

Uncovering overlapping clusters in biological networks

Pierre Latouche, Etienne Birmelé, Christophe Ambroise

Laboratoire Statistique et Génome, UMR CNRS 8071-INRA 1152-UEVE
La Genopole, Tour Evry 2, 523 place des Terrasses 91000 Evry France
pierre.latouche@genopole.cnrs.fr

Abstract: *In the last few years, there has been a growing interest in studying biological networks. Many deterministic and probabilistic clustering methods have been developed. They aim at learning information from the presence or absence of links between pairs of vertices (genes or proteins). Given a network, almost all these techniques partition the vertices into disjoint clusters, according to their connection profile. However, recent studies have shown that these methods were too restrictive and that most of the existing biological networks contained overlapping clusters. To tackle this issue, we present in this paper a latent logistic model, that allows each vertex to belong to multiple clusters, as well as an efficient approximate inference procedure based on global and local variational techniques. We show the results that we obtained on a transcriptional regulatory network of yeast.*

Keywords: Biological networks, clustering methods, overlapping clusters, global and local variational approaches.

1 Introduction

Networks are used in many scientific fields such as biology, social science, and information technology. In this context, a lot of attention has been paid on developing models to learn knowledge from the presence or absence of links between pairs of objects. Both deterministic and probabilistic strategies have been proposed. Among these techniques, random graph models [?,?], based on mixture distributions, have recently received a growing interest. In particular, they have been shown capable of characterizing the complex topology of real networks, that is, a majority of vertices with none or very few links and the presence of hubs which make networks locally dense. A drawback of such methods is that they all partition the vertices into disjoint clusters, while lots of objects in real world applications typically belong to multiple groups or communities. For instance, many genes are known to participate in several functional categories, and actors might belong to several groups of interests. Thus, a graph clustering method should be able to uncover overlapping clusters. This issue has received growing attention in the last few years, starting with an algorithmic approach based on small complete sub-graphs developed by Palla and al. [?]. More recent works [?] proposed a mixed membership approach. In this paper, we present a new mixture model [?] for generating networks, depending on $(Q+1)^2+Q$ parameters, where Q is the number of components in the mixture. A latent $\{0, 1\}$ -vector of length Q is assigned to each vertex, drawn from products of Bernoulli distributions whose parameters are not vertex-dependent. Each vertex may then belong to several components, allowing overlapping clusters, and each edge probability depends only on the components of its endpoints.

2 Model and Notations

We consider a directed binary random graph \mathcal{G} , where V denotes a set of N fixed vertices and $\mathbf{X} = \{X_{ij}, (i, j) \in V^2\}$ is the set of all the random edges. We assume that \mathcal{G} does not have any self loop, and therefore, the variables X_{ii} will not be taken into account.

For each vertex $i \in V$, we introduce a latent vector \mathbf{Z}_i , of Q independent Boolean variables $Z_{iq} \in \{0, 1\}$, drawn from Bernoulli distributions:

$$\mathbf{Z}_i \sim \prod_{q=1}^Q \mathcal{B}(Z_{iq}; \alpha_q) = \prod_{q=1}^Q \alpha_q^{Z_{iq}} (1 - \alpha_q)^{1-Z_{iq}}, \quad (1)$$

and we denote $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_Q\}$ the vector of class probabilities. Note that in the case of a usual mixture model, \mathbf{Z}_i would be generated according to a multinomial distribution with parameters $(1, \boldsymbol{\alpha})$. Therefore, the vector \mathbf{Z}_i would see all its components set to zero except one such that $Z_{iq} = 1$ if vertex i belongs to class q . The model would then verify $\sum_{q=1}^Q Z_{iq} = \sum_{q=1}^Q \alpha_q = 1, \forall i$. In this paper, we relax these constraints using the product of Bernoulli distributions (1), allowing each vertex to belong to multiple classes. We point out that \mathbf{Z}_i can also have all its components set to zero.

Given two latent vectors \mathbf{Z}_i and \mathbf{Z}_j , we assume that the edge X_{ij} is drawn from a Bernoulli distribution:

$$X_{ij} | \mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}(X_{ij}; g(a_{\mathbf{z}_i, \mathbf{z}_j})) = e^{X_{ij} a_{\mathbf{z}_i, \mathbf{z}_j}} g(-a_{\mathbf{z}_i, \mathbf{z}_j}),$$

where

$$a_{\mathbf{z}_i, \mathbf{z}_j} = \mathbf{Z}_i^\top \mathbf{W} \mathbf{Z}_j + \mathbf{Z}_i^\top \mathbf{U} + \mathbf{V}^\top \mathbf{Z}_j + W^*, \quad (2)$$

and $g(x) = (1 + e^{-x})^{-1}$ is the logistic sigmoid function. \mathbf{W} is a $Q \times Q$ matrix whereas \mathbf{U} and \mathbf{V} are Q -dimensional vectors. The first term in (2) describes the interactions between the vertices i and j . If i belongs only to class q and j only to class l , then only one interaction term remains ($\mathbf{Z}_i^\top \mathbf{W} \mathbf{Z}_j = W_{ql}$). However, the interactions can become much more complex if one or both of these two vertices belong to multiple classes. Note that the second term in (2) does not depend on \mathbf{Z}_j . It models the overall capacity of vertex i to connect to other vertices. By symmetry, the third term represents the global tendency of vertex j to receive an edge. Finally, we use W^* as a bias, to model sparsity.

3 Variational Approximations

The log-likelihood of the observed data set is defined through the marginalization: $p(\mathbf{X} | \boldsymbol{\alpha}, \tilde{\mathbf{W}}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\alpha}, \tilde{\mathbf{W}})$. This summation involves 2^{NQ} terms and quickly becomes intractable. To tackle this issue, the Expectation-Maximization (EM) algorithm has been applied on many mixture models. However, the E-step requires the calculation of the posterior distribution $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ which can not be factorized in the case of networks. In order to obtain a tractable procedure, we propose some approximations based on global and local variational techniques.

3.1 The q -transformation (Variational EM)

Given a distribution $q(\mathbf{Z})$, the log-likelihood of the observed data set can be decomposed using the Kullback-Leibler divergence (KL):

$$\ln p(\mathbf{X} | \boldsymbol{\alpha}, \tilde{\mathbf{W}}) = \mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}) + \text{KL}(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})). \quad (3)$$

By definition $\text{KL}(\cdot \| \cdot)$ is always positive and therefore \mathcal{L} is a lower bound of the log-likelihood:

$$\ln p(\mathbf{X} | \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \geq \mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}), \forall q(\mathbf{Z}). \quad (4)$$

The maximum $\ln p(\mathbf{X} | \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ of \mathcal{L} is reached when $q(\mathbf{Z}) = p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$. Thus, if the posterior distribution $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ was tractable, the optimizations of \mathcal{L} and $\ln p(\mathbf{X} | \boldsymbol{\alpha}, \tilde{\mathbf{W}})$, with respect to $\boldsymbol{\alpha}$ and $\tilde{\mathbf{W}}$, would be equivalent. However, in the case of networks, $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})$ can not be calculated and \mathcal{L} can not be optimized over the entire space of $q(\mathbf{Z})$ distributions. Thus, we restrict our search to the class of distributions which satisfy:

$$q(\mathbf{Z}) = \prod_{i=1}^N q(\mathbf{Z}_i) = \prod_{i=1}^N \prod_{q=1}^Q \mathcal{B}(Z_{iq}; \tau_{iq}). \quad (5)$$

Each τ_{iq} is a variational parameter and corresponds to the posterior probability of node i to belong to class q . Note that we do not constrain the vectors $\boldsymbol{\tau}_i = \{\tau_{i1}, \dots, \tau_{iQ}\}$ to lay on the $Q - 1$ dimensional simplex, and thereby, each node can belong to multiple clusters.

The decomposition (3) and the factorization (5) lead to a variational EM algorithm. During the E-step, the parameters $\boldsymbol{\alpha}$ and $\tilde{\mathbf{W}}$ are fixed; and by optimizing the lower bound with respect to the τ_{iq} s, the algorithm looks for the best approximation of the posterior distribution. Then, during the M-step, $q(\mathbf{Z})$ is used to optimize the lower bound and to find new estimates of $\boldsymbol{\alpha}$ and $\tilde{\mathbf{W}}$.

3.2 The ξ -transformation

The lower bound of (3) is given by

$$\mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}) = \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X} | \mathbf{Z}, \tilde{\mathbf{W}})] + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z} | \boldsymbol{\alpha})] - \mathbb{E}_{\mathbf{Z}}[\ln q(\mathbf{Z})], \quad (6)$$

where the expectations are calculated according to the distribution $q(\mathbf{Z})$. The first term of (6) is given by:

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\alpha}, \tilde{\mathbf{W}})] = \sum_{i \neq j}^N \left\{ X_{ij} \tilde{\boldsymbol{\tau}}_i^T \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})] \right\}. \quad (7)$$

Unfortunately, since the logistic sigmoid function is non linear, $\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$ can not be computed analytically. Thus, we need a second level of approximation to carry out the variational E and M steps described previously (Sect. 3.1).

We use the bound $\ln g(x, \xi)$ on the log-logistic sigmoid function introduced by [?]. When applied on $\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$, it leads to:

$$\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})] \geq \ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j}, \xi_{ij}) = \ln g(\xi_{ij}) - \frac{(\tilde{\boldsymbol{\tau}}_i^T \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \xi_{ij})}{2} - \lambda(\xi_{ij}) \left(\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[(\tilde{\mathbf{Z}}_i^T \tilde{\mathbf{W}} \tilde{\mathbf{Z}}_j)^2] - \xi_{ij}^2 \right), \quad (8)$$

where $\lambda(\xi) = \frac{1}{4\xi} \tanh\left(\frac{\xi}{2}\right) = \frac{1}{2\xi} \left\{ g(\xi) - \frac{1}{2} \right\}$. Thus, for each edge (i, j) in the graph, we have introduced a lower bound which depends on a variational parameter ξ_{ij} . By optimizing each function $\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j}, \xi_{ij})$ with respect to ξ_{ij} , we obtain the tightest bounds to the functions $\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j}[\ln g(-a_{\mathbf{Z}_i, \mathbf{Z}_j})]$. These bounds are then used during the variational E and M steps to optimize an approximation of \mathcal{L} defined in (6).

4 Experiments

We consider the yeast transcriptional regulatory network described in [?] and we focus on a subset of 192 vertices connected by 303 edges. Nodes of the network correspond to operons, and two operons are linked if one operon encodes a transcriptional factor that directly regulates the other operon. Such networks are known to be relatively sparse which makes them hard to analyze. In this Section, we aim at clustering the vertices according to their connection profile. Using $Q = 6$ clusters, we apply our algorithm and we obtain the results in Table 1.

cluster	size	operons
1	2	STE12 TEC1
2	33	YBR070C MID2 YEL033W SRD1 TSL1 RTS2 PRM5 YNL051W PST1 YJL142C SSA4 YGR149W SPO12 YNL159C SFP1 YHR156C YPS1 YPL114W HTB2 MPT5 SRL1 DHH1 TKL2 PGU1 YHL021C RTA1 WSC2 GAT4 YJL017W TOS11 YLR414C BNI5 YDL222C
3	2	MSN4 MSN2
4	32	CPH1 TKL2 HSP12 SPS100 MDJ1 GRX1 SSA3 ALD2 GDH3 GRE3 HOR2 ALD3 SOD2 ARA1 HSP42 YNL077W HSP78 GLK1 DOG2 HXK1 RAS2 CTT1 HSP26 TPS1 TTR1 HSP104 GLO1 SSA4 PNC1 MTC2 YGR086C PGM2
5	2	YAP1 SKN7
6	19	YMR318C CTT1 TSA1 CYS3 ZWF1 HSP82 TRX2 GRE2 SOD1 AHP1 YNL134C HSP78 CCP1 TAL1 DAK1 YDR453C TRR1 LYS20 PGM2

Table 1. Classification of the operons into $Q = 6$ clusters. Operons in bold belong to multiple clusters.

First, we notice that the clusters 1, 3, and 5 contain only two operons each. These operons correspond to hubs which regulate respectively the nodes of clusters 2, 4, and 6. More precisely, the nodes of cluster 2 are regulated by STE12 and TEC1 which are both involved in the response to glucose limitation, nitrogen limitation and abundant fermentable carbon source. Similarly, MSN4 and MSN2 regulate the nodes of cluster 4 in response to different stress such as freezing, hydrostatic pressure, and heat acclimation. Finally, the nodes of cluster 6 are regulated by YAP1 and SKN7 in the presence of oxygen stimulus. In the case of sparse networks, one of the clusters often contains most of the vertices having weak connection profiles, and is therefore not meaningful. Conversely, with our approach, the vectors \mathbf{Z}_i can have all their components set to zero, corresponding to vertices that do not belong to any cluster. Thus, we obtained 85 unclassified vertices. Our algorithm was able to uncover two overlapping clusters (operons in bold in Table. 1). Thus, SSA4 and TKL2 belong to cluster 2 and 4. Indeed, they are co-regulated by (STE12, TEC1) and (MSN4 and MSN2). Moreover, HSP78, CTT1, and PGM2 belong to cluster 4 and 6 since they are co-regulated by (MSN4, MSN2) and (YAP1, SKN7).

5 Conclusion

In this paper, we proposed a new latent logistic model to uncover overlapping clusters. We used both local and global variational techniques and we derived a variational EM algorithm to optimize the model parameters. We analyzed a transcriptional regulatory network of yeast and we showed that our model was able to handle sparsity. We discovered two overlapping clusters corresponding to co-regulated operons.