Journal de la Société Française de Statistique
Vol. 152 No. 2 (2011)

A review of statistical models for clustering networks with an application to a PPI network.

Titre: Modèles Statistiques pour la classification non-supervisée des sommets d'un graphe et application à un réseau d'interactions de protéines

Jean-Jacques Daudin[1]

Abstract: Clustering the nodes of a graph allows to analyze the topology of a network. At least three scientific communities (Computer science, Physics and Statistics) proposed some methods to go ahead. We give here an overview about the last developments about heterogeneous random graph models proposed by the statisticians. The Stochastic Block Model is applied to analyze a large Protein-Protein Interaction network

Résumé : La classification non-supervisée des noeuds d'un graphe donne des éléments essentiels sur l'architecture d'un réseau. Il existe des différences d'approche entre les communautés scientifiques (informaticiens, physiciens et statisticiens) qui se sont attaqués à cette question. Nous présentons ici les travaux récents de la communauté des statisticiens, basés sur des modèles de graphes aléatoires hétérogènes et nous analysons un grand réseau d'interactions de protéines avec un modèle de ce type.

Keywords: Biological Networks, Clustering, Random Graph, Mixture Model, Variational Estimation.
Mots-clés : Classification non supervisée, Estimation Variationnelle, Modèle de Mélange, Graphes aléatoires, Réseaux Biologiques.
AMS 2000 subject classifications: 05C80 Random graphs , 62H30 Classification and discrimination; cluster analysis, 60B20 Random matrices , 92B05 General biology and biomathematics.

1. Introduction

Complex networks are more and more 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.

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*. [24] show applications of this idea to biological networks and [23] and [14] to social networks.

The computer scientists build algorithms to cluster the network, such as hierarchical clustering, Spectral clustering based on the Laplacian of the network [30], or Markov Chain Clustering

[1] UMR 518 AgroParisTech-INRA.
E-mail: daudin@agroparistech.fr

 ISSN: 2102-6238

Algorithm based on random walks along the graph, [29]. This is not an exhaustive list at all. Physics scientists maximize a criteria such as Edge-betweenness or the modularity [13]. Statisticians propose a probabilistic model which is supposed to take into account the random variability in the data. This paper presents the most recent results from this school of thought.

The first model for random network is the well known Erdos-Renyi model which assumes that each edge exists with probability p. The probabilistic properties of this model have been largely studied. However it does not fit to any real network because it is too simple and is completely homogeneous.

The first probabilistic model which integrates explicitly heterogeneity in the network topology, the Stochastic block model called SBM in this paper, has been proposed by statisticians working in the domain of social science, Snijders and Nowicki, [26]. Since 1997, much progress has been made about its properties and the estimation procedures and a large part of this paper is devoted to the presentation of the last results about this model.

SBM is a 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.

More flexible models have been proposed recently, which allows a node to pertain to more than one group. Each vertex pertains partially to several groups, so the mixture is at the individual level and not at the population level. This class of model has been developed in social science for usual multivariate data, under the name of Grade of Membership (see [19] and [10]). 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.

The first part of the paper is devoted to the last developments about SBM and the second one to the Grade of Membership models for networks. The third part illustrates the interest of SBM on a large Protein-Protein Interaction network.

2. Recent developments about the SBM Model

2.1. SBM: a mixture model for random graphs

The SBM model is the following:

$i = 1, n$ nodes pertains to $q = 1, Q$ classes. The class of each node is defined by a hidden discrete latent variable, $Z_i = q$ if node i pertains to class q, with Probability Distribution Function (pdf) given by $Z_i \sim \mathscr{M}(1, \alpha_1, \alpha_2, ...\alpha_Q)$ and $\mathscr{M}$ is a multinomial pdf.

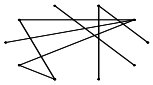
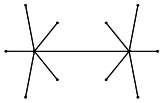
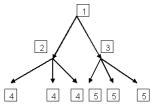
$X_{ij} = 1$ if there is an edge from node i to node j and 0 if there is no edge, and conditionally to Z, X_{ij} are independent Bernoulli random variable with

$$P(X_{ij} = 1/Z_i = q, Z_j = l) = \pi_{ql}$$

The Table 2.1 shows that the model is very flexible for it is able to produce hubs, separate communities or hierarchical structures. For example the second line of table 2.1 present two hubs each one connected to 5 vertices. This network is modelled with a four-classes SBM and the two hubs are respectively the second and third classes in the matrix π.

 ISSN: 2102-6238

TABLE 1. *Examples of SBM models*

Description	Graph	Q	π
Erdos		1	p
Hubs		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
communities		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$
Hierarchical		5	$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

2.2. Recent results about SBM

Model Identifiability The first results on the identifiability of the parameters of SBM have been obtained by Allman et al ([2, 3]), for undirected network (symmetric matrix π): for $Q = 2$ and $n \geq 16$, if the values of π are distinct, the parameters of SBM are identifiable up to label switching. Moreover, for any Q and $n \geq (Q - 1 + \frac{(Q+2)^2}{4})^2$ for n even, the parameters of SBM are generically identifiable (identifiable excepted on a subspace of null Lebesgue measure).

More recently Celisse et al., [6] proved the identifiability for directed and undirected networks for $Q = 2$ and $n \geq 4$ and generic identifiability for any Q and $n \geq 2Q$.

Estimation of parameters

Snijders and Nowicki [26] used Markov Chain Monte Carlo method for estimating the parameters. This method has two limits: its computational complexity is very high, so it cannot be used for networks with more than 200 nodes, and the label switching problem associated with bayesian methods applied to mixture models creates some difficulty to classify the nodes.

Daudin et al (2007) [8] used a variational method of estimation allowing to analyze network up to 3000 nodes. Let $b_{ij}(q,l) = x_{ij} \log \pi_{ql} + (1 - x_{ij}) \log(1 - \pi_{ql})$. In the following, the subscript $[n]$ indicates that the indexed mathematical object is defined for n nodes.

The Log-Likelihood $\mathscr{L}(x_{[n]}; \alpha, \pi) = \log \left\{ \sum_{z_{[n]} \in \mathscr{Z}_n^c} e^{\left[\sum_{i,j \neq i}^n b_{ij}(z_i, z_j)\right]} P_{Z_{[n]}}(z_{[n]}) \right\}$ is not computable even for networks with moderate size because the sum $\sum_z$ runs over Q^n terms.

The Variational log-likelihood approximation is the following, see [8]:

$$\mathscr{J}(x_{[n]}; \tau_{[n]}, \pi, \alpha) = \sum_{(i,j \neq i)=1}^{n} \sum_{(q,l)=1}^{Q} b_{ij}(q,l) \tau_{iq} \tau_{jl} + \sum_{i=1}^{n} \sum_{q=1}^{Q} \tau_{iq} (\log \alpha_q - \log \tau_{iq})$$

for any $\tau \in S_n$, a continuous version of z's, with $S_n = \{u \in ([0,1]^Q)^n \ : \ \forall i = 1:n, \ \sum_{q=1}^{Q} u_{iq} = 1\}$. Note that the variational likelihood is a mean field approximation. In other words, the approximation comes to the fact that $P(Z_i, Z_j | X)$ is assumed to be a product, and the τ_{iq} can be interpreted as approximations of $P(Z_i | X)$.

 ISSN: 2102-6238

The Maximum Likelihood estimates are not computable. The E step of the Expectation-Maximisation (EM) method is highly computationally intensive because it needs to compute n sums of Q^{n-1} terms. It cannot be achieved for networks of size greater than 20 nodes, but an iterative algorithm (see [8]) is available to obtain the variational estimates, $(\widehat{\tau},\widehat{\pi}) = \arg\max_{\tau_{[n]},\pi} \mathscr{J}(x_{[n]};\tau_{[n]},\pi,\alpha)$. There is no proof that the algorithm gives a global maximum, but in practice some simulations indicate that this is the case when n/Q is greater than 20. For smaller values of n/Q, the result may depend on the initial values.

Under some mild conditions about (α,π), and if the number of classes is fixed, the variational estimates of (α,π) are consistent and asymptotically equivalent to the maximum-likelihood estimates, [6]. Moreover most of the labels of the nodes can be retrieved exactly: the number of bad classified nodes divided by n tends to zero when $n \to \infty$.

The proof uses the properties of $\mathscr{J}$, concentration inequalities and an extension of classical methods for proving consistency in [28]. There are two main properties of networks data and the SBM model which are important in this proof: (i) there are n^2 data which helps greatly for obtaining strong concentration inequalities (ii) the asymptotic pdf of $Z|X$ pertains to the factorized class of p.d.f.s in which the variational approximation is searched. These two properties are rarely shared by other data sets and models, so the properties of the variational estimates may be quite specific to random networks.

At the end of 2010 three papers proposed asymptotically consistent methods for SBM for unweighed undirected graphs:

- Bickel and Chen [5] proved a similar result for undirected networks, using a profile likelihood and showed that the profile likelihood is a better criteria than the modularity defined by Girvan and Newman,
- Choi et al. [7] consider a conditional version of SBM with $Z = z$ fixed. They show that the fraction of misclassified nodes converges in probability to zero under maximum likelihood fitting when the number of classes is allowed to grow as the root of n and the average network degree grows at least poly-logarithmically with n.
- Rohe et al. [25] consider the same conditional version of SBM with $Z = z$ fixed and classified the node using the Spectral Clustering algorithm. They show that the fraction of misclassified nodes converges almost surely to zero when the number of classes is allowed to grow.

Latouche et al.([17] and Gazal et al. [12] proposed to make the inference using Variational Bayesian method. This method gives results similar to the frequentist variational method and allows to obtain credibility intervals for the parameters.

Extension to weighted networks and online estimation Mahendra et al [21] have extended the variational estimation method to weighted networks with pdf of the weights pertaining to the exponential family. Ambroise and Matias [4] proposed a consistent and asymptotically normal method for estimating the parameters of weighted random graph mixture models. Zanghi et al. [31] developed an on-line estimation procedure for weighted and unweighed networks.

3. Recent developments about Grade of membership Models

SBM model can be written under the form

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

 ISSN: 2102-6238

where $z_{iq} = 1$ if the vertex i is in class q, and 0 if not, which gives the matrix relation

$$P = ZAZ',$$

with Z the (n,Q)-matrix containing the z_{iq}. If we allow z_{iq} to be in $[0,1]$ (and not in $(0,1)$) then each vertex does not pertain to only one group, which bears more flexibility to the model. This leads to the following CSBM (Continuous-SBM) developed in [9].

3.1. Model CSBM

Vertices Consider a graph with n vertices, labeled in $\{1,\dots,n\}$. 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 $Z_i = (z_{i1},\dots,z_{iQ})$, with $z_{iq} \geq 0$ and $\sum_q z_{iq} = 1$. Q is assumed to be a fixed constant with $Q << n$. The Q extreme vertices are put at the end of the canonical unit vectors $(1,0...0),(0,1,0...0)$ $...(0...0,1)$, in $\mathbb{R}^Q$ in an arbitrary order. The set of vertices $\{1,\dots,n\}$ 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 which 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

$$X \sim \mathbb{B}(ZAZ') \tag{1}$$

where $\mathbb{B}$ 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 semi-parametric statistical models, for each individual (vertex) has its own set of parameters $(z_{i1},...z_{iQ})$. Using statistical models, it is generally impossible to estimate as many parameters as the number

 ISSN: 2102-6238

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, $(x_{i1},...x_{in})$, available to estimate the $Q-1$ linearly independent parameters contained in the vector $(z_{i1},...z_{iQ})$.

We can choose whether the graph is directed or undirected 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 undirected, 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 $i=1,n$).

As defined so far, the model is not identifiable. Daudin et al. [9], proposed 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 in order to confer upon them some reality. This closeness between EHV and some vertices is naturally provided by the maximization of $Tr(ZZ')$.

3.2. Parameter Estimation

The log-likelihood is

$$L=\sum_{i\neq j} x_{ij}\log(\sum_{q,l=1,Q} z_{iq}A_{ql}z_{jl})+(1-x_{ij})\log(1-\sum_{q,l=1,Q} z_{iq}A_{ql}z_{jl}) \tag{2}$$

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. An algorithm is proposed in [9] to maximize the Log-Likelihood under the above constraints and the choice of Q is made using Akaike criterion.

3.3. Comparison between CSBM and SBM

In SBM, the variables Z are random and are equal to 0 or to 1. In CSBM, the variables Z are fixed parameters, and take their values in the simplex S_Q^n. In SBM, each vertex is assumed to pertain to only one group. The mixture model is a mixture of populations of pure vertices. In CSBM, 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:

- 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 CSBM. Note that $E(Z/X=x)$ in the mixture model, and Z in CSBM, take their values in the same set S_Q^n.
- The connectivity matrix A is the key element for the description and interpretation of groups in the two models, see [8] [9]. In SBM, A is the mean connectivity matrix in the sense that

 ISSN: 2102-6238

TABLE 2. *Summary of the models* $X \sim \mathbb{B}(P = f(UAV'))$, *with* $S_Q = \{x, \in [0,1]^Q, \sum_{q=1}^{Q} x_q = 1\}$, $\mathbb{B}(.)$ *is the Bernoulli pdf and* $\mathbb{M}(.)$ *is the multinomial pdf with one trial and* Q *classes*

Model	f	Z	A	P
CSBM	Id	$Z \in S_Q^n$	$A \in [0,1]^{Q^2}$	P=f(ZAZ')
RDPG	$f : \mathbb{R} \to [0,1]$, monotone	$U, V \in \mathbb{R}^{nQ}$	$A = Id$	P=f(UAV')
DEDICOM	Id	$Z \in \mathbb{R}^{nQ}$,$Z'Z = I$	$A \in \mathbb{R}^{Q^2}$	P=f(ZAZ')
MMB	$f(x) = \rho x$, $\rho \in]0,1]$	$Z \in S_Q^n$, $U_{ij} \sim \mathbb{M}(Z_i)$, $V_{ji} \sim \mathbb{M}(Z_j)$	$A \in [0,1]^{Q^2}$	$P_{ij} = f(U'_{ij} A V_{ji})$
OSBM	$f(x) = (1+e^{-x})^{-1}$	$Z \in (0,1)^{nQ}$, $Z_{iq} \sim \mathbb{B}(\alpha_q)$	$A \in \mathbb{R}^{Q^2}$	$P = f(Z'AZ)$

the probability of connection is the weighted mean of the connections between the vertices. In the CSBM, however, A represents an extreme connectivity matrix. As a result A is more contrasted in CSBM than in the mixture model.

3.4. *Other models with continuous latent variables*

Several models have been proposed with a functional form similar to $X \sim \mathbb{B}(ZAZ')$: the Random Dot Product Graphs, DEDICOM, the Mixed Membership Stochastic Blockmodel (MMB) and the Overlapping Stochastic Blockmodel (OSBM). The Table 2 summarizes the functional definition of the different models.

3.4.1. *Random Dot Product Graphs*

The multidimensional scaling (MS) method, applied to the similarity matrix P, 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 MS method is not well suited for modeling P, with two major drawbacks: TT' does not lie in $[0,1]^{n^2}$ if $T \in \mathbb{R}^k$ and TT' is symmetric so it is not suited for the modeling of directed graphs

The Random Dot Product Graph (RDPG) defined in [20] is

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

f is a simple threshold in [20]: $f(x) = 0$ if $x < 0$, $f(x) = x$ if $0 \leq x \leq 1$ and $f(x) = 1$ if $x > 1$.

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, called DEDICOM, was proposed by [15] and well described in [27]. This model uses only one vector for each vertex but inserts a non-symmetric (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 [16]).

 ISSN: 2102-6238

3.4.2. *Mixed Membership Stochastic Blockmodel*

The Mixed Membership Stochastic Blockmodel (MMB, see [1]) is similar to CSBM, with a more complex setting:

- The lines of Z (i.e. the random vectors of weights $Z_i = (z_{i1}...z_{iQ})$) 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 $\rho U'_{i\to j} A V_{i\leftarrow j}$

CSBM 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 CSBM integrates U,V from this structure to obtain $X|A,Z$. Moreover the CSBM model does not need the adhoc sparsity parameter ρ.

3.4.3. *Overlapping Stochastic Blockmodel (OSBM)*

Latouche et al. [18] propose another extension of SBM. They relax the constraints $\sum_q Z_{iq} = 1$ and use a logistic link function between $P(X_{ij} = 1)$ and a quadratic function of Z.

4. Analysis of a large PPIN

MS-Interactome (Ewing et al, [11]), represents the first large-scale study of protein-protein interactions in human cells using a mass spectrometry approach. A total of 6,463 interactions between 2,235 distinct proteins is available. The MS-Interactome includes human protein-protein interactions identified by a combination of immunoprecipitation and high-throughput mass spectrometry. Protein complexes in Human kidney cells were pulled by immuno-precipitation using 338 bait proteins, then identified by LC-ESI-MS/MS. Non specific interactions and false positives were filtered out based on control experiments, quality control parameters and repeat experiments. Bait proteins were chosen based on known functional annotation and implied disease association. About one third of the 338 bait proteins are disease-related ones, and mainly involved in cancer. The complete dataset comprises bait-prey pairs with associated confidence values (complete details are in Ewing et al. 2007). We have analyzed the complete dataset using Mixnet (`http://stat.genopole.cnrs.fr/software/mixnet/`) with the variationnal algorithm described at the bottom of page 3. We present here the results obtained on a subset of the interactions possessing a level of confidence exceeding 0.2 (the scale goes from 0 or NA to 1). This reduced dataset contains 3,494 interactions between 1,561 proteins.

4.1. Number of groups

In the context of mixture model for graphs and using the variational estimation method, Daudin et al. ([8]) propose to use ICL for choosing the number of groups.

$$ICL = \mathscr{J}(x_{[n]}; \tau_{[n]}, \pi, \alpha) - (Q-1)\log n - \frac{Q(Q+1)}{2}\log\left[\frac{n(n-1)}{2}\right]$$

 ISSN: 2102-6238

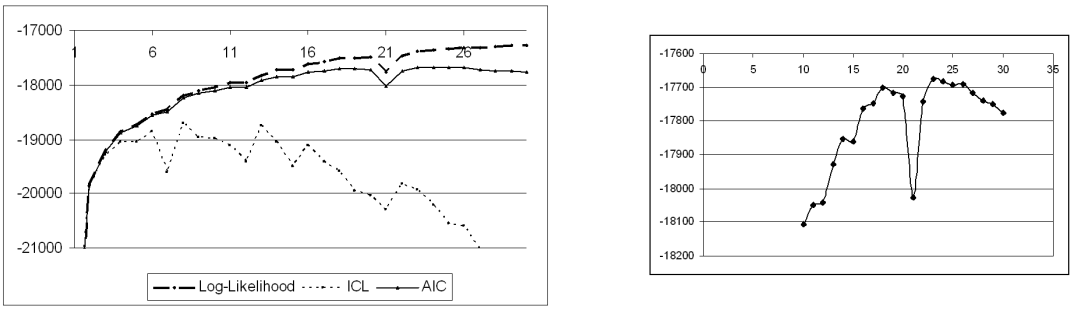


FIGURE 1. *Left side: Log-Likelihood, AIC and ICL for the MS20 PPIN. x-axis : number of groups. Right side: zoom on models* $Q = 10$ *to* $Q = 30$*. y-axis AIC for the MS20 PPIN. x-axis : number of groups*

We have found in practice that ICL has a tendency to underestimate the true number of groups and that AIC give better results for moderate ratios Q/n.

$$AIC = \mathscr{J}(x_{[n]}; \tau_{[n]}, \pi, \alpha) - (Q-1) - \frac{Q(Q+1)}{2}$$

Note that ICL and AIC defined above are approximations of the true corresponding criteria, because the Log-Likelihood is replaced by its variational approximation. However the approximation is precise if n is sufficiently high, because the variationnal approximation of the log-likelihood is asymptotically equivalent to the log-ligelihood, [6]. Figure 1 shows that the best choice using AIC (respectively ICL) is $Q = 23$ (resp. $Q = 8$).

4.2. Results for 18 groups

We have first computed the results for $Q = 1$ to $Q = 20$. The best model using AIC is $Q = 18$, and we have proceed to the complete analysis of each of the 18 groups. Then we examined the values of AIC from $Q = 20$ to $Q = 30$ to see how the curve decreases with Q, and we discovered that AIC was better for $Q = 23$ than for $Q = 18$. However we had no time to reanalyze the 23 new groups for this paper, so we present here the results with $Q = 18$. We have used the *GO term Finder* application from Lewis-Sigler Institute to characterize the groups obtained by Mixnet. The Gene Ontology project (see `http://www.geneontology.org/`) provides an ontology of defined terms representing gene product properties. The ontology covers three domains; cellular component, the parts of a cell or its extracellular environment; molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. We present above the results obtained with the last domain (Biological process). The other two domains have also been tested with some interesting results (data not shown). The P-Values for testing the association between a group and a GO term is obtained by the exact Fisher test which compares the proportion

 ISSN: 2102-6238

TABLE 3. *Description of the 18 groups. The proteins have been affected to one group if their probability of pertaining to the group is greater than 0.5. The* 19^{th} *group contains the unclassified proteins*

group	# proteins	# unrecognized proteins	GO Term	Corrected P-Value
1	44	2	Cellular metabolic Process & Apoptose	4.10^{-7}
2	79	11	RNA Processing	5.10^{-3}
3	12		cell proliferation	8.10^{-3}
4	211	24	intracellular transport	9.10^{-8}
5	55	11	macromolecule localization	1.10^{-4}
6	4		protein targeting and transport	1.10^{-6}
7	353	57	Cellular metabolic Process	5.10^{-12}
8	111	12	macromolecule modification	3.10^{-16}
9	372	73	protein complex assembly	3.10^{-8}
10	96	14	phosphorylation	7.10^{-7}
11	5	2	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	1.10^{-5}
12	15		negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.10^{-38}
13	2		RNA metabolic process	1.10^{-2}
14	8	1	induction of apoptosis by intracellular signals	5.10^{-3}
15	8	1	ribosome biogenesis	1.10^{-3}
16	110	27	translation	4.10^{-25}
17	2		regulation of cellular process	8.10^{-2}
18	19	1	translational elongation	4.10^{-38}
19	55			

of proteins associated to the GO term in one group with the same proportion in the reference set composed of all the annotated proteins of the *goa-human-hgnc* database. We have also computed the exact Fisher test by comparing to another reference set, composed of the 1561 proteins of this study, with similar results (not shown). The P-Values are corrected for multiple testing.

Table 3 shows that each group can be identified by at least one GO term with low corrected P-values excepted for very small groups such as groups 13 and 17 containing only 2 proteins. It is interesting to note that some proteins were not recognized by *GO term Finder*. This means that one can use the results of Mixnet to propose a classification for unknown protein. This possibility concerns a total of 234 proteins. The larger groups have quite general GO terms: for example the group 7, which contains 353 proteins, is characterized by the GO term "Cellular metabolic Process" and group 9, which contains 372 proteins, is characterized by "protein complex assembly". On the opposite small groups are characterized by GO terms which are more precise. For example groups 11 and 12 containing respectively 5 and 15 proteins are characterized by "negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle", and group 17 containing only two proteins is characterized by "regulation of cellular process".

Table 4 shows the connectivity between groups. One can see that some groups are highly connected, such as group 2 with group 17, group 5 with groups 13 or 14, group 6 with groups 10 and 13, and group 15 with group 18. Large groups such as 8 and 9 are loosely connected with other groups.

Figure 2 summarizes the connections between groups using a threshold value of 0.015. One can see that some small groups are connected to many other groups, such as groups 1(Cellular

 ISSN: 2102-6238

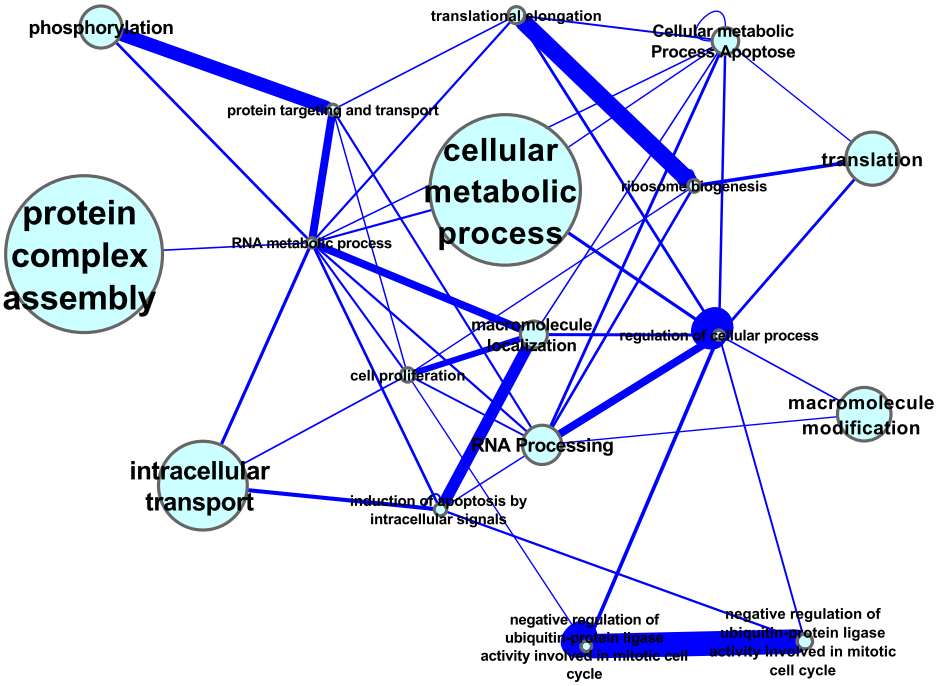


FIGURE 2. *Representation of the 18 groups obtained with Mixnet. Edges between two nodes are present only if the probability of connection between them is greater than 0.015. The size of each node and the size of the police are proportional to the number of proteins contained in it. The width of the edges are proportional to the probability of connection between the corresponding nodes*

 ISSN: 2102-6238

TABLE 4. 100*(Probability of connection between the groups)*

group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	2	7	1	0	2	0	2	1	0	0	0	1	2	1	0	2	6	4
2	7	0	4	0	0	5	0	2	0	0	1	0	5	2	6	0	25	0
3	1	4	0	3	19	2	0	1	1	0	2	0	4	0	2	0	0	0
4	0	0	3	0	0	0	0	0	0	0	0	0	8	11	0	0	1	0
5	2	0	19	0	0	0	0	1	0	0	1	0	24	38	0	0	8	0
6	0	5	2	0	0	0	1	1	0	45	0	0	25	0	0	0	0	3
7	2	0	0	0	0	1	0	0	0	0	1	0	5	0	0	0	8	0
8	1	2	1	0	1	1	0	1	1	0	1	0	0	1	1	0	3	1
9	0	0	1	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0
10	0	0	0	0	0	45	0	0	0	0	0	0	6	0	1	0	0	0
11	0	1	2	0	1	0	1	1	0	0	80	84	0	0	0	0	10	1
12	1	0	0	0	0	0	0	0	0	0	84	0	0	5	0	0	4	0
13	2	5	4	8	24	25	5	0	2	6	0	0	0	6	0	0	0	5
14	1	2	0	11	38	0	0	1	0	0	0	5	6	4	0	1	0	1
15	0	6	2	0	0	0	0	1	0	1	0	0	0	0	3	10	0	58
16	2	0	0	0	0	0	0	0	0	0	0	0	0	1	10	0	7	0
17	6	25	0	1	8	0	8	3	0	0	10	4	0	0	0	7	100	7
18	4	0	0	0	0	3	0	1	0	0	1	0	5	1	58	0	7	0

metabolic Process & Apoptose), 3 (cell proliferation), 13 (RNA metabolic process), 14 (induction of apoptosis by intracellular signals), 15 (ribosome biogenesis) and 17 (regulation of cellular process). On the opposite large groups are less connected.

Interestingly we note that the 17^{th} group is composed of two proteins highly related with tumor progression: the Von Hippel Lindau (VHL) tumor suppression protein and MCC, which blocks cell cycle progression. A similar comment may be made for group 13, composed of two proteins Tgfb1i4 (transforming growth factor beta 1 induced transcript), which is a growth factor, and RNSP1, which is a part of a post-splicing multiprotein complex regulating exons. This is consistent with the fact that about one third of the 338 bait proteins of the dataset are disease-related ones, and mainly involved in cancer.

These results show that it is possible to use a mixture model such as Mixnet to cluster large networks in one pass. This method gives interesting results which deserve to be compared with the ones obtained by the two steps procedures proposed by Marras et al ([22]). We cannot make a more precise comparison because the classification obtained in [22] is not available. A bigger dataset (7385 proteins) described by [22] has also been analyzed with Mixnet (not shown). However this analysis required 7 days.

5. Conclusion

Model-based clustering for networks do not give the same type of results than the communities or cliques research algorithms. For example in table 2.1 the second line networks contains two hubs. A community-research algorithm will give two communities whereas SBM give 4 groups. SBM give more insight into the topology, because a node which is connected to many others ones is basically different from the nodes which are only connected with it. SBM can detect hubs, hierarchical structures, not only communities. Note that these models can also be used to predict the existence of an unknown edge between two nodes.

The algorithms associated with model-based clustering are time consuming. The variational method which is the most efficient cannot deal with more than 5 000 nodes. This may be enough for most of biological and ecological networks and some social networks. However very large networks with more than 100 000 nodes remain a very hard challenge for these algorithms. On-line SBM see [31] is promising and some other algorithms are under study.

The asymptotic behavior of the estimates of SBM and CSBM is under study and some results are given in [6]. The framework is special because we may have an infinite number of parameters (of order n) with n^2 observations. Some results have been obtained with an increasing number of groups, $Q(n)$. The use of covariates about the nodes and the evolution along time are also two challenging problems.

 ISSN: 2102-6238

References

[1] EM. Airoldi, DM. Blei, S. Fienberg, and EP. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[2] E.S. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37:3099–3132, 2010.

[3] E.S. Allman, C. Matias, and J.A. Rhodes. Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference*, doi:10.1016/j.jspi.2010.11.022, 2010.

[4] C. Ambroise and C. Matias. New consistent and asymptotically normal estimators for random graph mixture models. *http://arxiv.org/abs/1003.5165*, 2010.

[5] P.J. Bickel and A. Chen. A nonparametric view of network models and newman-girvan and other modularities. *PNAS*, pages 1–6, 2010.

[6] A. Celisse, J.J. Daudin, and L. Pierre. Consistency of maximum likelihood and variational estimators in mixture models for random graphs. $http://hal.archives-ouvertes.fr/hal-00593644/fr/$, 2011.

[7] D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with growing number of classes. *http://arxiv.org/abs/1011.4644/*, 2010.

[8] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Stat Comput*, 18:173–183, 2008.

[9] J.-J. Daudin, L. Pierre, and C. Vacher. Model for heterogeneous random networks using continuous latent variables and an application to a tree-fungus network. *Biometrics*, 66(4):1043–51, 2010.

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

[11] R. Ewing. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, 3(89):1–17, 2007.

[12] S. Gazal, J.-J. Daudin, and S. Robin. Accuracy of variational estimates for some random graph models. *CSDA*, to appear, 2011.

[13] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99:7821–7826, 2002.

[14] MS. Handcock, AE. Raftery, and JM. Tantrum. Model-based clustering for social networks. *JRSSA*, 54:301–354, 2007.

[15] RA. Harshman. Model for analysis of asymmetrical relationships among n objects or stimuli. *First Joint Meeting of the Psychometric Society and the Society of Mathematical Psychology*, 1978.

[16] HAL. Kiers, JMF. ten Berge, Y. Takane, and J. De Leeuw. A generalization of takane's algorithm for dedicom. *Psychometrika*, pages 151–158, 2002.

[17] P. Latouche, E. Birmele, and C. Ambroise. Bayesian methods for graph clustering. *Advances in Data Analysis, Data Handling and Business Intelligence, Springer*, 2009.

[18] P. Latouche, E. Birmele, and C. Ambroise. Overlapping stochastic block models. *arXiv:0910.2098v2 [stat.ME]*, 2010.

[19] KG. Manton, MA. Woodbury, and HD. Tolley. *Statistical Applications Using Fuzzy Sets*. 1994.

[20] DJ. Marchette and CE. Priebe. Predicting unobserved links in incompletely observed networks. *CSDA*, 52:1373–1386, 2008.

[21] M. Mariadassou, S. Robin, and C. Vacher. Uncovering structure in valued graphs: a variational approach. *Ann. Appl. Statist.*, 4(2):715–42, 2010.

[22] E. Marras, A. Travaglioney, and E. Capobiancoz. Sub-modular resolution analysis by network mixture models. *Statistical Applications in Genetics and Molecular Biology*, 9, 2010.

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

[24] F. Picard, V. Miele, J.-J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using mixnet. *BMC Bioinformatics*, 10, 2009.

[25] K. Rohe, S. Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional sbm. *Technical report Berkeley 791*, http://www.stat.berkeley.edu/25, 2010.

 ISSN: 2102-6238

[26] T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[27] NT. Trendafilov. Gipscal revisited. a projected gradient approach. *Statistics and Computing*, 12:135–145, 2002.

[28] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer Series in Statistics, 1996.

[29] S. van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008.

[30] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[31] H. Zanghi, F. Picard, V. Miele, and C. Ambroise. Strategies for online inference of network mixture. *Annals of Applied Statistics*, to appear, 2010.

 ISSN: 2102-6238