous study (154 versus 157 in the previous study Vacher, Piou, et al., 2008) and the slightly smaller number of interactions (543 versus 547).

We characterized each tree species by phylum (Magnoliophyta or Conipherophyta) and introduction status (alien or native). An estimate of the area covered by each tree species was also available (*Inventaire Forestier National*, 2000 census report, `http://www.ifn.fr/spip`). An estimate of the total number of times each tree species had been encountered and examined by foresters during their daily work was also available from the DSF database. This variable is called "sampling intensity" and is positively correlated with area, because foresters encounter abundant tree species more frequently than rare species during their daily work (Vacher, Piou,

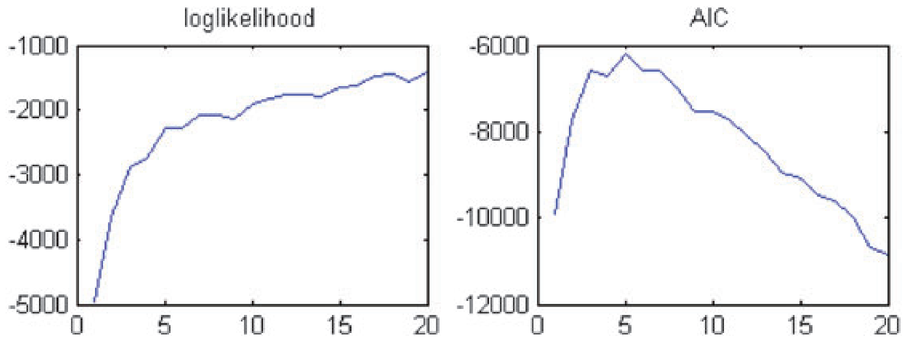

Figure 1. Evolution of the log likelihood and the AIC criteria as a function of the number of EHV, x-axis: number of extremal vertices, left side y-axis: log likelihood, right side y-axis: AIC. This figure appears in color in the electronic version of this article.

et al., 2008). The definition of tree species as aliens or species native to France was not an easy task, because the composition of European forests has been profoundly modified by human activities (Petit et al., 2004). In this study, we considered a tree species to be alien if it was introduced into France after the beginning of the modern era (which we define as the discovery of the New World by Columbus). Such recently introduced species are known as *neophytes*.

Each fungal species was characterized by phylum (Ascomycota or Basidiomycota), introduction status (alien or native), and life-history strategy. As suggested by Garcia-Guzman and Morales (2007), life-history strategies were described in terms of the parasitic lifestyle (biotrophic versus necrotrophic) and the plant organs and tissues attacked: (1) strict foliar necrotrophic parasites, (2) canker agents, (3) stem decay fungi, (4) obligate biotrophic parasites, (5) root decay fungi, (6) other foliar and twig necrotrophic parasites, (7) stem blue stain agents, (8) parasites of fine roots, (9) wilting agents, and (10) other root fungi. The first five strategies accounted for 87% of the fungal species. As for the tree species, it was not a straightforward task determining which fungal species were aliens (Desprez-Loustau, 2009). In this study, we considered a fungal species to be alien if there was documentary evidence that this species was first described in France after 1850 and good evidence that it was introduced from elsewhere.

4.2 *Main Results*

The AIC criteria (Figure 1) indicated that the optimal number of EHVs for the tree–fungus network was 5. The connectivity matrix between the EHVs (Table 2) was symmetric because the network was indirected. It indicated that one of the EHVs (hereafter called FT0) had no connection with all the other EHVs whereas the four remaining EHVs (hereafter called F1, F2, T1, and T2) formed two pairs of highly connected vertices. The matrix of weights $\widehat{Z}$ indicated that each real vertex was a mixture of three EHVs only. All the real vertices representing tree species were a mixture of FT0, T1, and T2 whereas all the real vertices representing fungal species were a mixture of FT0, F1, and F2. Therefore, T1 and T2 were two virtual tree species. They were highly connected with the virtual fungal species F1 and F2, respectively. In the data X, the zeros between trees (respectively between fungal species) are structural ones, but this information is not included a priori in the model. This bipartite network structure is recovered in the results: the EHV are tree-EHV (T1 and T2) or fungi-EHV (F1 and F2) (except the isolated EHV FT0 with no connection with any other EHV), and the probability of connection between T1 and T2 is null and the same is true for F1 and F2.

Table 2
Connectivity matrix A between the five extremal vertices

	FT0	T1	T2	F1	F2
FT0	0	0	0	0	0
T1	0	0	0	0.996	0
T2	0	0	0	0	0.985
F1	0	0.996	0	0	0
F2	0	0	0.985	0	0

In the following the positions of the vertices in the triangular representations are "analyzed" graphically by annotating these classifications with several species descriptors (phylum, life-history strategy, introduction status, number of interactions). There are two possibilities to formalize these graphical analyses: the first one consists in using a linear model with the values of Z as (multivariate) response and the species descriptors as independent variables. The second one would be to extend the EVMRG model by including covariates giving some information about each vertex. The first type of analysis has been done and confirms the graphical results (data not shown), and we are working on the model extension approach.

The projection of phylogenetic data in the triangular representation of tree species showed that the tree species belonging to the Magnoliophyta (angiosperms) and the tree species belonging to the Conipherophyta (gymnosperms) had very different connection profiles (Figure 2). T1 was close to seven gymnosperm species which are highly represented in the French forests (*Abies alba, Abies grandis, Picea excelsa, Pinus laricio, Pinus pinaster, Pinus sylvestris, and Pseudotsuga menziesii*). All the other gymnosperm species were located on the line joining T1 and FT0, suggesting that they all had a subset of the interactions realized by the seven tree species close to T1. This result confirmed the nested pattern of interactions found in a previous study (Vacher, Piou, et al., 2008). T2 was close to six tree taxa belonging to the Magnoliophyta, which are also dominant in the French forests (large maples, cultivated poplars, beech [*Fagus*

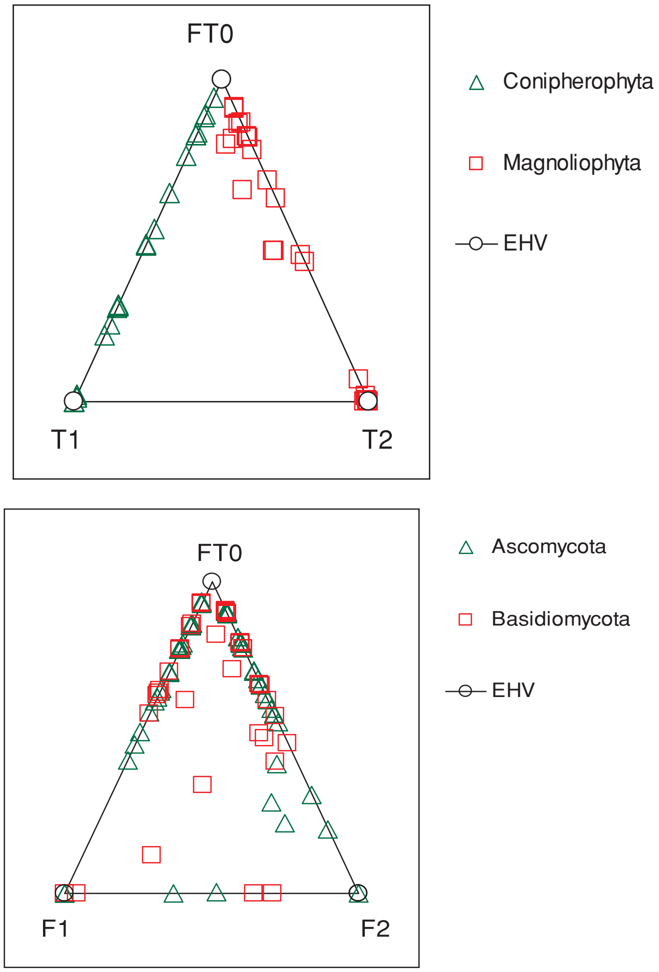

Figure 2. Triangular representations of tree species (top) and fungal species (bottom) as a function of their phylogenetic origin. This figure appears in color in the electronic version of this article.

silvatica], and three species of oaks [*Quercus petraea, Q. pubescens and Q. rubra*]. Five species belonging to the Magnoliophyta were not located on the line joining T2 and FT0, suggesting that the associations of angiosperms with parasitic fungi were slightly more diverse than those of gymnosperm species. This result is consistent with other studies (Vacher, Piou, et al., 2008) and may be accounted for by the wider distributional range of angiosperm species. Among the five species mentioned above, three belonged to the Rosaceae family (*Prunus avium, Sorbus aria, Sorbus torminalis*). It is noteworthy that these three species were classified with gymnosperm species in a previous analysis of the tree–fungus network in which each vertex was assumed to pertain to one group (Vacher, Piou, et al., 2008). The approach used here revealed that the connection profiles of these three species were mixtures between the typical profile of angiosperms (T2) and the typical profile of gymnosperms (T1), but were actually closer to the typical profile of angiosperms.

In contrast, the projection of phylogenetic data in the triangular representation of fungal species (Figure 2) showed that the species belonging to the Ascomycota and the Basiodiomycota had similar connection profiles. F1 was close to one generalist fungal species belonging to the Basidiomycota (*Armillaria ostoyae*) and two generalist fungal species belonging to the Ascomycota (*Sphaeropsis sapinea and Sydowia polyspora*). According to the connectivity matrix, these parasitic fungal species were highly specialized on gymnosperms. The species located on the line joining F1 and FT0 also belonged both to the Ascomycota and the Basiodiomycota. F2 was close to one species only, which belonged to the Ascomycota (*Botryosphaeria stevensii).* However, the species located on the line joining F2 and FT0, which were specialized on angiosperm species according to the connectivity matrix, belonged both to the Ascomycota and the Basidiomycota. Therefore, the phylogenetic history of fungal species did not account for their specialization on gymnosperms or angiosperms.

4.3 *Discussion and Conclusions about the Example*

Our results confirmed that the heterogeneous structure of the network mostly results from the deep evolutionary history of seed plants (Vacher, Piou, et al., 2008). Angiosperm species and gymnosperm species had very contrasted interaction profiles, except when they covered low areas and were consequently not intensively monitored for their fungal diseases (see Web Appendix B1). In contrast, parasitic fungal species belonging to the Ascomycota and the Basidiomycota had very similar interaction profiles. A possible explanation for this asymmetric phylogenetic signal may be that, to survive, parasitic species had no other choice than to adopt an opportunistic feeding behavior, which decreased the relationship between their phylogenetic similarity and the similarity in their interaction profiles. In contrast, the relationship between the phylogenetic similarity of tree species and the similarity in their interaction profiles may have been maintained because adaptations allowing tree species to defend against or avoid parasitic fungal species were least favored by natural selection. Our results (see Web Appendix B2) also showed that the parasitic fungal species having the most opportunistic feeding behavior (i.e., able to attack both angiosperms and gymnosperms) were mainly fungal species with high saprophytic abilities, belonging to stem or root decay fungi. Therefore our results confirmed that the ability to survive well without a host may increase the opportunities for and the likelihood of host shifts (Parker and Gilbert, 2004). Finally, our results (see Web Appendix B3) showed that alien tree species and alien fungal species were well integrated into the network. This rapid integration was unexpected for a plant–pathogen network, because selection is supposed to act continually on plants, favoring the emergence of defenses against new pathogens, and impeding the development of new interactions (Parker and Gilbert, 2004; Thompson, 2006).

Our study showed that the amount of information obtained from the EVMRG model in the case of a host–parasite network was considerable. The EVMRG model therefore appears as a good approach for synthesizing the heterogeneity of ecological networks. Applying the EVMRG model to the network of interactions between tree species and parasitic fungal species of the French forests confirmed, with a single analysis, several results obtained in previous studies (Vacher, Vile, et al., 2008) through different analyses. It also suggested that one of the results obtained previously—the classification of three angiosperm tree species belonging to the Rosaceae family in a module containing only gymnosperm species (Vacher, Piou, et al., 2008)—is likely to be false. By allowing the species to have intermediate positions in the simplified representation of the network, the EVMRG model revealed that the interaction profiles of the three species were actually closer to that of angiosperm species. Therefore, previous discussions (Vacher, Piou, et al., 2008) concerning the surprising interaction profiles of tree species belonging to the Rosaceae family should not be given too much importance.

5. Conclusion

The mixture model is a method of choice for modeling heterogeneous random graphs because it contains most of the known structures of heterogeneity: hubs, hierarchical structures, or community structure. One of the weaknesses of mixture models on random graphs is that, at the present time, there is no computationally feasible estimation method that is completely satisfying from a theoretical point of view. The discrete nature of Z implies that one has to explore a space of dimension Q^n, a task that is highly computationally intensive. The discrete values for Z are replaced by continuous ones in the EVMRG model, which leads to an easier optimization problem and allows us to obtain the maximum-likelihood estimates with an efficient algorithm. Moreover the continuous nature of Z allows us to alleviate the assumption of pure units, pertaining only to one group. The EVMRG model is more flexible than the usual mixture model for it includes the possibility for a vertex to have intermediate connectivity properties. This model, which needs a vector of parameters for each vertex, is tractable because we have n data for each vertex. A MATLAB package called CMixnet, allowing one to analyze a network using the EVMRG model, is available at http://www.agroparistech.fr/mia/doku.php?id=productions:logiciels. However, some additional work is necessary to understand the behavior of the maximum-likelihood estimates of n parameters and n^2 observations when $n \to \infty$.

6. Supplementary Materials

Web Appendices A and B, referenced respectively in Sections 2.3 and 4, are available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

Acknowledgements

We thank the *Département Santé des Forêts (DSF)* for allowing us to use their database and the French research consortium *Interactions Biotiques dans les communautés (GDR ComEvol)* for funding.

References

Airoldi, E. M., Blei, D. M., Fienberg, S., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9,** 1981–2014.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22,** 719–725.

Darwin, C. E. (1869). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, 5th edition, p. 611. London: John Murray.

Daudin, J. J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing* **18**(2), 173–183.

Desprez-Loustau, M. L. (2009). The alien fungi of Europe. In *Handbook of Alien Species in Europe*, W. Nentwig, P. Hulme, P. Pysek, and M. Vila (eds), vol. 3, p. 400. Berlin: Springer-Verlag.

Erosheva, E. (2005). Comparing latent structures of the grade of membership, Rasch and latent class model. *Psychometrika* **70**(4), 619–628.

Erosheva, E., Fienberg, S., and Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* **101,** 5220–5227.

Garcia-Guzman, G. and Morales, E. (2007). Life-history strategies of plant pathogens: Distribution patterns and phylogenetic analysis. *Ecology* **88,** 589–596.

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Annales de l'Institut Henri Poincaré*, **38,** 897–906.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A* **54,** 301–354.

Harshman, R. A. (1978). Model for analysis of asymmetrical relationships among N objects or stimuli. *First Joint Meeting of the Psychometric Society and the Society of Mathematical Psychology*, Hamilton, Ontario, Canada.

Hoff, D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association* **100,** 286–295.

Hoff, D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approach to social network analysis. *Journal of the American Statistical Association* **97,** 1090–1098.

Kiers, H. A. L., ten Berge, J. M. F., Takane, Y., and De Leeuw, J. (2002). A generalization of Takane's algorithm for DEDICOM. *Psychometrika* **55,** 151–158.

Lewinsohn, T. M., Prado, P. I., Jordano, P., Bascompte, J., and Olesen, J. M. (2006). Structure in plant-animal interaction assemblages. *Oikos* **113,** 174–184.

Manton, K. G., Woodbury, M. A., and Tolley, H. D. (1994). *Statistical Applications Using Fuzzy Sets.* New York: Wiley Interscience.

Marchette, D. J. and Priebe, C. E. (2008). Predicting unobserved links in incompletely observed networks. *CSDA* **52,** 1373–1386.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Mixnet (2009). http://stat.genopole.cnrs.fr/software/mixnet/, accessed November 21, 2009.

Nowicki, K. and Snijders, T. (2001). Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association* **96,** 1077–1087.

Olesen, J. M., Bascompte, J., Dupont, Y. L., and Jordano, P. (2007). The modularity of pollination networks. *Procedings of the National Academy of Sciences* **104,** 19891–19896.

Parker, I. M. and Gilbert, G. S. (2004). The evolutionary ecology of novel plant-pathogen interactions. *Annual Review of Ecology Evolution and Systematics* **35,** 675–700.

Petit, R. J., Bialozyt, R., Garnier-Gere, P., and Hampe, A. (2004). Ecology and genetics of tree invasions: From recent introductions to quaternary migrations. *Forest Ecology and Management* **197,** 117–137.

Picard, F., Miele, V., Daudin, J. J., Cottret, L., and Robin, S. (2009). Deciphering the connectivity structure of biological networks using MixNet. *BMC Bioinformatics* **10,** `http://www.biomedcentral.com/1471-2105/10/S6/S17`, accessed June 16, 2009.

Thompson, J. N. (2006). Mutualistic webs of species. *Science* **312,** 372–373.

Trendafilov, N. T. (2002). GIPSCAL revisited. A projected gradient approach. *Statistics and Computing* **12,** 135–145.

Vacher, C., Piou, D., and Desprez-Loustau, M.-L. (2008). Architecture of an antagonistic tree/fungus network: The asymmetric influence of past evolutionary history. *PLoS ONE* **3,** e1740.

Vacher, C., Vile, D., Helion, E., Piou, D., and Desprez-Loustau, M. L. (2008). Distribution of parasitic fungal species richness: Influence of climate versus host species diversity. *Diversity and Distributions* **14,** 786–798.

Young, S. J. and Scheinerman, E. R. (2007). Random dot product graph models for social networks. *Conf WAW*, `http://dx.doi.org/10.1007/978-3-540-77004-6_11`.

Received May 2009. *Revised October* 2009.
Accepted October 2009.