Properties of variational estimates of a mixture model for random graphs

Jean-Jacques Daudin[1] and Alain Celisse[2] and Steven Gazal[1] and Stephane Robin[1]

[1] UMR 518 AgroParisTech/INRA, F-75005, Paris, France
[2] UMR 8524 CNRS – Univ. Lille 1 59655 Villeneuve d'Ascq Cedex - France

Abstract. Mixture models for random graphs have a complex dependency structure and a likelihood which is not computable even for moderate size networks. Variational and variational Bayes techniques are useful approaches for statistical inference of such complex models but their theoretical properties are not well known. We give a result about the consistency of variational estimates of the parameters of the model and we propose variational Bayes estimates. We compare the accuracy of the two variational methods through simulation studies and show an application to a large Protein-Protein interaction network.

Key words: biological network, mixture model, random graph model, variational inference, variational Bayes inference

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 similar connectivity behavior between several vertices leads to their grouping in one *meta-vertex*, without too much information loss. Then, the initial complex network may be reduced to a simpler *meta-network*, with few *meta-vertices* connected by few *meta-edges*. Picard et al. [15] show applications of this idea to biological networks and Nowicki and Snijders [14] to social networks.

In mixture models, discrete latent variables assign each vertex to a group, and each vertex is supposed to pertain to one group. Nowicki and Snijders [14] were among the first to propose what they called a Stochastic block structure model because their model was on the line of an older non stochastic block structure model largely developed in social science. Their estimation method is bayesian MCMC algorithms for networks with less than 200 vertices. Daudin et al. [4] have given more insight on the same model, the degree distribution

and the clustering coefficient, and used a variational method for estimating the parameters.

Bickel and Chen give a general view about the consistency of algorithms maximizing some modularity criteria, such as Newman-Girvan or profile-likelihood modularities [2]. The profile-likelihood is the likelihood of the Stochastic block structure model, with some parameters replaced by their estimates. They used a label switching algorithm to maximize modularity criteria and give some asymptotic results of consistency and speed of convergence. In particular they proved that, using the profile-likelihood, one can recover exactly the class of each node when the number of nodes, n, tends to infinity, a result which was given first by Snijders and Nowicki [17] for two classes.

The variational method allows to deal with several thousand vertices and gives good results in practice ([12]). However the statistical properties of variational estimates are not well known. They maximize a pseudo-likelihood and no general properties have been established. Gunawardana and Byrne [7] show that the variational estimates are consistent only for degenerate cases. However the variational estimates have been proved to be consistent in some cases ([8], [9], [3], [22]) and not consistent in other ones ([21]).

In this paper we give some results about the consistency of variational estimates under three identifiability conditions, and their asymptotic equivalence to the maximum likelihood estimates. Then we build variational Bayes estimates for the same model and compare the two variational methods through simulations. Marras et al. ([11]) propose to analyze large Protein-Protein Interaction networks (PPIN) using two steps: first a deterministic method allows to find large core and community structures and second a stochastic method (such as mixture model) permits to uncover fine-grained interactome components. We show that it is possible to analyze the same large PPIN (6,463 interactions between 2,235 distinct proteins) in one step, using variational algorithm to estimate the parameters of the mixture model, obtaining both large and small clusters from the same model.

2 Model and log-likelihood

2.1 MixNet Model

The definition of the model first proposed by [14], developed by [4] and illustrated by [15] is the following:

Let $i = 1, \ldots, n$ vertices pertaining to $q = 1, \ldots, Q$ classes. Let $X_{ij} = 1$ if there is an edge from node i to node j and $X_{ij} = 0$ if not. X may be symmetric (undirected network) or not (directed network). Let $Z_{[n]} = (Z_1, ...Z_i...Z_n)$, $Z_i \in \{1, \ldots, Q\}$ a sequence of independent random integers with $P(Z_i = q) = \alpha_q$, and $\alpha = (\alpha_1, ..., \alpha_Q)$. In some cases we will use the notation Z_{iq}, with $Z_{iq} = 1$ if $Z_i = q$ and $Z_{iq} = 0$ if $Z_i \neq q$.

The model is the following

Conditionally to Z, X_{ij} are independent Bernoulli random variables with

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

π is the $Q \times Q$ matrix of the parameters π_{ij} and X is the random matrix composed of the X_{ij}.

The parameters of the model are α and π. In the following, this model is denoted MixNet model. Note that the graph of conditional dependency structure of $(Z_1 ... Z_n)/X$ is a clique.

2.2 Log-likelihood

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 $\mathcal{L}(x_{[n]}; \alpha, \pi) = \log \left\{ \sum_{z_{[n]} \in \mathcal{Z}_n} e^{\left[\sum_{i,j \neq i}^{n} b_{ij}(z_i, z_j)\right]} P_{Z_{[n]}}(z_{[n]}) \right\}$ 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 [4]:

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

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 $P(Z_i, Z_i | X)$ is assumed to be a product, and the τ_{iq} can be interpreted as approximations of $P(Z_i|X)$.

The Maximum Likelihood estimates are not computable. The E step of the 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 [4]) is available to obtain the variational estimates, $(\widehat{\tau}, \widehat{\pi}) = \arg\max_{\tau_{[n]}, \pi} \mathcal{J}(x_{[n]}; \tau_{[n]}, \pi, \alpha)$. There is no proof that the algorithm gives a global maximum, but in practice some simulations indicates 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.

3 Properties of the variational estimates

Condition C1
$\forall (q \neq q') \quad \exists l \in (1, Q) \ : \ \pi_{ql} \neq \pi_{q'l}$ or $\pi_{lq} \neq \pi_{lq'}$.
Condition C2
$\exists a > 0 \ : \ \min(\min(\pi_{ql} > 0), \min(1 - \pi_{ql}) > 0) \geq a$.
Condition C3
$\exists b > 0 \ : \ \min(\alpha_q) \geq b$.

Theorem 1. *Assume that C1, C2 and C3 are true. Then the variational estimates of (π, α) are consistent and asymptotically equivalent to the maximum likelihood estimates. Moreover, when $n \to \infty$,*

$\frac{1}{n^2} \left[\mathcal{L}(x_{[n]}; \alpha, \pi) - \mathcal{J}(x_{[n]}; \tau_{[n]}, \pi, \alpha) \right] \xrightarrow{P} 0$ *and* $\widehat{\tau_{[n]}} \xrightarrow{P} z^*$, *with z^* being the true value for Z.*

The proof is tricky and will appear in another paper. It uses the properties of $\mathcal{J}$, concentration inequalities and an extension of classical methods for proving consistency in [20]. There are two mains properties of networks data and the MixNet 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 pdfs in which the variational approximation is searched. These two properties are rarely shared by other data sets and models so the properties of theorem 1 may be quite specific to random networks.

4 Variational Bayes method

The variational approximation can be applied in a Bayesian setting where parameters are viewed as unobserved variables. It turns out to be helpful to handle the two sets of unobserved variables: Z and $\theta = (\alpha, \pi)$. Using conjugate priors, [10] give closed-form approximate conditional distributions of both Z and θ. We show that the same results can be obtained as an application of the general variational Bayes method with exponential families given by [1].

Proposition 1. *If the proportions α and the connexion probabilities $\{\pi_{q\ell}\}$ have all independent prior Dirichlet and Beta distributions*

$$\alpha \sim \mathcal{D}(n^0), \qquad \pi_{q\ell} \sim B(\eta^0_{q\ell}, \zeta^0_{q\ell}),$$

where $n^0 = (n^0_1, \ldots, n^0_Q)$, then the VB approximate conditional mean of Z_{iq} satisfies

$$\tau^{VB}_{iq} \propto e^{\psi(\widetilde{n}_q) - \psi\left(\sum_{l=1}^{Q} \widetilde{n}_\ell\right)} \prod_{j \neq i}^{n} \prod_{l=1}^{Q} e^{\tau^{VB}_{j\ell}\left\{\psi(\widetilde{\zeta}_{q\ell}) - \psi(\widetilde{\eta}_{q\ell} + \widetilde{\zeta}_{q\ell}) + X_{ij}[\psi(\widetilde{\eta}_{q\ell}) - \psi(\widetilde{\zeta}_{q\ell})]\right\})}$$

and the VB approximate posterior distributions of the parameters are

$$(\alpha|X) \approx \mathcal{D}(\widetilde{n}), \qquad (\pi_{q\ell}|X) \approx B(\widetilde{\eta}_{q\ell}, \widetilde{\zeta}_{q\ell})$$

where ψ is the digamma function,

$$\begin{aligned}
\widetilde{n}_q &= n^0_q + \sum_i \tau^{VB}_{iq}, \\
\widetilde{\eta}_{q\ell} &= \eta^0_{q\ell} + \left(1 - \frac{1}{2}\mathbb{1}_{q=l}\right) \sum_{i \neq j} X_{ij} \tau^{VB}_{iq} \tau^{VB}_{j\ell}, \\
\widetilde{\zeta}_{q\ell} &= \zeta^0_{q\ell} + \left(1 - \frac{1}{2}\mathbb{1}_{q=l}\right) \sum_{i \neq j} (1 - X_{ij}) \tau^{VB}_{iq} \tau^{VB}_{j\ell}.
\end{aligned}$$

Proof. We follow the strategy described in [1] which aims at maximizing

$$\mathcal{J}(X) = \log P(X) - KL(\mathcal{R}, P(\theta, Z|X)) \text{ with } \mathcal{R}(Z, \theta) = \mathcal{R}_Z(Z)\mathcal{R}_\theta(\theta) \qquad (1)$$

whith KL the Kullback-Leibler divergence and $\mathcal{R}(Z,\theta)$ the variational approximation of $P(Z,\theta|X)$, omitting the subscript X for $\mathcal{R}$. $\mathcal{R}_Z$ and $\mathcal{R}_\theta$ are approximate conditional distributions given the data X. As we use conjugate priors, $P(\theta)$ is proportional to $\exp[\phi'(\theta)\nu]$ and $P(X,Z|\theta)$ is proportional to $\exp[\phi'(\theta)u(X,Z)]$ where $\phi(\theta)$, ν and $u(X,Z)$ are vectors with dimension $Q+2\times Q(Q+1)/2$ defined as

$$\begin{aligned}
\phi(\theta) &= [\{\log\alpha_q\}_q \quad \{\log\pi_{q\ell}\}_{q\leq l} \qquad\qquad \{\log(1-\pi_{q\ell})\}_{q\leq l}],\\
\nu &= [(n_q-1)_q \quad (\eta_{q\ell}-1)_{q\leq l} \qquad\qquad (\zeta_{q\ell}-1)_{q\leq l}],\\
u(X,Z) &= \left[\{\textstyle\sum_i Z_{iq}\}_q \ \left\{(1-\frac{1}{2}\mathbb{1}_{q=l})\sum_{i\neq j}X_{ij}Z_{iq}Z_{j\ell}\right\}_{q\leq l} \ \left\{(1-\frac{1}{2}\mathbb{1}_{q=l})\sum_{i\neq j}(1-X_{ij})Z_{iq}Z_{j\ell}\right\}_{q\leq l}\right].
\end{aligned}$$

The variational approximate distribution $\mathcal{R}$ is searched in the family of factorized distributions. More precisely, the factorizations $\mathcal{R}(Z,\theta)=\mathcal{R}_Z(Z)\mathcal{R}_\theta(\theta)$, $\mathcal{R}_\theta(\theta)=\mathcal{R}_\alpha(\alpha)\mathcal{R}_\pi(\pi)$ and $\mathcal{R}_Z(Z)=\prod_i\mathcal{R}_{Z_i}(Z_i)$ are assumed. The optimization of (1) under the factorization assumptions leads to

$$\mathcal{R}_\theta(\theta)\propto\exp\{\phi(\theta)'[\nu+\overline{u}(X)]\} \qquad \text{and} \qquad \mathcal{R}_Z(Z)\propto\exp\{\overline{\phi}'u(X,Z)\},$$

where

$$\overline{u}(X)=\sum_z u(X,z)\mathcal{R}_Z(z) \qquad \text{and} \qquad \overline{\phi}=\int\phi(t)\mathcal{R}_\theta(t)\mathrm{d}t$$

and the result follows after standard calculations. ■

5 Simulations

We now present a simulation study to compare the accuracy of the two methods, variational (VEM) and variational Bayes (VB), in terms of parameter inference. We have restricted the study to algorithms which can deal with networks containing more than 200 nodes. Therefore we have not included the MCMC method in our comparison. For small networks, MCMC may give good results. However, the presentation of the results is difficult because the label switching creates some confusion, see [18]. For example the package StOCNET [19] does not give the class of each node but only if two nodes are frequently in the same class. Therefore the results of the MCMC algorithm cannot be compared on the ground of the classification of the nodes. It would be possible to use a relabeling algorithm, but they are numerous, it is difficult to choose between them and the result depends on this choice. Note that the VBEM algorithm selects a particular labeling code, so each node can be classified. This discrepancy between two bayesian algorithms for mixture models is surprising and has been recognized by some authors [16].

Simulation design For computing time reasons, we have made the simulations on small networks with with n nodes, with n from 2 to 50. This choice may seem to be inconsistent with the claim that we want to work on large networks. Actually we use estimation procedures which are able to deal with more than

one thousand of nodes, but we test them on small networks. In practice the important parameter is the number of nodes in each class. The convergence is very good once this number is greater than 20. Therefore we considered a 2-group model with the following parameters.

$$\alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}, \pi = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}.$$

The model involves 4 independent parameters: α_1, π_{11}, π_{12} and π_{22}. We simulated 500 graphs for each graph size.

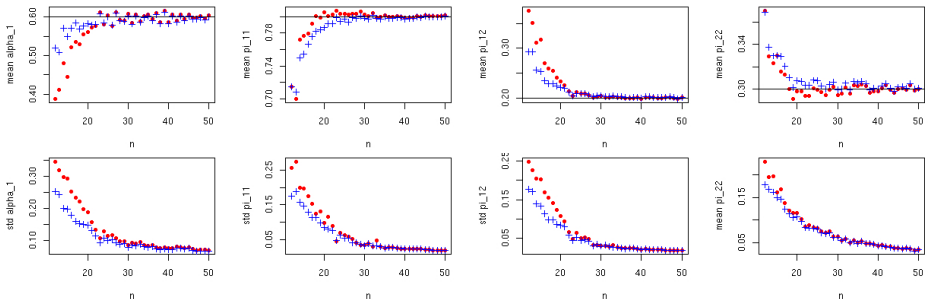

Fig. 1. Mean (top) and standard deviations (bottom) of the estimates. From left to right: α_1, π_{11}, π_{12}, π_{22}. VEM: red circles, VB: blue crosses.

Comparison of the two estimates The results displayed on Figure 1 confirm that VEM provides consistent estimates (see section 3) and that seems to be true also for VB. The consistency of VB has already been proved for some simpler models (see [9] and [3]). In view of our simulations, VB estimates were more robust to extremal samples, resulting in smaller standard deviation than VEM for networks with few nodes. For networks with 25 to 50 nodes there is no real difference between the two methods. A similar study with $Q = 3$ groups was made, from which similar conclusions were drawn (not shown).

VB Credibility intervals The approximation of the posterior distribution provided by VB permits to construct (approximate) credibility intervals. Figure 2 presents the actual credibility of these intervals. The actual credibility was found to be similar to the nominal one as soon as the graph includes $n = 25$ nodes.

Convergence rate of the VB estimates We studied the rate of convergence of the VB estimates. Some parameters of the MixNet model are related to the nodes (i.e. the proportions α_q), whereas some other (i.e. the connexion probabilities $\pi_{q\ell}$) are related to the edges. Information does not accumulate at the same speed for both of them. The ICL model selection criterion developed in [4] involves two distinct penalty terms: one in $\log(n)$ for the proportions and one in $\log[n(n+1)/2]$

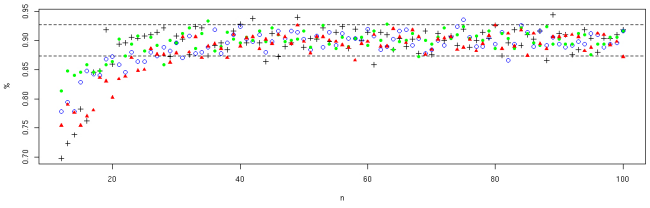

Fig. 2. Proportion of the simulations where interval with credibility 90% contain the true value of the parameter. α_1: black crosses, π_{11}: red triangles, π_{12}: blue circles, π_{22}: green solid circles. Binomial confidence interval: dotted lines.

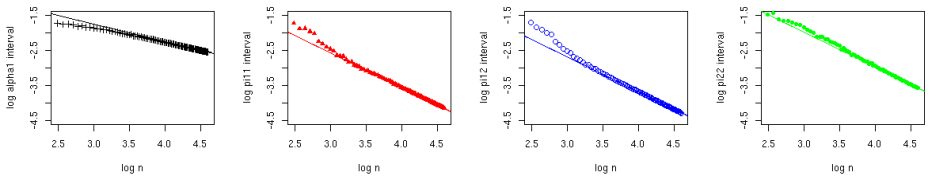

Fig. 3. Width of the 90% credibility as a function of the graph size (in log scale). From left to right: α_1, π_{11}, π_{12} and π_{22}. Straight lines have slope -0.5 α_1 and -1 for the three others.

(for undirected networks) for the connexion probabilities. Figure 3 presents the evolution of the width of the credibility interval as n increases.

Interestingly the empirical convergence rate was found close to $n^{-1/2}$ for the proportion α_1 and close to n^{-1} for the $\pi_{q\ell}$. These rates have been found by [2] in a similar context and are consistent with the penalty terms of the ICL criterion: the relevant sample size for the proportion is the number of nodes, whereas it is the number of edges for the connexion probabilities.

6 Analysis of a large PPIN

MS-Interactome (Ewing et al, [5]), with online materials and protocols available at [13], 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 (IP) and high-throughput mass spectrometry (HTMS). 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, and a summary is reported in the online supplementary file TableMS-Int-Exp-details.doc from [11]). We have analyzed the complete dataset using Mixnet ([12]) with the VEM algorithm. 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.

6.1 Number of groups

In the context of mixture model for graphs and using the variational estimation method, Daudin et al. ([4]) propose to use ICL for choosing the number of groups.

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

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.

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

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, thanks to theorem 1. Figure 4 shows that the best choice using AIC (respectively ICL) is $Q = 23$ (resp. $Q = 8$).

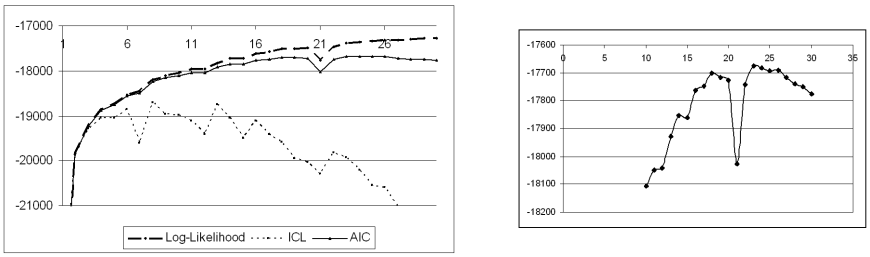

Fig. 4. Left side: Log-Likelihood, AIC and ICL for the MS20 PPIN. x-axis : number of groups. Right side: zoom on models $Q = 10$ to $Q = 30$. y-axis AIC for the MS20 PPIN. x-axis : number of groups

6.2 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 [6]) 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 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 1 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 2 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 5 summarize 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 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 ([11]). We cannot make a more precise comparison because the classification obtained in [11] is not available. A bigger dataset (7385 proteins) described by [11] has also been analyzed with Mixnet (not shown). However this analysis required 7 days.

Table 1. 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			

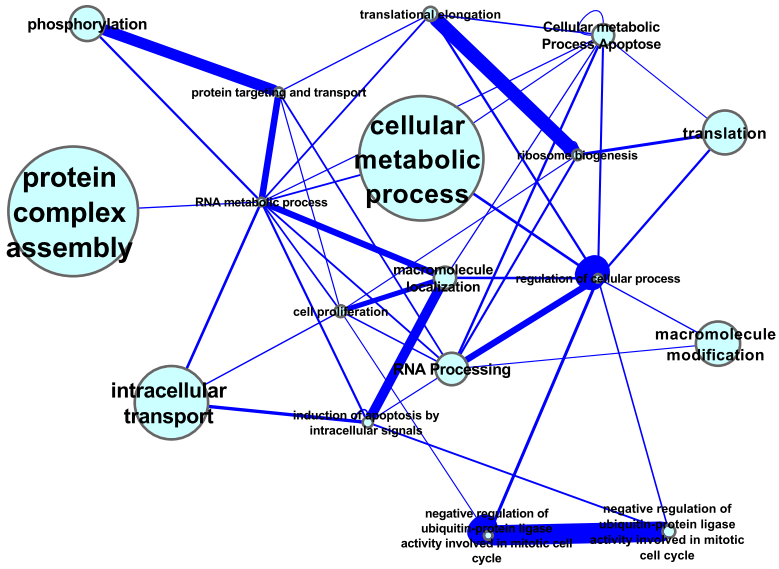

Fig. 5. Representation of the 18 groups obtained with Mixnet. Edges between two nodes are present only if the probability of connection between them is greater than 0.015. The size of each node and the size of the police are proportional to the number of proteins contained in it. The width of the edges are proportional to the probability of connection between the corresponding nodes

Table 2. 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

References

1. Beal, M.J., Ghahramani, Z. : The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. In: Bayesian Statistics 7, Oxford University Press, 543–552. (2000)
2. Bickel, P.J., Chen, A. : A nonparametric view of network models and Newman-Girvan and other modularities. PNAS, 1–6. (2010)
3. Consonni, G., Marin, J.M. : Mean-field variational approximate Bayesian inference for latent variable models. CSDA 52, 790–798. (2007)
4. Daudin, J.J., Picard, F., Robin, S. : A mixture model for random graphs. Stat Comput 18, 173–183 (2008)
5. Ewing R.M., Chu P., Elisma F., Li H., Taylor P., Climie S., McBroom- Cerajewski L., Robinson M.D., OConnor L., Li M., Taylor R., Dharsee M., Ho Y., Heilbut A., Moore L., Zhang S., Ornatsky O., Bukhman Y.V., Ethier M., Sheng Y., Vasilescu J., Abu-Farha M., Lambert J.P., Duewel H.S., Stewart I.I., Kuehl B., Hogue K., Colwill K., Gladwish K., Muskat B., Kinach R., Adams S.L., Moran M.F., Morin G.B., Topaloglou T., and Figeys D. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol. Syst. Biol., 3(89), 1-17. (2007)
6. `http://www.geneontology.org/`
7. Gunawardana, A., Byrne, W. : Convergence Theorems for Generalized Alternating Minimization Procedures. JMLR. (2005)
8. Hall, P., Humphreys, K., Titterington, D.M. : On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. JRSSB 64(3), 549–564. (2002)

9. Humphreys, K., Titterington, D.M. : Approximate Bayesian inference for simple mixtures. In: Proc. Computational Satistics, Physica-Verlag, 331–336. (2000)
10. Latouche, P., Birmele, E., Ambroise, C. : Bayesian Methods for Graph Clustering. SSB Research Report 17 (2008)
11. Marras, E., Travaglioney, A., Capobiancoz, E. : Sub-Modular Resolution Analysis by Network Mixture Models. Statistical Applications in Genetics and Molecular Biology, 9,1,19 (2010)
12. Mixnet, http://stat.genopole.cnrs.fr/software/mixnet/
13. MS-PPIN, http://www.nature.com/msb/journal/v3/n1/full/msb4100134.html
14. Nowicki, K., Snijders, T. : Estimation and prediction for stochastic block-structures. J. Am. Stat. Assoc. 96, 1077–1087 (2001)
15. Picard, F., Miele, F., Daudin, J.J., Cottret, L., Robin, S. : Deciphering the connectivity structure of biological networks using MixNet. BMC Bioinformatics. 10, (2009)
16. eprints.pascal-network.org/archive/00006297/01/ida09mcmcpointcr.pdf
17. Snijders, T.A.B., Nowicki, K. :Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. Journal of Classification 14, 75–100. (1997)
18. Stephens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society, Series B, Methodological 62, 795809.
19. http://stat.gamma.rug.nl/stocnet/
20. van der Vaart, Aad.W., Wellner, J.A. : Weak Convergence and Empirical Processes With Applications to Statistics. Springer Series in Statistics, (1996)
21. Wang, B., Titterington, D.M. : Lack of consistency of mean field and variational Bayes approximations for state space models. Neural Processing Letters 20(3), 151–170. (2004)
22. Woolrich, M.W., Behrens, T.E.: Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Bayesian Analysis 1, 625–650. (2006)