ERMG, a new probabilistic model for random graphs

J-J. Daudin[†], V. Miele[⋆], F. Picard[⋆] and S. Robin[†]

[†]UMR INA P-G/ENGREF/INRA MIA 518, Paris,

[⋆]UMR CNRS-8071/INRA-1152, Statistique et Génome, Évry.

picard@genopole.cnrs.fr

The Erdös-Rényi model of a network is simple and possesses many explicit expressions for average and asymptotic properties [1], but it does not fit well to real-world networks. The vertices of these networks are often structured in *prior* unknown classes (functionally related proteins or social communities) with different connectivity properties. We define a generalization of the Erdös-Rényi model called ERMG for Erdös-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 a statistical criterion to select the number of classes. This method is applied to uncover the structure of social and biological networks.

Model

Let $\mathbf{X}$ be the adjacency matrix of a random graph such that $\{X_{ij} = 1\}$ if vertices $\{i\}$ and $\{j\}$ are connected. We suppose that nodes are spread among Q hidden classes, with *prior* probability $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$. These classes are used to model the heterogeneity of connectivity which is observed in real complex networks. We introduce a sequence of hidden independent variables $\mathbf{Z} = (Z_{iq})$ such that $\{Z_{iq} = 1\}$ if vertex $\{i\}$ is in class q. These variables are independent and distributed according to a multinomial distribution:

$$[Z_{i1}, \ldots, Z_{iQ}] \sim \mathcal{M}(1; \alpha_1, \ldots, \alpha_Q).$$

Then we define the conditional distribution of X_{ij} given the class of the vertices. For this purpose we introduce the connectivity matrix $\boldsymbol{\pi} = (\pi_{ql})$ such that:

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

ERMG : Erdös-Rényi Mixture for Graphs

Properties

Degree distribution.

In the ERMG model, the degree distribution is a mixture of binomial distributions approximated by a mixture of Poisson, such that:

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

with $\lambda_q = (n-1)\bar{\pi}_q$ and $\bar{\pi}_q = \sum_l \alpha_l \pi_{ql}$. An interesting feature of ERMG is that classes can also be interpreted in terms of their average degree of connection with parameter λ_q.

Clustering Coefficient.

This coefficient (c) is supposed to measure the aggregative trend of a graph. The clustering coefficient has a probabilistic definition: it 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 $\{1, \ldots, n\}^3$: $c = \Pr\{X_{ij}X_{jk}X_{ki} = 1 \mid X_{ij}X_{ik} = 1\}$. Under the ER model, this probability equals p. In the ERMG model, the clustering coefficient is:

$$c = \sum_{q,l,m} \alpha_q \alpha_l \alpha_m \pi_{ql} \pi_{qm} \pi_{lm} \Big/ \sum_{q,l,m} \alpha_q \alpha_l \alpha_m \pi_{ql} \pi_{qm} .$$

Parameter estimation and model selection

We propose to estimate the parameters of ERMG by maximizing the likelihood $\Pr(\mathbf{X})$ undirectly, using the conditional expectation of the complete-data log-likelihood defined as:

$$\mathcal{Q}(\mathbf{X}) = \mathbb{E}\left[\log \Pr(\mathbf{X}, \mathbf{Z}) | \mathbf{X}\right].$$

However, $\Pr(\mathbf{Z}|\mathbf{X})$ is unknown but can be approximated by $\mathcal{R}[\mathbf{Z}]$, such that $KL(\mathcal{R}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X}))$ is minimal. The principle of variational method is to optimize an approximation of $\Pr(\mathbf{X})$ noted $\mathcal{J}(\mathcal{R}[\mathbf{Z}])$ which depends on $\mathcal{R}$ such that:

$$\mathcal{J}(\mathcal{R}[\mathbf{Z}]) = \log \Pr(\mathbf{X}) - KL(\mathcal{R}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X})).$$

- $\mathcal{J}(\mathcal{R}[\mathbf{Z}])$ has a unique maximum at $\mathcal{R}[\mathbf{Z}] = \Pr(\mathbf{Z}|\mathbf{X})$,
- In practice, $\mathcal{J}(\mathcal{R}[\mathbf{Z}])$ is maximized over a limited set of distributions, and we choose the multinomial distribution $\log \mathcal{R}[\mathbf{Z}] = \sum_i \log h(\mathbf{Z}_i; \tau_i)$, where τ_i is called the variational parameter.

Iterative algorithm.

(h) Optimizing $\mathcal{J}(\mathcal{R}[\mathbf{Z}])$ with respect to $\mathcal{R}[\mathbf{Z}]$ leads to a fixed point equation:

$$\tilde{\tau}_{iq} = \Pr\{Z_{iq} = 1 | \mathbf{X}, \tilde{\mathbf{Z}}^i\}.$$

(h+1) Optimizing $\mathcal{J}(\mathcal{R}[\mathbf{Z}])$ with respect to $(\boldsymbol{\alpha}, \boldsymbol{\pi})$:

$$\tilde{\alpha}_q = \sum_i \tilde{\tau}_{iq}/n, \quad \tilde{\pi}_{ql} = \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl}X_{ij} / \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl}.$$

A statistical criterion to select the number of classes.

We derive a statistical criterion to select the number of classes in a Bayesian setting. This criterion is based on the penalization of the integrated complete-data likelihood :

$$\log \Pr(\mathbf{X}, \mathbf{Z}|m_Q) = \int_{\boldsymbol{\Theta}} \log \Pr(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, m_Q) g(\boldsymbol{\theta}|m_Q) \mathrm{d}\boldsymbol{\theta}.$$

This quantity can be split into two terms [2], $\log \Pr(\mathbf{X}|\mathbf{Z}, m_Q)$ and $\log \Pr(\mathbf{Z}|m_Q)$ which can be calculated separately. This leads to an integrated classification criterion (ICL) for ERMG such that:

$$\mathrm{ICL}(m_Q) = \max_{\boldsymbol{\theta}} \log \Pr(\mathbf{X}, \tilde{\mathbf{Z}}|\boldsymbol{\theta}, m_Q) - \frac{Q(Q+1)}{4} \log \frac{n(n-1)}{2} - \frac{Q-1}{2} \log(n).$$

Applications

Social Network

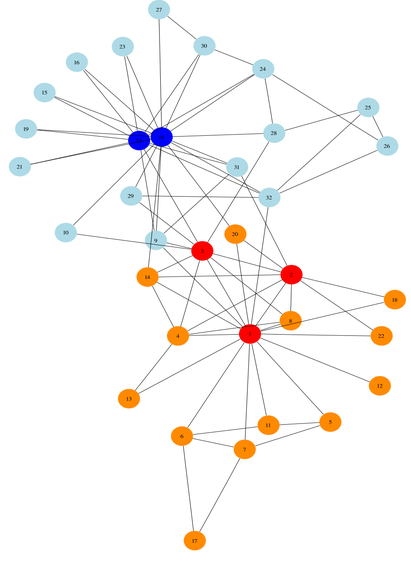
This network represents the connections between 34 members of a karate club over a period of two years. During the course of the study, the administrator and the instructor disagreed, which resulted in a split of the club members into 2 distinct clubs [3]. There are 4 real classes of vertices in Zachary's network : two classes of leaders (vertices $\{1, 2, 3\}$, and vertices $\{33, 34\}$ respectively), and two classes of members differentially associated with classes of leaders. ERMG allows us to uncover this structure perfectly.

Empirical

π_{ql}	1	2	3	4
1	1	0.53	0.15	0.17
2		0.14	0	0.04
3			0.08	0.75
4				1
$\boldsymbol{\alpha}$	0.088	0.382	0.470	0.058

ERMG

π_{ql}	1	2	3	4
1	1	0.53	0.16	0.16
2		0.12	0	0.07
3			0.08	0.73
4				1
$\boldsymbol{\alpha}$	0.089	0.368	0.484	0.058



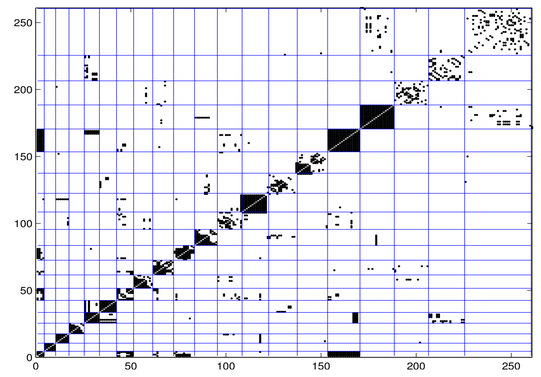
Zachary's karate club network.

Biological Network

This network represents the metabolic network of E. Coli (from http://www.biocyc.org/) which has $n = 605$ vertices (reactions) and $1\,782$ edges. 2 reactions $\{i\}$ and $\{j\}$ are connected if the product of $\{i\}$ is the substrate of $\{j\}$ (cofactors excluded). A compound (chorismate, pyruvate, ATP, *etc*) can be associated to each group. The structure of the metabolic network is governed by the compounds. The ERMG model splits reactions which use Pyruvate according to their connection with other sets of reactions involving CO_2 and Acetyl-Coenzyme A.

π_{ql}	Pyr.	CO_2	A.CoA	Pyr.
Pyr.	**1.0**			
CO_2	.11	.65		
A.CoA	.43		.67	
Pyr.	**1.0**	.01		**1.0**

Estimated connectivity matrix



Adjacency matrix re-organized according to classes

References

[1] Erdös, P. and Rényi, A. (1959), On random graphs, *Publicationes Mathematicae*,6,290–297.

[2] Biernacki, C. and 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.

[3] Girvan, M. and Newman, M.E.J. (2002), Community structure in social and biological networks, *PNAS*, 99(12),7821–7826.