Properties of variational estimates of a mixture model for random graphs

Jean-Jacques Daudin
UMR AgroParisTech-INRA518

ECCS Lisbonne 9/16/2010,

Why do people represent data by a network (1) ?

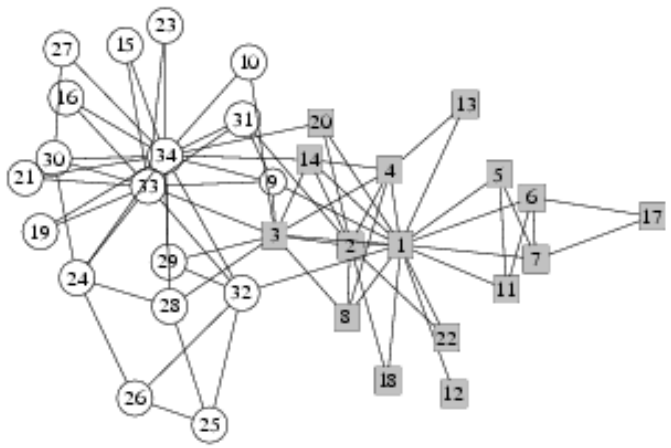


FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).

- Real networks do exist: electric, transport or www networks...They have been represented for a long time by virtual networks.
- Virtual network is a nice way for representing or even "modelling" many scientific phenomenons: social relations, metabolic pathways, chemical reactions...

Why do people represent data by a network (2) ?

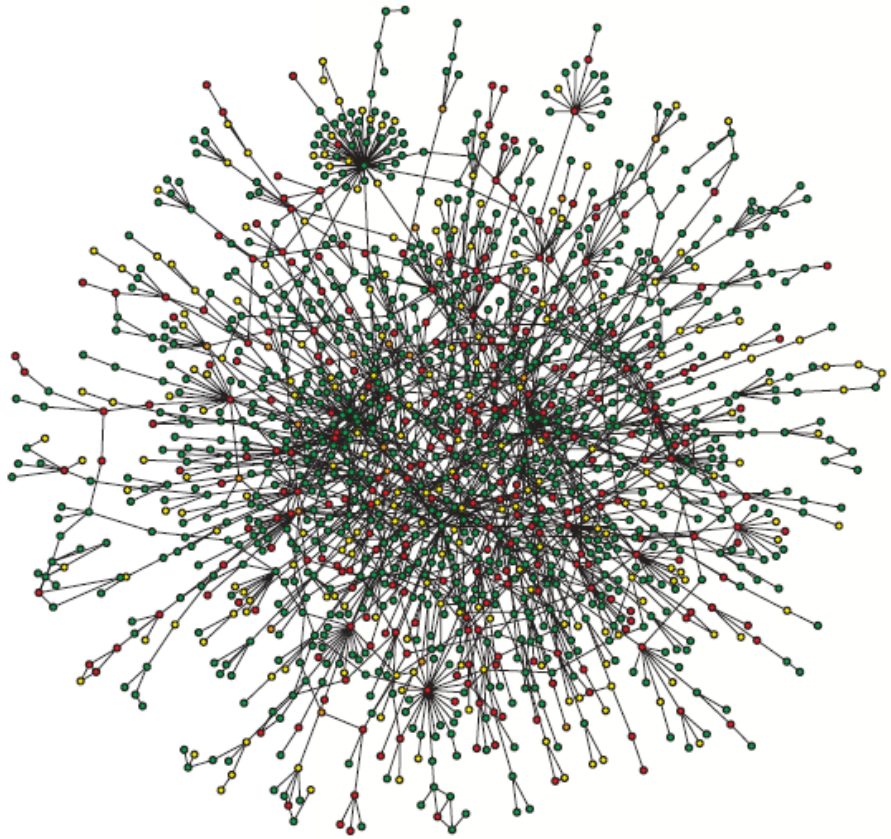


Figure 2 | **Yeast protein interaction network.** A map of protein–protein interactions[18] in *Saccharomyces cerevisiae*, which is based on early yeast two-hybrid measurements[23], illustrates that a few highly connected nodes (which are also known as hubs) hold the network together. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown). Reproduced with permission from REF. 18 © Macmillan Magazines Ltd.

- an overall representation of the interactions between many nodes
- the plot reveals the topology of the networks
- nodes may be colored, adding more information

An unusual data set structure

Usual i.i.d. structure

item	X_1	...	...	X_p
1	x_{11}	...	...	x_{1p}
2	x_{21}	...	...	x_{2p}
.	.	.	.	.
n	x_{n1}	...	...	x_{np}

Structure for relational data

item1	item2	R
1	2	r_{12}
1	3	r_{13}
.	.	.
$n-1$	n	$r_{n-1,n}$

- In the relational data set, the core information is the relation between two items.
- lines are not independent
- the data structure is similar to distance, similarity, covariance or correlation matrices

What do we want?

A simple representation of a complex graph, using meta-vertices and meta-edges.

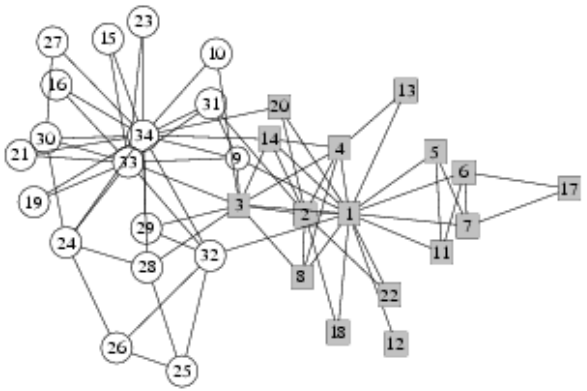
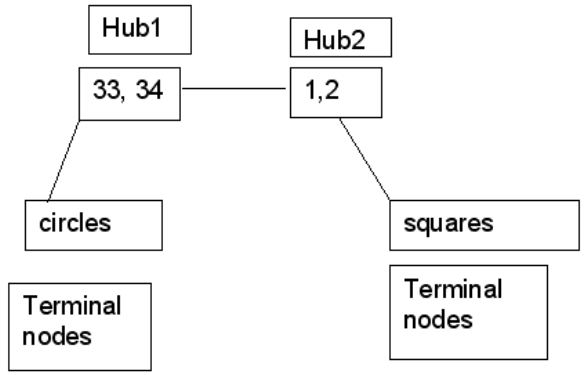


FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).

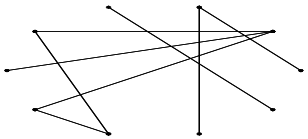
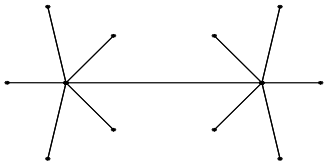
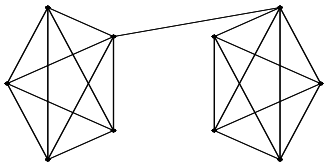
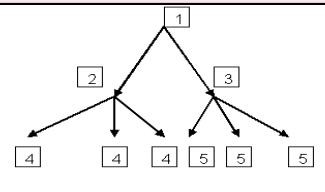


Stochastic Block Model (SBM) a mixture model for random graphs, Snijders and Nowicki (1997), Daudin(2008)

- $i = 1, n$ nodes
- $q = 1, \mathrm{Q}$ classes
- $\mathrm{X}_{ij} = 1$ if there is an edge from node i to node j.
- $\mathrm{Z} = \mathrm{Z}_{iq}$ discrete latent variable, $\mathrm{Z}_{iq} = 1$ if node i pertains to class q
- $(\mathrm{Z}_{i1}, \mathrm{Z}_{i2} ... \mathrm{Z}_{i\mathrm{Q}}) \sim \mathcal{M}(1, \alpha_1, \alpha_2, ... \alpha_{\mathrm{Q}})$
- Conditionally to Z, X_{ij} are independent Bernoulli RV with

$$\mathrm{P}(\mathrm{X}_{ij} = 1 / \mathrm{Z}_i = q, \mathrm{Z}_j = l) = \pi_{ql}$$

SBM: a flexible model

Description	Graph	Q	π
Erdos, no cluster		1	p
Hubs		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$
cluster in "usual sense"		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$
Hierarchical		5	$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

Log-Likelihood is untractable

- **Complete data likelihood**

$$\mathscr{L}(\mathbf{X},\mathbf{Z}) = \sum_i \sum_q Z_{iq} \ln \alpha_q + \sum_{i<j} \sum_{q,l} Z_{iq} Z_{jl} \ln b(\pi_{ql}, X_{ij})$$

where $b(\pi_{ql}, X_{ij}) = \pi_{ql}^{X_{ij}} (1-\pi_{ql})^{(1-X_{ij})}$

- **Observed data likelihood**

$$\mathscr{L}(\mathbf{X}) = \ln \sum_{\mathbf{Z}} \exp \mathscr{L}(\mathbf{X},\mathbf{Z})$$

- observed data likelihood requires a sum over Q^n terms : untractable
- EM-like strategies require $\Pr(\mathbf{Z}|\mathbf{X})$: untractable (no conditional independence).

Variational Inference

Main Idea: Replace complicated $\Pr(\mathbf{Z}|\mathbf{X})$ by a simple $\mathscr{R}_{\mathbf{X}}[\mathbf{Z}]$ such that $\mathrm{KL}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X}))$ is minimal.

- Optimize in $\mathscr{R}_{\mathbf{X}}$ the function $\mathscr{J}(\mathscr{R}_{\mathbf{X}})$ given by :

$$\begin{aligned} \mathscr{J}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}]) &= \mathscr{L}(\mathbf{X}) - \mathrm{KL}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}], \Pr(\mathbf{Z}|\mathbf{X})) \\ &= \mathscr{H}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}]) - \sum_{\mathbf{Z}} \mathscr{R}_{\mathbf{X}}[\mathbf{Z}] \mathscr{L}(\mathbf{X}, \mathbf{Z}) \end{aligned}$$

- For simple $\mathscr{R}_{\mathbf{X}}$, $\mathscr{J}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}])$ is tractable,
- At best, $\mathscr{R}_{\mathbf{X}} = \Pr(\mathbf{Z}|\mathbf{X})$ and $\mathscr{J}(\mathscr{R}_{\mathbf{X}}[\mathbf{Z}]) = \mathscr{L}(\mathbf{X})$.

2 Steps Iterative Algorithm, (Daudin et al., 2008)

- Step 1 **Optimize** $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ **w.r.t.** $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}]$:
  - → Restriction to a "comfortable" class of functions,
  - → $\mathcal{R}_{\mathbf{X}}[\mathbf{Z}] = \prod_i h(\mathbf{Z}_i; \boldsymbol{\tau}_i)$, with $h(.; \boldsymbol{\tau}_i)$ the multinomial distribution,
  - → τ_{iq} is a variational parameter to be optimized using a fixed point algorithm:

$$\tilde{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{l=1}^{Q} b(\pi_{ql}, X_{ij})^{\tilde{\tau}_{jl}}$$

- Step 2 **Optimize** $\mathcal{J}(\mathcal{R}_{\mathbf{X}}[\mathbf{Z}])$ **w.r.t.** $(\boldsymbol{\alpha}, \boldsymbol{\pi})$:
  - → Constraint: $\sum_q \alpha_q = 1$

$$\begin{aligned} \tilde{\alpha}_q &= \sum_i \tilde{\tau}_{iq}/n \\ \tilde{\pi}_{ql} &= \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl}X_{ij} / \sum_{ij} \tilde{\tau}_{iq}\tilde{\tau}_{jl} \end{aligned}$$

Toy-Example: Karate Club

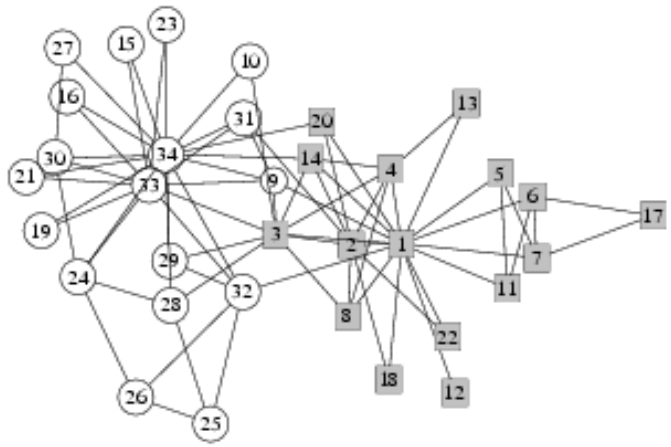
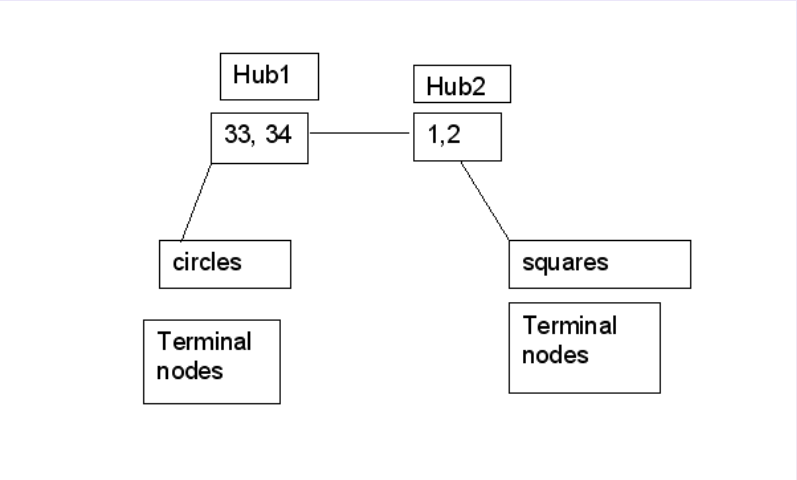


FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).

- nodes are members of the club
- edges between 2 members if they have social relation outside the club
- known properties: the club has split away in two parts (cercles and squares).

Data from W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977).

SBM results for Karate Club



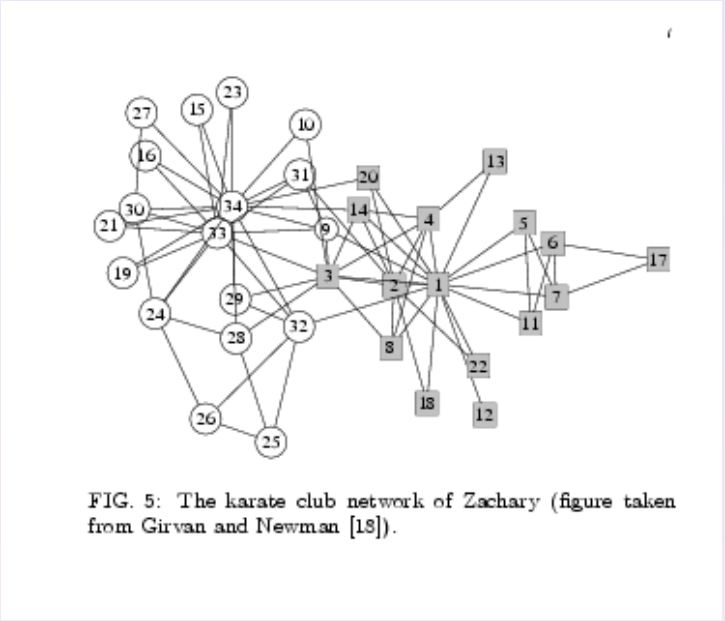


FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).

	SBM Classes			
	1	2	3	4
1	100	53	16	16
2	53	12	0	7
3	16	0	8	73
4	16	7	73	100
$n * \alpha$	3	13	16	2

The split is recovered and the role of the leaders is underlined.

Known results about the consistency of variational estimates

- The variational method is practically effective with many applications till $n = 3000$ (see Picard et al. and `http://stat.genopole.cnrs.fr/software/mixnet/`).
- But till now no theoretical property is known

No general result about variational estimates: they maximize a pseudo-likelihood and no general properties have been established.

- Gunawardana and Byrne show that the variational estimates are consistent only for degenerate cases,
- VE have been proved to be consistent in some cases, Markovian models (Hall et al.), latent variable models (Consonni et al. and normal mixture model (Woolrich et al.)
- and not consistent for state space models (Wang et al.)

Consistency of variational estimates

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

Under C1, C2 and C3, the variational estimates $(\widehat{\pi}, \widehat{\alpha})$ are consistent and asymptotically equivalent to the maximum likelihood estimates. Moreover, when $n \to \infty$,

$$\frac{1}{n^2} \left[\mathscr{L}(x_{[n]}; \alpha, \pi) - \mathscr{J}(x_{[n]}; \tau_{[n]}, \pi, \alpha) \right] \xrightarrow{P} 0$$

and

$$\widehat{\tau_{[n]}} \xrightarrow{P} z^*,$$

with z^* being the true value for Z.

Idea of proof

The proof uses

- the properties of $\mathscr{J}$,
- concentration inequalities
- an extension of classical methods using Empirical Processes for proving consistency.

Two properties of networks data and the model are important in this proof:

- there are n^2 data $\rightarrow$ strong concentration inequalities
- the asymptotic pdf of Z|X pertains to the factorized class of pdfs in which the variational approximation is searched.

These properties are rarely shared by other data sets and models so the proof is specific to random networks.

Variational Bayes(1)

- Bayesian setting: parameters are viewed as unobserved variables
- two sets of unobserved variables: Z and $\theta = (\alpha, \pi)$.
- Conjugate priors, Latouche et al. give closed-form approximate conditional distributions of both Z and θ.
- same results can be obtained as an application of the general variational Bayes method with exponential families given by Beal and Ghahramani.

Variational Bayes(2)

Prior $\alpha \sim \mathscr{D}(n^0)$, $\pi_{q\ell} \sim B(\eta^0_{q\ell}, \zeta^0_{q\ell})$, $n^0 = (n^0_1, \ldots, n^0_Q)$ $\mathscr{D}$: Dirichlet, B: Beta.

Approximate conditional mean of Z_{iq}

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

ψ digamma function.

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

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

Simulation design

- $n = 2, 4, \ldots 50$
- $\alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}$, $\pi = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}$.
- 500 graphs for each graph size.

Consistency and precision of VEM and VB

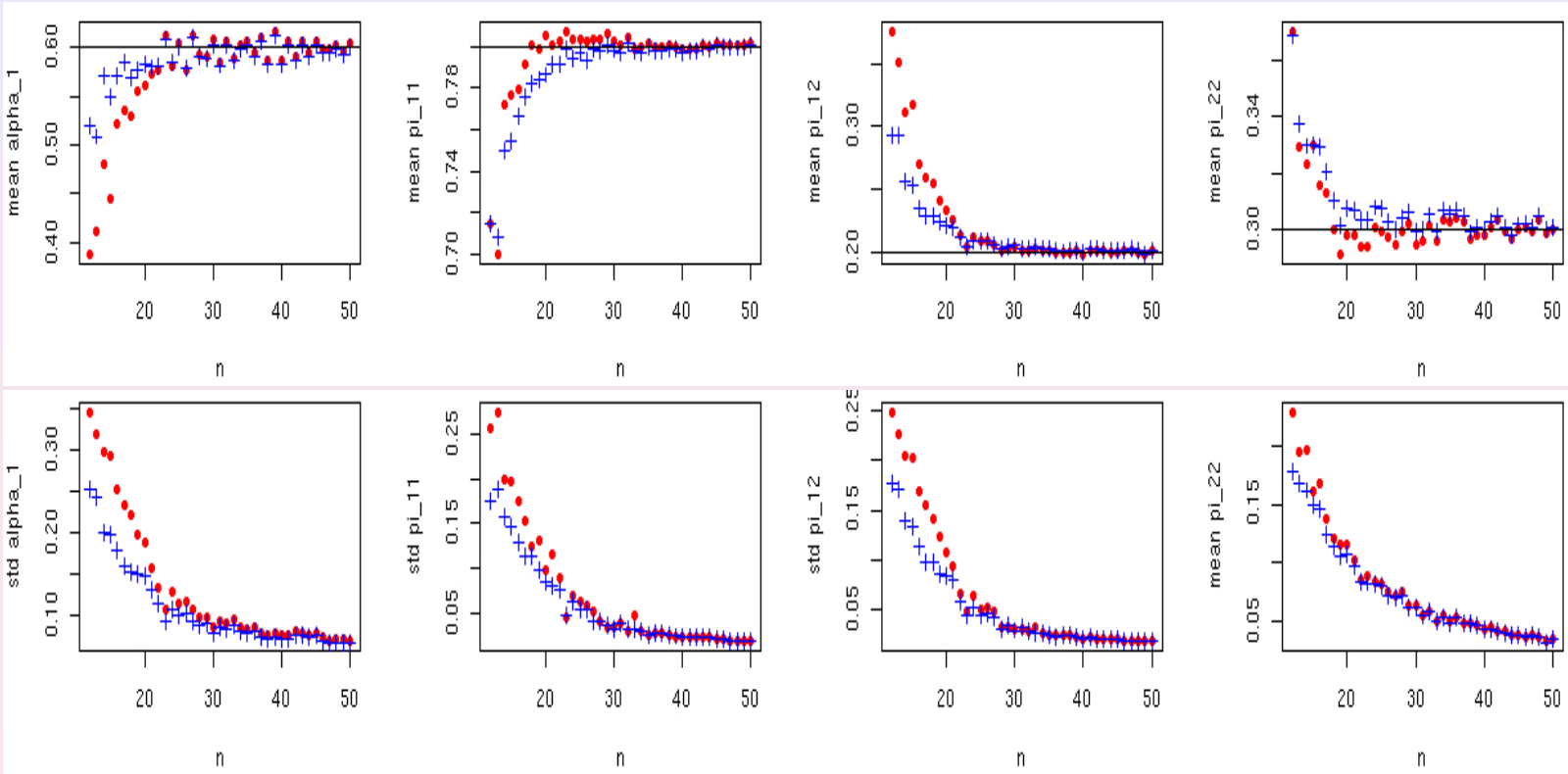


Figure: Mean (top) and standard deviations (bottom) of the estimates. From left to right: α_1, π_{11}, π_{12}, π_{22}. VEM: red circles, VB: blue crosses.

Credibility intervals

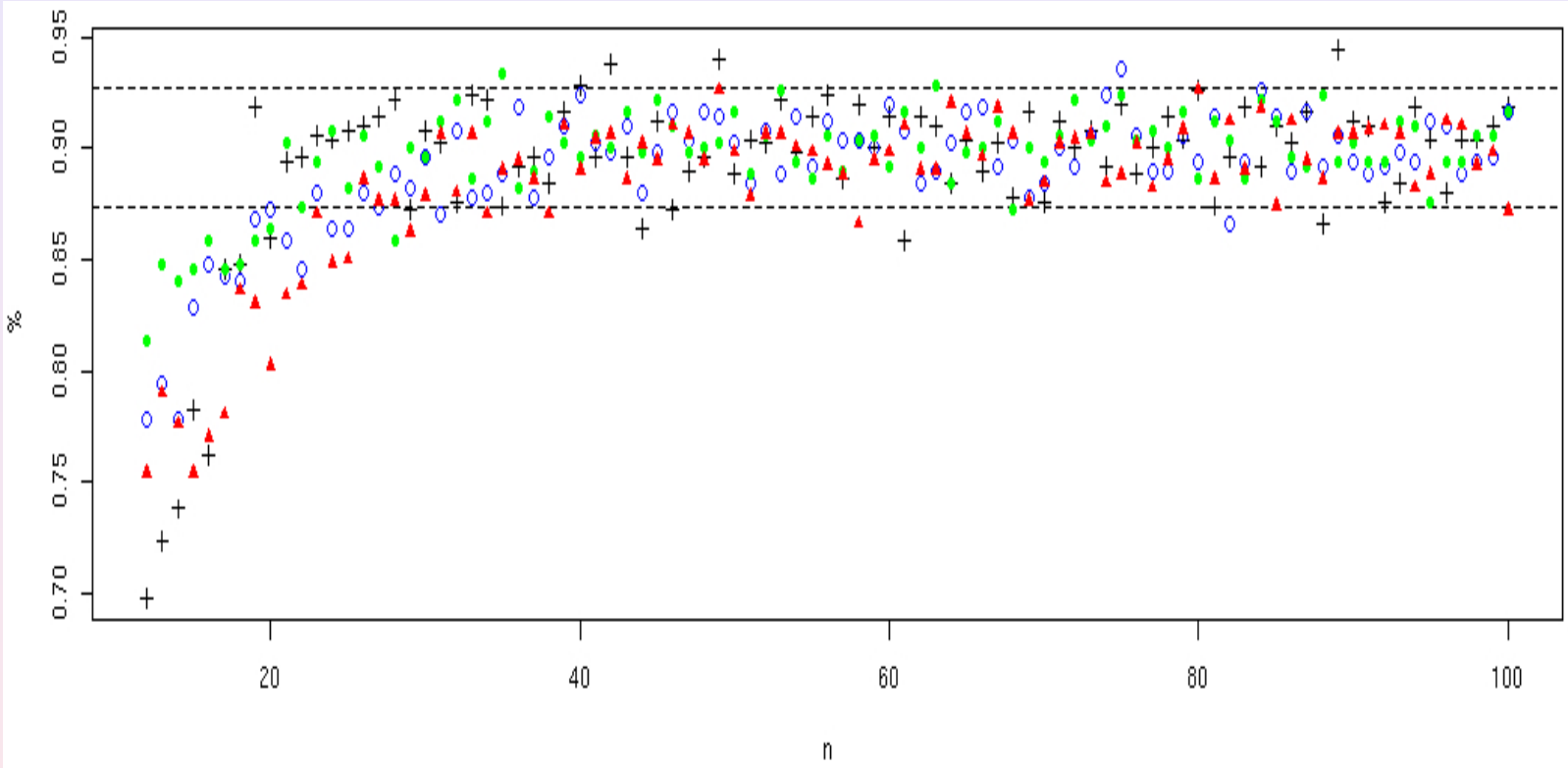


Figure: Proportion of the simulations where interval with credibility 90% contain the true value of the parameter. α_1: black crosses, π_{11}: red triangles, π_{12}: blue circles, π_{22}: green solid circles. Binomial confidence interval: dotted lines.

Speed of convergence

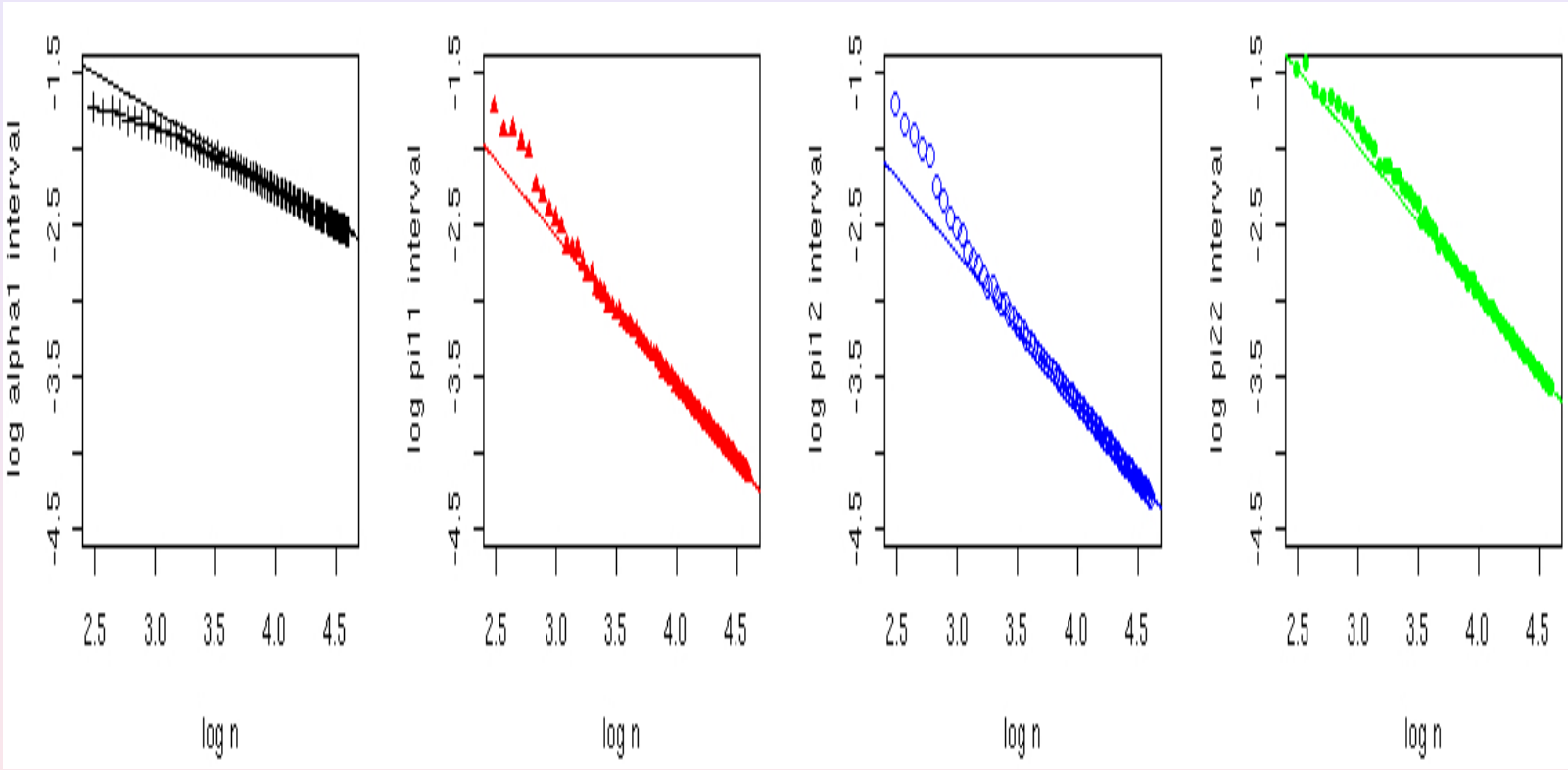


Figure: Width of the 90% credibility as a function of the graph size (in log scale). From left to right: α_1, π_{11}, π_{12} and π_{22}. Straight lines have slope -0.5 for α_1 and -1 for the three others.

MS-Interactome data

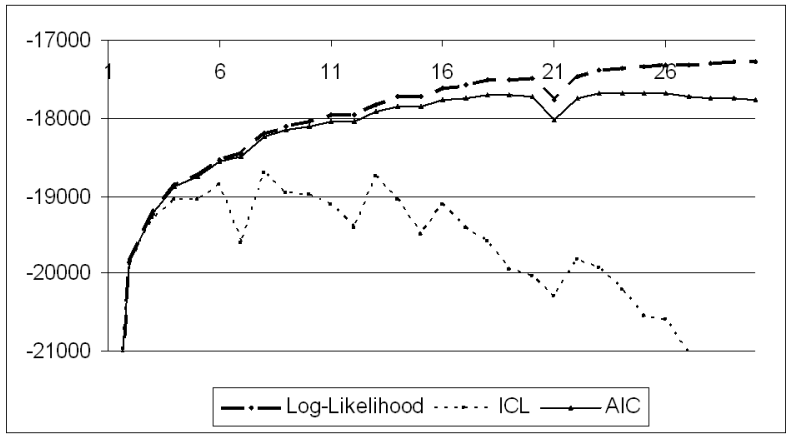
- MS-Interactome (Ewing et al.): first large-scale study of protein-protein interactions in human cells using a mass spectrometry approach.
- 3,494 interactions between 1,561 proteins
- Bait proteins chosen based on known functional annotation and implied disease association.
- One third of the 338 bait proteins are disease-related ones, mainly involved in cancer
- Data previously analyzed by Marras et al. using a two-steps procedure: 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.
- The following analysis is made using VEM method using package Mixnet.

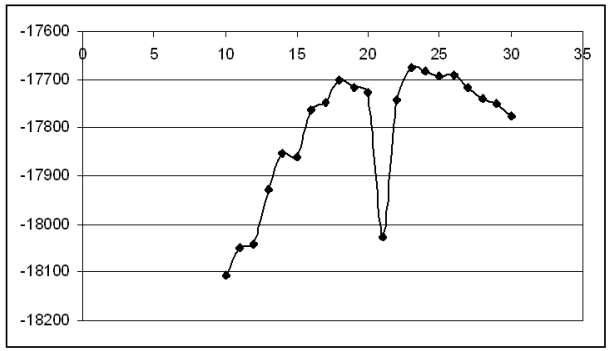
Number of groups

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

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

Best choices: $Q = 23$ (AIC) and $Q = 8$ (ICL).





GO-Characteristics of the groups

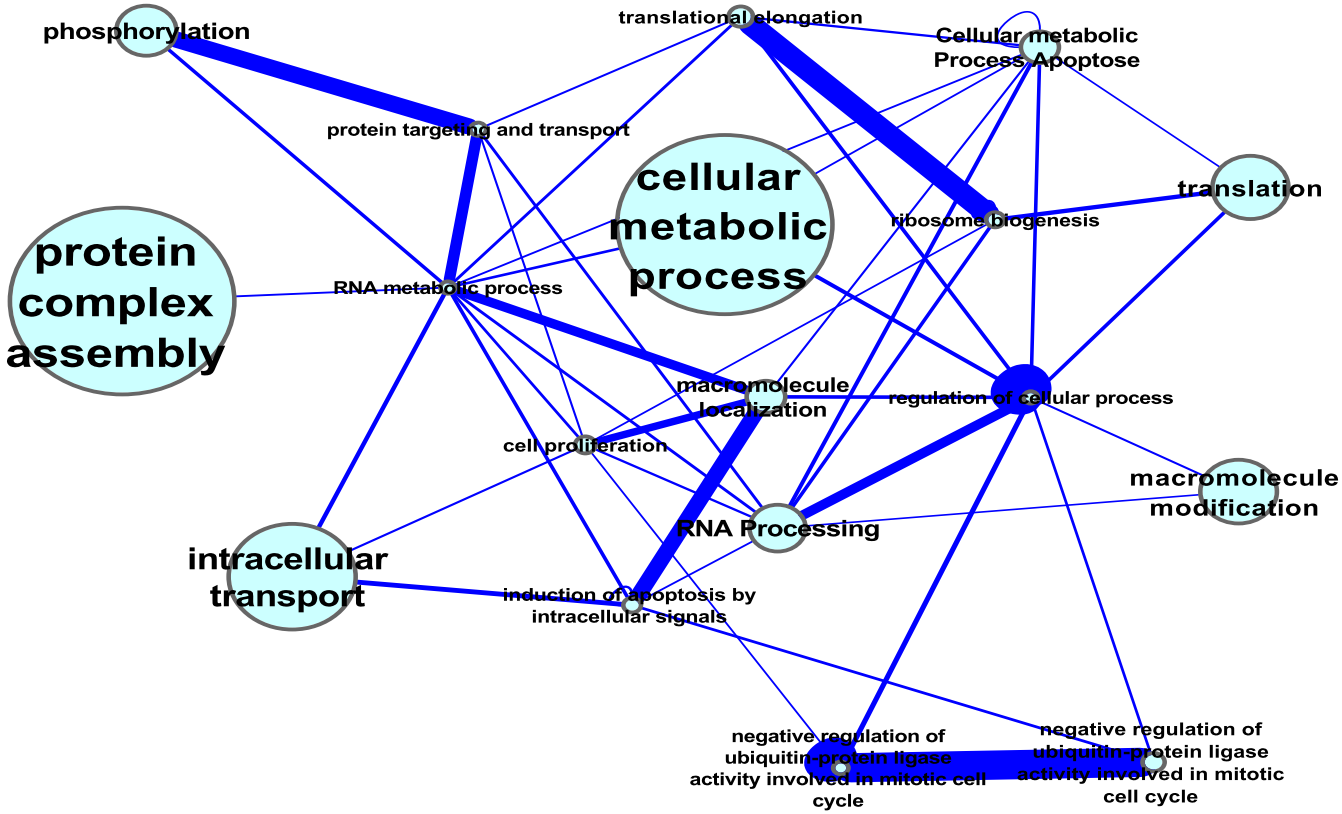
Description of the first groups. The proteins have been affected to one group if their probability of pertaining to the group is greater than 0.5.

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}

More about the groups

- most of the groups can be identified by at least one GO term with low corrected P-values
- 234 proteins were not recognized by *GO term Finder* → SBM proposes a classification for unknown proteins.
- 17^{th} group composed of two proteins highly related with tumor progression: the Von Hippel Lindau (VHL) tumor suppression protein and MCC, which blocks cell cycle progression.
- 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.

Meta-Network obtained with SBM



Conclusions

- SBM is a flexible model which allows to replace a complicated network by a simple meta-network,
- VEM and VB are theoretically validated,
- It is possible to use SBM to cluster large networks,
- Freely available package, URL: http://pbil.univ-lyon1.fr/software/MixNet.

Common work with

Alain Celisse		UMR CNRS 8524,Univ. Lille 1	Consistency of variationnal Es
Steven Gazal		INSERM, Paris	Simulations
Stéphane Robin		UMR 518 AgroParisTech/INRA	Variationnal Bayes

SBM

- Daudin, J.J., Picard, F., Robin, S. : A mixture model for random graphs. Stat Comput 18, 173–183 (2008)
- Latouche, P., Birmele, E., Ambroise, C. : Bayesian Methods for Graph Clustering. SSB Research Report 17 (2008)
- Mixnet, `http://stat.genopole.cnrs.fr/software/mixnet/`
- Nowicki, K., Snijders, T. : Estimation and prediction for stochastic block-structures. J. Am. Stat. Assoc. 96, 1077–1087 (2001)
- Picard, F., Miele, F., Daudin, J.J., Cottret, L., Robin, S. : Deciphering the connectivity structure of biological networks using MixNet. BMC Bioinformatics. 10, (2009)
- 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)

Variational method

- Beal, M.J., Ghahramani, Z. : The Variational Bayesian EM Algorithm for Incomplete Data, In: Bayesian Statistics 7, Oxford University Press, (2000)
- Consonni, G., Marin, J.M. : Mean-field variational approximate Bayesian inference for latent variable models. CSDA 52, (2007)
- Gunawardana, A., Byrne, W. : Convergence Theorems for Generalized Alternating Minimization Procedures. JMLR. (2005)
- 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), (2002)
- Wang, B., Titterington, D.M. : Lack of consistency of mean field and variational Bayes approximations for state space models. Neural Processing Letters 20(3), (2004)
- 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, (2006)

PPIN Example

- Ewing R.M. et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol. Syst. Biol., 3(89), 1-17. (2007)
- 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)