UNIVERSITY OF CALIFORNIA
SANTA CRUZ

SCALABLE, HIERARCHICAL AND DYNAMIC MODELING OF COMMUNITIES IN NETWORKS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS AND APPLIED MATHEMATICS

by

Pedro Regueiro Martinez

December 2017

The Dissertation of Pedro Regueiro Martinez
is approved:

Professor Abel Rodriguez, Chair

Professor Athanasios Kottas

Professor Rajarshi Guhaniyogi

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Scalable, hierarchical and dynamic modeling of communities in networks

by

Pedro Regueiro Martinez

The class of Bayesian stochastic blockmodels has become a popular approach for modeling and prediction with relational network data. This is due, in part, to the fact that inference on structural properties of networks follows naturally in this framework. Here, we study the problem of community detection under stochastic blockmodels in different settings.

First, we evaluate a stochastic gradient variational algorithm for stochastic models. Stochastic gradient variational algorithms have become a popular tool for approximate posterior inference in the statistics and machine learning literatures. We develop a new version of the algorithm and compare its performance to that of Markov chain Monte Carlo, the conventional method used to fit Bayesian stochastic blockmodels. We show that, although the SGV algorithm is scalable, its performance can be very poor, specially when there is substantial uncertainty in the community structure in the data.

Then, we turn our attention to the study of multilevel community structures in network data. That is, arrangements in which vertices group to form communities and, in turn, communities group into supercommunities. We propose a Bayesian hierarchical extension of stochastic blockmodel that is capable of identifying and recovering multilevel communities when these are present on the data. Markov chain Monte Carlo as well as variational algorithms are derived and evaluated.

Finally, we introduce a new dynamic stochastic blockmodel that allows us to study the evolution of communities across time. Our approach models both shifts in community membership using a fragmentation-coagulation prior, and changes in the propensities of interaction among communities using a variant of the autoregressive process. Computation is performed using a Markov chain Monte Carlo algorithm.

All models and algorithms are illustrated using both real and simulated data.

Acknowledgments

First an foremost I want to thank my advisor Abel Rodriguez for all his guidance, support, patience, and always spot-on intuition that saved me from wandering many unfruitful paths.

I would also like to thank my committee members, professor Athanasios Kottas and Rajarshi Guhaniyagui for their thoughtful and insightful comments.

To all my friends, old and older. Thanks for those conversations that improved this dissertation, and for all those times that kept me from finishing it sooner, they were all worth it.

To my family whose encouragement kept me going and, in particular, to my mother, without her support this simply would not have been possible.

Finally, I cannot even begin to thank my wife Agueda. I am beyond grateful that you decided to join me in this adventure. Thank you for always believing in me and for your continuous support and encouragement throughout this five years.

Chapter 1

Introduction

Complex systems of interrelated components are an essential part of our everyday lives. Interactions between proteins and other metabolites regulate our biological functions, individuals and organizations develop different types of relationships, and we rely on the Internet for communication and commercial purposes, to cite some examples. Networks constitute an adequate mathematical object to model interactions among components of such systems and, thus, their study can provide important qualitative information on the effect that the structure of interactions has over the system.

For this reason, networks have been extensively studied for many years now; in fact, their analysis is usually traced back to Leonard Euler in eighteenth century. Moreover, research in networks has been developed across various fields of science. Such is the case of mathematics, physics, statistics, sociology, computer science and, more recently, machine learning. This has led to a vast and diverse body of literature studying various properties and characteristics of networks. A good overview of the development of the field can be

found in Newman (2003) or the more extensive treatment of Newman (2010).

Now, probabilistic modeling of networks is much more recent than the study of networks. Early work in the area is found in Solomonoff and Rapoport (1951) and Luce et al. (1955), but it was not until the 1960s that this approach drew attention. With Erdős and Rényi (1959) and subsequent work, Paul Erdős and Alfréd Rényi popularized what today is the most widely studied model for networks, the *random graph*. Since then, a variety of extensions have been proposed; most notably, the *configuration model* of Bender and Canfield (1978) and the *exponential random graph* of Frank and Strauss (1986). These simple models have helped grasp an understanding of many features of networks, such as the distribution of vertex *degrees* (number of ties), vertex *centrality* (relative "importance" of a vertex), or *transitivity* (propensity of two vertices to be connected when they share a neighbor). Also, models like *preferential attachment* (Price, 1976; Barabási and Albert, 1999) and the *small world model* of Watts and Strogatz (1998) have shed light into the process of network formation, while Grassberger (1983) and Pastor-Satorras and Vespignani (2001) have studied how processes evolve in a network. A good overview of the statistically oriented literature can be found in Goldenberg et al. (2010).

Perhaps the best well studied problem within network analysis is that of *community detection*, which refers to the splitting a graph into *clusters* or *communities*. As accounted in Fortunato (2010); Newman (2004); Porter et al. (2009); Schaeffer (2007), many different solutions to this problem have been proposed in the literature. Among the most successful approaches are those based on the ideas of *agglomerative clustering*, *modularity* (density of subsets of vertices), and *betweenness* (extent to which a vertex lies on the path between

other vertices). However, the majority of the methods developed in this direction have been deterministic algorithms which are not able to provide a measure of the uncertainty associated with the solution they produce. Another drawback from this literature is the fact that these algorithms are usually developed to recover *assortative* structures; that is, communities with high density of connections within the vertices of the same community but few interactions across communities.

Among probabilistic models that can be used for community detection are *latent social space models* of Hoff et al. (2002) where induced communities are found using a mixture model in the latent Euclidean space as in Handcock et al. (2007). In contrast, the present work explores the topic of community detection based on *stochastic blockmodels* of Holland et al. (1983). Stochastic blockmodels naturally lend themselves to the problem of community detection as they are based on the idea of modeling interaction probabilities by partitioning the network into groups of structurally equivalent vertices. Furthermore, stochastic blockmodels posses the appealing features of being capable of simultaneously recovering assortative and disassortative mixing.

In particular, we look at three different problems within this context. First, given the heavy computational requirements of *Markov chain Monte Carlo* algorithms, in Chapter 2 we explore an alternative computational method to fit the model. *Stochastic variational* inference combines ideas from stochastic optimization (Robbins and Monro, 1951) and variational Bayes algorithms (Saul et al., 1996), and it has been applied to a wide range of models including topic models in the original paper by Hoffman et al. (2013) and the mixed membership stochastic blockmodels of Airoldi et al. (2009) in Gopalan et al. (2013). Our

goal here is to evaluate the performance of the stochastic variational algorithm in the specific setting of community detection under the stochastic blockmodel and compare it to that of the MCMC. These comparisons are carried out both in terms of computational efficiency, and posterior inference and predictive accuracy.

Secondly, a feature that is commonly observed in network data is the hierarchical structure of communities. That is, nested arrangements in which vertices in the network group to form communities and, in turn, communities group into so-called *supercommunities*. With exception of agglomerative clustering based methods, little attention has been paid to having mechanisms that are capable of recovering this kind of multilevel community structure. Agglomerative clustering methods, such as the work of Clauset et al. (2007), place a probability distribution directly over the space of dendrograms with the network's vertices as leaves. Instead, in Chapter 3 we introduce a hierarchical extension of the stochastic blockmodel that is able to capture the multilevel structure of communities, our work is closest to that of Ho et al. (2012), though we use a fundamentally different approach to introduce the hierarchy in the community parameters.

Finally, in Chapter 4 we turn our attention to the problem of modeling the evolution of communities in dynamic networks. That is, in a setting where a network is observed repeatedly across multiple points in time. A common strategy to deal with dynamic network data has been the generalization of static models. In this way, works like Sarkar and Moore (2005), Westveld and Hoff (2011), Durante and Dunson (2014) and Sewell and Chen (2015) have extended latent space models, while Guo et al. (2007) and Hanneke et al. (2010) introduced temporal versions of the exponential random graph. See also Goldenberg et al. (2010)

for a good overview of the early work in the area. Dynamic extensions of the stochastic blockmodel have also been proposed in works like Rodríguez (2012) and Betancourt et al. (2015) where an extension based on hidden Markov models models is introduced. Here, we propose an extension based on the fragmentation-coagulation processes of Bertoin (2006).

We begin the discussion in Section 1.1 by introducing the main ideas behind stochastic blockmodels.

1.1 Stochastic blockmodels

The stochastic blockmodel (Holland et al., 1983) is a simple, yet very flexible model that allows to represent different kind of interactions among different types of agents in a complex system. In this section we concentrate on the case of *simple*, *unweighted* and *undirected* networks, which can be characterized in terms of their *adjacency matrix* or *sociomatrix*, $Y \in \Re^{I \times I}$, given by

$$Y_{i,j} = \begin{cases} 1 & \text{if there is an edge between vertices } j \text{ and } i \\ 0 & \text{otherwise} \end{cases}$$

where I represent the number of vertices in the network. For any undirected network the adjacency matrix is, by construction, symmetric and, therefore, is possible to disregard the observations below (or above) the main diagonal. Furthermore, in the case of a network without self interactions, it is also possible to disregard the observations on the main diagonal as they ate all assumed to be structural zeros. Consequently, the set of observations is taken to be

$$\mathcal{Y} = \{y_{i,j} : 1 \leq i < j \leq I,\ i, j \in \mathbb{N}\}.$$

The binary nature of interactions naturally suggests these are modeled trough a Bernoulli distribution

$$y_{i,j} \mid \lambda_{i,j} \sim \mathcal{B}er(\lambda_{i,j}); \qquad 1 \leq i < j \leq I. \tag{1.1}$$

Now, the basic idea behind the stochastic blockmodel is that the network can be partitioned into $K \leq I$ groups or communities, where two vertices are in the same community only if they have equal interaction probabilities across the network. Formally,

$$\lambda_{i,j} = g(\theta_{\xi_i,\xi_j})$$

where the block indicators $\xi_1, \xi_2, \ldots, \xi_I$ take value in the set $\{1, 2, \ldots, K\}$, the elements of the set $\{\theta_{k,l}\}_{k,l=1}^{K}$ are usually referred to as the community parameters, and g is an appropiate link function. Notice that, again, because of symmetry, attention can be constraint to the set

$$\Theta = \{\theta_{k,l} : 1 \leq k \leq l \leq K,\ k, l \in \mathbb{N}\}.$$

Assuming conditional independence in the interactions both within and across actors the likelihood can be expressed as

$$p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}) = \prod_{i=1}^{I-1} \prod_{j=i+1}^{I} p\left(y_{i,j} \mid \theta_{\phi(\xi_i,\xi_j)}\right). \tag{1.2}$$

where henceforth $\phi : \Re^2 \to \Re^2$ denotes $\phi(u, v) = (\min\{u, v\}, \max\{u, v\})$, which ensures mapping to the elements in Θ. In turn, (1.2) implies that

$$p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}) = \prod_{k=1}^{K} \prod_{l=k}^{K} \{g(\theta_{k,l})\}^{s_{k,l}} \{1 - g(\theta_{k,l})\}^{n_{k,l} - s_{k,l}} \tag{1.3}$$

with $s_{k,l} = \sum_{\mathcal{S}_{k,l}} y_{i,j}$ and $n_{k,l} = \sum_{\mathcal{S}_{k,l}} 1$, and the sum is taken over the set

$$\mathcal{S}_{k,l} = \{(i, j) :\ i < j,\ (k, l) = \phi(\xi_i, \xi_j)\}.$$

With respect to the maximum number of communities K, a nonparametric approach can be taken as in the *infinite relational model* of Kemp et al. (2006). This model allows for the effective number of communities in the network $K^\star \leq K$ to be learned from the data, being able to take any integer value between 1 and I. Specifically, $\boldsymbol{\xi}$ is assumed to follow a *Chinese restaurant process* (CRP) prior, which implies that its distribution is given by Ewens sampling formula (Ewens, 1972), *i.e.,*

$$p(\xi_1, \xi_2, \ldots, \xi_I) = \frac{\Gamma(\alpha)\alpha^{K^\star}}{\Gamma(\alpha + I)} \prod_{k=1}^{K^\star} \Gamma(n_k).$$

Alternatively K can be fixed trying to overestimate the number of communities in the network, thus leading to a finite mixture model as in Nowicki and Snijders (2001). Here, for simplicity, we take this later approach. However, for the block indicators we assume that the entries of $\boldsymbol{\xi}$ are exchangeable and follow a Categorical distribution in $\{1, 2, \ldots, K\}$

$$Pr(\xi_i = k \mid w_k) = w_k; \qquad i = 1, 2, \ldots, I, \tag{1.4}$$

with weights vector $\boldsymbol{w}$ satisfying

$$\boldsymbol{w} \sim \mathcal{D}ir\left(\boldsymbol{\alpha}_{\boldsymbol{w}}\right). \tag{1.5}$$

which, as discussed in Ishwaran and Zarepour (2000) and Neal (2000), if the parameter vector is chosen as $\boldsymbol{\alpha}_{\boldsymbol{w}} = \left(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$, as $K \to \infty$, leads to a model that approximates the infinite relational model.

Regarding the community parameters, a simple and computationally convenient choice of prior is achieved by taking g to be the identity function and the elements of Θ independent a priori from a common Beta distribution

$$\theta_{k,l} \sim \mathcal{B}eta(a, b). \tag{1.6}$$

Assuming conditional independence in the elements of Θ leads to a model where communities and actors are exchangeable in the network. The *hyperparameters* a and b control the propensity of interactions within and across communities a priori; they could be fixed, for example to imply a non-informative Uniform distribution with $a = b = 1$. Otherwise, an additional hierarchical layer can be added to allow further information pulling by assigning a prior distribution $\pi(a, b)$ such as independent Exponential distributions.

Alternatively, a logit structure can be used under a Gaussian prior for the elements of Θ. Specifically, if g is taken to be the canonical link, that is, $\theta_{\xi_i,\xi_j} = \log\left(\frac{\lambda_{i,j}}{1-\lambda_{i,j}}\right)$, each element in $\mathcal{Y}$ satisfies

$$p(y_{i,j} \mid \theta_{\xi_i,\xi_j}) = \frac{(\exp\{\theta_{\phi(\xi_i,\xi_j)}\})^{y_{i,j}}}{1 + \exp\{\theta_{\phi(\xi_i,\xi_j)}\}},$$

and in this case the likelihood reduces to

$$p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}) = \prod_{k=1}^{K} \prod_{l=k}^{K} \frac{(\exp\{\theta_{k,l}\})^{s_{k,l}}}{(1 + \exp\{\theta_{k,l}\})^{n_{k,l}}}. \tag{1.7}$$

The community parameters can then be assumed conditionally independent from a common Gaussian prior

$$\theta_{k,l} \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2). \tag{1.8}$$

In this case μ affects the overall density of the network, while σ^2 control the variability among the propensity of interaction between the different clusters. Thus, setting $\mu = 0$ centers the interaction probabilities at $\frac{1}{2}$, while, considering the transformation, choosing $\sigma^2 = 1$ leaves approximately 95% of the mass in $[0.12, 0.88]$ a priori for all $\lambda_{i,j}$. If instead a hyperprior, $\pi(\mu, \sigma^2)$, is to be placed in these parameters, one computationally convenient option is choosing conditionally conjugate distributions

$$\mu \sim \mathcal{N}(\mu_\mu, \sigma_\mu^2) \quad \text{and} \quad \sigma^2 \sim \mathcal{IG}(\alpha_\sigma, \beta_\sigma). \tag{1.9}$$

The concentration parameter, α, controls number of occupied communities in the network $K^\star$. In the limit case of the infinite relational model, from Antoniak (1974), it is known that the distribution satisfies

$$Pr(K^\star = k \mid \alpha) = S(I,k)\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + I)}$$

where S represents the unsigned Sterling numbers of the first kind. Thus,

$$\mathbb{E}[K^\star \mid \alpha] = \alpha \left[\Psi(\alpha + I) - \Psi(\alpha)\right] \approx \alpha \log\left(\frac{\alpha + I}{\alpha}\right)$$

and

$$\mathbb{V}[K^\star \mid \alpha] = \alpha \left[\Psi(\alpha + I) - \Psi(\alpha)\right] + \alpha^2 \left[\Psi'(\alpha + I) - \Psi'(\alpha)\right] \approx \alpha \log\left(\frac{\alpha + I}{\alpha}\right)$$

with the first order approximations valid for large I.

In the parametric case we explore the effect of α via simulation. As an example, Figure 1.1 shows the empirical CDF of $K^\star$ for four different values of α in the case where $I = K = 100$. From this figure is possible to observe that, in this case, the behavior expected in the nonparametric model is also present for a finite K; namely, that the expected number of clusters increases with α in the order of $\alpha \log(I/\alpha)$.

If alternatively, α is to be learned from the data, a common choice of hyperprior is

$$\alpha \sim \mathcal{G}(\alpha_\alpha, \beta_\alpha) \tag{1.10}$$

with the Gamma distribution parametrized in terms of shape and rate.

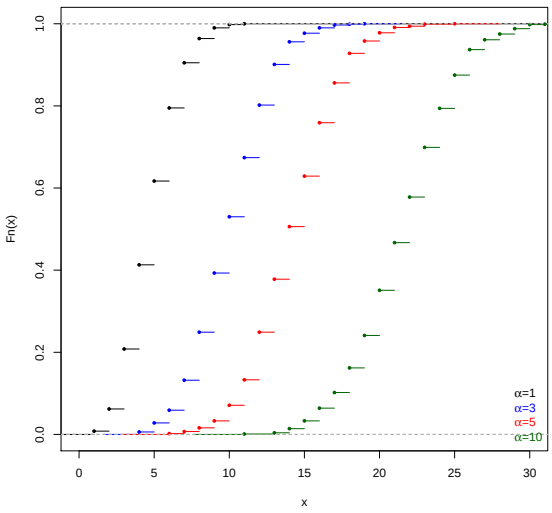

Figure 1.1: CDF of the effective number of communities $K^\star$ implied by the prior for four different values of α, $\alpha = 1$ (black), $\alpha = 3$ (blue), $\alpha = 5$ (red), and $\alpha = 10$ (green) when $K = I = 100$

1.2 Posterior inference using Markov chain Monte Carlo algorithms

The model described above does not lead to closed form posteriors and, thus, some form of approximation is required. The most usual way to fit this models is posterior sampling via Markov chain Monte Carlo (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; Gelfand and Smith, 1990). In this section we describe an MCMC algorithm for the stochastic blockmodel under the Beta prior. For brevity in the exposition, consider the case in which α, a and b are assumed know.

Making use of Bayes theorem, equations (1.3) to (1.6) can be combined to express

the posterior distribution as

$$p(\Theta, \boldsymbol{\xi}, \boldsymbol{w} \mid \mathcal{Y}) \propto \prod_{k=1}^{K} \prod_{l=k}^{K} \{\theta_{k,l}\}^{a+s_{k,l}-1} \{1-\theta_{k,l}\}^{b+n_{k,l}-s_{k,l}-1} \prod_{k=1}^{K} w_k^{\frac{\alpha}{K}+n_k-1} \tag{1.11}$$

where $n_k = n_k(\boldsymbol{\xi}) = \sum_{\mathcal{S}_k} 1$ and the sum is taken over $\mathcal{S}_k = \{i : \xi_i = k\}$. From equation (1.11) is possible to obtain the full conditional distributions of the model parameters. Particularly, in the case of $\boldsymbol{\xi}$,

$$Pr(\xi_i = k \mid \Theta, \boldsymbol{\xi}_{-i}, \boldsymbol{w}, \mathcal{Y}) \propto w_k \prod_{\substack{j=1 \\ j \neq i}}^{I} \left\{\theta_{\phi(\xi_j,k)}\right\}^{y_{\phi(i,j)}} \left\{1-\theta_{\phi(\xi_j,k)}\right\}^{1-y_{\phi(i,j)}} \tag{1.12}$$

for $k \in \{1, 2, \ldots, K\}$. Thus, for every $i = 1, 2, \ldots, I$, ξ_i can be sampled from a Categorical distribution with weights vector given by normalizing the RHS of equation (1.12).

Now, in the case of Θ,

$$p(\theta_{k,l} \mid \Theta_{-kl}, \boldsymbol{\xi}, \boldsymbol{w}, \mathcal{Y}) \propto \theta_{k,l}^{a+s_{k,l}-1}(1-\theta_{k,l})^{b+n_{k,l}-s_{k,l}-1} \tag{1.13}$$

which is easily identified as the kernel of a Beta distribution with parameters $a + s_{k,l}$ and $b + n_{k,l} - s_{k,l}$. From this distribution is interesting to observe that, in the case of an empty component, that is if no actor is assigned to either community k or l, the full conditional distribution reduces to the prior.

Finally, the weights can be sampled from

$$p(\boldsymbol{w} \mid \Theta, \boldsymbol{\xi}, \mathcal{Y}) \propto \prod_{k=1}^{K} w_k^{\frac{\alpha}{K}+n_k-1} \tag{1.14}$$

a Dirichlet with parameter vector $\left(\frac{\alpha}{K} + n_1, \frac{\alpha}{K} + n_2, \ldots, \frac{\alpha}{K} + n_K\right)$.

Chapter 2

Stochastic variational inference for the stochastic blockmodel

The methods described in Section 1.2 possess certain desirable properties. Namely, they are relatively easy to implement, the Markov chain implicitly defined is guaranteed to eventually converge to the posterior distribution, and and they allow to control the accuracy level of the approximation by controlling the number of samples. In practice, however, the rate of convergence may be slow and, as the number of parameters grows, the computational burden can make this approach infeasible. Specific to the setting of network analysis, notice that as a network grows the number of interactions grows $\mathcal{O}(I^2)$, and, although perhaps not at the same rate, the number of communities, and hence the number of parameters, is expected to increase accordingly. Thus MCMC algorithms can result impractical, even for moderately large networks. This chapter explores an alternative class of methods to approximate the posterior distribution.

2.1 Variational approximations

Consider the problem of approximating an unknown function that can be evaluated up to a proportionality constant (p), by another function (q) that is restricted to be a member of a certain family of functions. To this end, it is possible to define a functional measure of "dissimilarity" between p and q, and use calculous of variations techniques to minimize that measure, thus finding the q in such family that is "closest" to p. The idea of applying this technique to case where p is chosen to be a posterior distribution and q is a density can be traced back to the mid 90's, in works like Saul et al. (1996), and is now known in the literature as variational Bayes. A good overview of the early development in this topic can be found in Jordan et al. (1999).

Briefly, the main idea can be summarized as follows. Let $\boldsymbol{\varphi}$ be a set of parameters and $p(\boldsymbol{\varphi} \mid \mathbf{x})$ its posterior distribution after some data $\mathbf{x}$ has been observed. The purpose is to approximate $p(\boldsymbol{\varphi} \mid \mathbf{x})$ with $q(\boldsymbol{\varphi})$. Specifically, if the Kullback–Leibler divergence is chosen as a measure of dissimilarity, the problem becomes

$$\min_{q} \int q(\boldsymbol{\varphi}) \log \frac{q(\boldsymbol{\varphi})}{p(\boldsymbol{\varphi} \mid \mathbf{x})} d\boldsymbol{\varphi}. \tag{2.1}$$

Now, it is easily shown that

$$\int q(\boldsymbol{\varphi}) \log \frac{q(\boldsymbol{\varphi})}{p(\boldsymbol{\varphi} \mid \mathbf{x})} d\boldsymbol{\varphi} = \log p(\mathbf{x}) - \int q(\boldsymbol{\varphi}) \log \frac{p(\boldsymbol{\varphi}, \mathbf{x})}{q(\boldsymbol{\varphi})} d\boldsymbol{\varphi}$$

and, thus, minimizing $\mathrm{KL}[q \mid\mid p]$ is equivalent to maximizing $\mathbb{E}_{q(\boldsymbol{\varphi})}\left[\log \frac{p(\boldsymbol{\varphi},\mathbf{x})}{q(\boldsymbol{\varphi})}\right]$ which is known in physics as the *free energy* and in the computer science literature as the *evidence lower bound* (ELBO). Note that the ELBO can be decomposed as

$$F(q, \mathbf{x}) = \mathbb{E}_{q(\boldsymbol{\varphi})}[\log p(\mathbf{x}, \boldsymbol{\varphi})] + H\left[q(\boldsymbol{\varphi})\right]$$

where H denotes the Shannon entropy. Furthermore, if q is assumed to satisfy the mean field assumption

$$q(\boldsymbol{\varphi}) = \prod_i q_i(\varphi_i)$$

where $q_i(\varphi_i)$ are the marginal variational densities, the solution of this problem satisfies

$$\log q_i^\star(\varphi_i) \propto \mathbb{E}_{q(\boldsymbol{\varphi}_{-i})} \left[\log p(\boldsymbol{\varphi}, \mathbf{x})\right] \tag{2.2}$$

which leads to a coordinate optimization algorithm. In particular, when in the exponential family of distributions, the solution to (2.2) is readily available.

It is important to notice that the choice of the Kullback–Leibler divergence is an arbitrary one. Naturally, other choices could be taken, such is the case of L_1 or L_2 divergence. Even more, since Kullback–Leibler is not symmetric, it would also be possible to consider the minimization of $\mathrm{KL}[p \mid\mid q]$, which has been studied under the name of *expectation propagation*. In practice, the choice described in (2.1) is computationally convenient since it leads to closed form optimization, which is not generally true for other measures.

Consider the Beta blockmodel with fixed hyperparameters as discussed in Section 1.2. Before going into the variational algorithm, notice that when interest lies in the community indicators only, it is possible to marginalize (*collapse*) over the weight parameters to obtain

$$p(\boldsymbol{\xi}) = \frac{\Gamma(\alpha)}{\left[\Gamma\left(\frac{\alpha}{K}\right)\right]^K \Gamma(I+\alpha)} \prod_{k=1}^{K} \Gamma\left(\frac{\alpha}{K} + n_k\right) \tag{2.3}$$

and, thus,

$$Pr(\xi_i = k \mid \boldsymbol{\xi}_{-i}) = \frac{n_k^{-i} + \frac{\alpha}{K}}{(I-1)+\alpha} \text{ for all } i \in \{1, 2 \ldots, I\} \text{ and } k \in \{1, 2, \ldots, K\} \tag{2.4}$$

where $n_k^{-i} = \sum_{j \neq i} \mathbb{1}_{\{\xi_j = k\}}$

In this way, it is possible to find the variational approximation $q(\Theta, \boldsymbol{\xi})$ to the marginal posterior distribution $p(\Theta, \boldsymbol{\xi} \mid \mathcal{Y}) \propto p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}) p(\Theta) p(\boldsymbol{\xi})$ rather that approximating the complete posterior distribution $p(\Theta, \boldsymbol{\xi}, \boldsymbol{w} \mid \mathcal{Y})$. This, effectively reduces the number of constrains imposed by the mean field assumptions and, thus, potentially improving the approximation. Now, for $1 \leq k \leq l \leq K$, the solution to (2.1) satisfies

$$\log q^{\star}(\theta_{k,l}) \propto \mathbb{E}_{q(\Theta_{-kl}, \boldsymbol{\xi})}[\log p(\mathcal{Y}, \Theta, \boldsymbol{\xi})] \propto \mathbb{E}_{q(\Theta_{-kl}, \boldsymbol{\xi})}[\log p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi})] + \log p(\theta_{k,l})$$

which implies that

$$q^{\star}(\theta_{k,l}) \propto \exp\left\{ \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} [y_{i,j} \log \theta_{k,l} + (1 - y_{i,j}) \log(1 - \theta_{k,l})] r_{k,l}^{i,j} \right\} \theta_{k,l}^{a-1} (1 - \theta_{k,l})^{b-1} \quad (2.5)$$

where

$$r_{k,l}^{i,j} = \begin{cases} q(\xi_i = k) q(\xi_j = l) + q(\xi_i = l) q(\xi_j = k) & \text{if } k \neq l \\ q(\xi_i = k) q(\xi_j = l) & \text{if } k = l. \end{cases}$$

That is, the variational distribution of $\theta_{k,l}$ is a $\mathcal{B}eta(a_{k,l}^{\star}, b_{k,l}^{\star})$ with

$$a_{k,l}^{\star} = a + \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} r_{k,l}^{i,j} y_{i,j} \quad \text{and} \quad b_{k,l}^{\star} = b + \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} r_{k,l}^{i,j} (1 - y_{i,j}). \quad (2.6)$$

Also, for $i \in \{1, 2, \ldots, I\}$

$$\log q^{\star}(\xi_i) \propto \mathbb{E}_{q(\Theta, \boldsymbol{\xi_{-i}})}[\log p(\mathcal{Y}, \Theta, \boldsymbol{\xi})] \propto \mathbb{E}_{q(\Theta, \boldsymbol{\xi_{-i}})}[\log p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi})] + \mathbb{E}_{q(\boldsymbol{\xi_{-i}})}[\log p(\xi_i \mid \boldsymbol{\xi}_{-i})].$$

Following Kurihara et al. (2007) the second term in this last expression can be approximated using the second order Delta method and the fact that $n_{\xi_i}^{-i}$ is a sum of independent Bernoulli random variables. Specifically, for any i

$$\mathbb{E}_{q(\boldsymbol{\xi_{-i}})}\left[\log\left(n_{\xi_i}^{-i} + \frac{\alpha}{K}\right)\right] \approx \log\left(\sum_{j \neq i} q(\xi_j = \xi_i) + \frac{\alpha}{K}\right) - \frac{\sum_{j \neq i} q(\xi_j = \xi_i)(1 - q(\xi_j = \xi_i))}{2\left[\sum_{j \neq i} q(\xi_j = \xi_i)\right]^2}$$

Therefore, for any $k \in \{1, 2, \ldots, K\}$

$$\log q^{\star}(\xi_i = k) \propto \sum_{j \neq i} \left[y_{\phi(i,j)} \chi_{k,1} + (1 - y_{\phi(i,j)}) \chi_{k,2} \right] + \log \left(\sum_{j \neq i} q(\xi_j = k) + \frac{\alpha}{K} \right) - \frac{1}{2} \frac{\sum_{j \neq i} q(\xi_j = k)(1 - q(\xi_j = k))}{\left[\sum_{j \neq i} q(\xi_j = k) \right]^2} \quad (2.7)$$

where

$$\chi_{k,1} = \sum_{l=1}^{K} \left[\left(\psi \left(a^{\star}_{\phi(k,l)} \right) - \psi \left(a^{\star}_{\phi(k,l)} + b^{\star}_{\phi(k,l)} \right) \right) q(\xi_j = l) \right],$$

$$\chi_{k,2} = \sum_{l=1}^{K} \left[\left(\psi \left(b^{\star}_{\phi(k,l)} \right) - \psi \left(a^{\star}_{\phi(k,l)} + b^{\star}_{\phi(k,l)} \right) \right) q(\xi_j = l) \right]$$

and ψ represents the Digamma function. Finally, in a similar fashion, the free energy is given by

$$\begin{aligned} F(q, \mathcal{Y}) \approx{}& \frac{1}{2} K(K+1) \left(\log \Gamma(a+b) - \log \Gamma(a) - \log \Gamma(b) \right) + \log \Gamma(\alpha) - \log \Gamma(I + \alpha) \\ & - K \log \Gamma \left(\frac{\alpha}{K} \right) + \sum_{k=1}^{K} \sum_{l=k}^{K} \log \Gamma(a^{\star}_{k,l}) + \log \Gamma(b^{\star}_{k,l}) - \log \Gamma(a^{\star}_{k,l} + b^{\star}_{k,l}) \\ & + \sum_{k=1}^{K} \left[\log \Gamma \left(\sum_{i=1}^{I} q(\xi_i = k) + \frac{\alpha}{K} \right) + \psi_1 \left(\sum_{i=1}^{I} q(\xi_i = k) + \frac{\alpha}{K} \right) \left(\sum_{i=1}^{I} q(\xi_i = k)(1 - q(\xi_i = k)) \right) \right] \\ & - \sum_{i=1}^{I} \sum_{k=1}^{K} q(\xi_i = k) \log q(\xi_i = k) \end{aligned}$$

where ψ_1 denotes the Trigamma function.

In general, variational Bayes algorithms tend to converge to a local optimum in a relatively small number of iterations, making them fast in comparison to Markov chain Monte Carlo algorithms. However, because the calculations required in the computation of $\mathbb{E}_{q(\boldsymbol{\varphi}_{-i})} [\log p(\boldsymbol{\varphi}, \mathbf{x})]$ are similar to those of the corresponding full conditional, $p(\varphi_i \mid \boldsymbol{\varphi}_{-i}, \mathbf{x})$, these methods also suffer from scalability issues. Essentially, the difficulty lies on the fact that, for each parameter, calculations involving the entire data matrix are required.

An extension of this algorithm, introduced by Hoffman et al. (2013) to address these scalability concerns is *stochastic variational inference*. Using stochastic optimization ideas, this algorithm speeds computations producing noisy estimates of the (natural) gradient within a coordinate ascent approach. This algorithm relies on a split of the parameter set into local and global parameters $\boldsymbol{\varphi} = (\boldsymbol{\varphi_l}, \boldsymbol{\varphi_g})$ in such a way that $p(\boldsymbol{\varphi_l}, \mathbf{x} \mid \boldsymbol{\varphi_g}) = \prod_i p(\varphi_i, x_i \mid \boldsymbol{\varphi_g})$, and $p(\boldsymbol{\varphi_l}, \mathbf{x} \mid \boldsymbol{\varphi_g})$ and $p(\boldsymbol{\varphi_g})$ are conjugate in the exponential family. Specifically, if

$$p(\varphi_i, x_i \mid \boldsymbol{\varphi_g}) = h(\varphi_i, x_i) \exp\{\boldsymbol{\varphi_g}^T t(\varphi_i, x_i) - a_l(\boldsymbol{\varphi_g})\}$$

and

$$p(\boldsymbol{\varphi_g}) = h(\boldsymbol{\varphi_g}) \exp\{\mathbf{h}^T t(\boldsymbol{\varphi_g}) - a_g(\mathbf{h})\},$$

conjugacy implies that $t(\boldsymbol{\varphi_g})^T = (\boldsymbol{\varphi_g}, -a_l(\boldsymbol{\varphi_g}))$ and thus

$$p(\boldsymbol{\varphi_g} \mid \boldsymbol{\varphi_l}, \mathbf{x}) \propto h(\boldsymbol{\varphi_g}) \exp\{\boldsymbol{\eta}^T(\boldsymbol{\varphi_l}, \mathbf{x}) t(\boldsymbol{\varphi_g})\}$$

where $\boldsymbol{\eta}^T(\boldsymbol{\varphi_l}, \mathbf{x}) = \mathbf{h}^T + (\sum_i^N t(\varphi_i, x_i), N)$, with N the dimension of $\mathbf{x}$. Furthermore, in that case the solution to the variational distribution is in the same exponential family

$$q(\boldsymbol{\varphi_g}) \propto h(\boldsymbol{\varphi_g}) \exp\{\mathbf{h_q}^T t(\boldsymbol{\varphi_g}) - a_g(\mathbf{h_q})\}$$

and, therefore, the natural gradient of the free energy can be written as

$$\hat{\nabla}_{\mathbf{h_q}} F(q, \mathbf{x}) = \mathbb{E}_{q^\star(\boldsymbol{\varphi_l})}[\boldsymbol{\eta}^T(\boldsymbol{\varphi_l}, \mathbf{x})] - \mathbf{h_q}$$

where $q^\star(\boldsymbol{\varphi_l})$ is the regular variational distribution for the local parameters. This result can be used in a Robbins and Monro (1951) algorithm taking $x^{(L)}$ a random subsample of the

data, $\boldsymbol{\varphi_l}^{(L)}$ the corresponding local parameters and $\mathcal{C}$ the appropriate scaling factor, and setting

$$\mathbf{h_q}^{(t)} = (1-\rho_t)\mathbf{h_q}^{(t-1)} + \rho_t \mathcal{C}\mathbb{E}_{q^\star(\boldsymbol{\varphi}_l^{(L)})}[\boldsymbol{\eta}^T(\boldsymbol{\varphi_l}^{(L)}, \mathbf{x}^{(L)})] \tag{2.8}$$

which is guaranteed to converge to a local minimum as long as the positive real sequence ρ_t satisfies

$$\sum_{t=1}^{\infty} \rho_t \to \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \rho_t^2 < \infty. \tag{2.9}$$

A common choice for the sequence of step sizes is given by $\rho_t = (t+\tau)^{-\kappa}$ where $\kappa \in \left(\frac{1}{2}, 1\right]$ represents the forgetting rate and $\tau \geq 0$ is known as the delay. This scheme is illustrated in Figure 2.1 for various choices of τ and κ. From this figure is possible to observe that κ determines how fast the step size declines, and τ affects the starting level of the sequence $\{\rho_t\}$.

Since the stochastic blockmodel previously discussed satisfies the assumptions of the stochastic variational inference algorithm, it is straightforward to derive the required updates from equations (2.2) and (2.8). The resulting algorithm is summarized as follows:

1. Randomly initialize the global parameters $a_{k,l}^{\star(0)}$, $b_{k,l}^{\star(0)}$ for $1 \leq k \leq l \leq K$.

2. Repeat

    2.1. Randomly obtain a subnetwork S by uniformly sampling the vertices in the original network.

    2.2. Update the local variational probabilities $q(\xi_i = k)$ for all $i \in S$ and all k using equation (2.7) with the corresponding scaling factor.

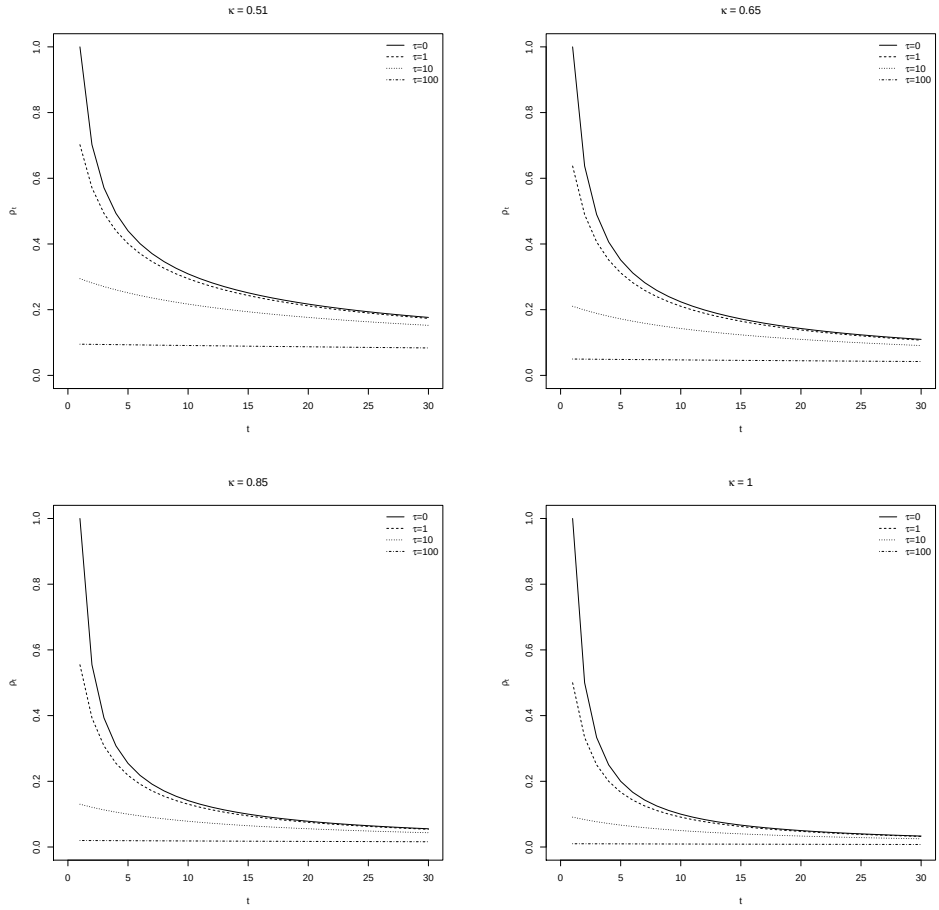

Figure 2.1: Evolution of the step size in the stochastic variational algorithm under the scheme $\rho_t = (t + \tau)^{-\kappa}$ for different choices of κ and τ.

2.3. Compute the intermediate global parameters $\hat{a}^{\star}_{k,l}$, $\hat{b}^{\star}_{k,l}$ using the noisy gradient into equation (2.6).

2.4. Update the estimates of the global variational parameters using (2.8).

$$a^{\star(t)}_{k,l} = (1 - \rho_t)a^{\star(t-1)}_{k,l} + \rho_t \mathcal{C}\hat{a}^{\star}_{k,l} \text{ and } b^{\star(t)}_{k,l} = (1 - \rho_t)b^{\star(t-1)}_{k,l} + \rho_t \mathcal{C}\hat{b}^{\star}_{k,l}.$$

Notice that some of the computations required to calculate the ELBO are precisely those avoided by the stochastic variational algorithm. For this reason, in order to asses convergence of the algorithm, we track a noisy estimate of the ELBO computed over a fixed

subnetwork composed of randomly selected vertices. Finally, it is worthwhile mentioning that the multimodality of the problem makes both MCMC and variational approximation algorithms susceptible to initial conditions. This problem is particularly salient in the standard variational Bayes, and is slightly ameliorated with the noisy gradient estimates of stochastic variational Bayes. Nonetheless multiple runs, which can be parallelized, are required.

2.2 Evaluation

2.2.1 Simulated data

In this section we evaluate the performance of the stochastic variational algorithm described in Section 2.1 in a setting in which the ground truth is known. To this end we make use of the simulated dataset represented graphically in figure 2.2. This network is constructed with $I = 350$ actors, split evenly in $K^\star = 7$ communities. In the experiments throughout this section we set the maximum number of communities to $K = 20$ and the model hyperparameters as $a = b = \alpha = 1$.

The first issue we address is the selection of the parameters in the sequence $\{\rho_t\}$ and the block size $|S|$. In principle, any sequence satisfying conditions (2.9) leads to an algorithm that is guaranteed to converge to a local mode. However, the choice of these step sizes can have an important effect in the rate of convergence. Furthermore, the effect of $\{\rho_t\}$ is also dependent in the block sizes $|S|$. Using this simulated dataset, setting the maximum number of communities $K = 20$ and choosing the hyperparameters $\alpha = a = b = 1$, while fixing the form of $\rho_t = (\tau + t)^{-\kappa}$, Figure 2.3 compares different choices of $\tau \geq 0$ and

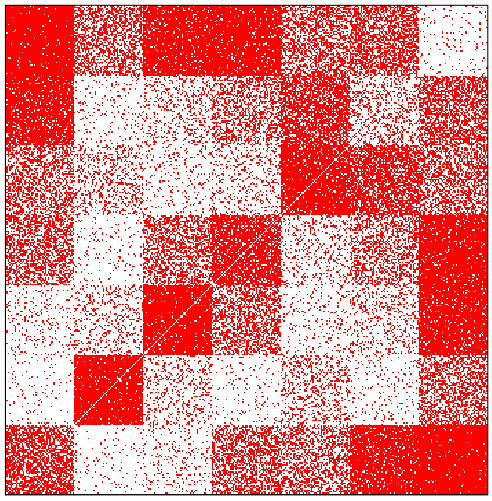

Figure 2.2: Pictorial representation of the adjacency matrix. Here actors in the network are placed along the x and y axis. $Y_{i,j} = 1$ is represented by a red dot, while a lack of interaction is shown in white.

$\kappa \in \left(\frac{1}{2}, 1\right]$. Here, for 32 random initial conditions and each combination of κ, τ and ω the stochastic variational algorithm is allowed to run for a *time budget* equal to the time that the standard variational algorithm takes to converge. The free energy and entropy are then calculated using the complete network and the difference is evaluated. The resulting boxplots, which show the variability due to the starting point, are displayed in this figure. In particular, notice that the case of $\omega = 1$ does not correspond to the variational Bayes algorithm as, even though the gradient is calculated with the whole dataset, the algorithm takes only a partial step in the direction of the gradient. With almost $3,000$ runs of the stochastic variational algorithm, this figure encompasses a vast amount of information for the dataset at hand and suggests the use of non extreme values of ω, τ and κ. In particular, very small fractions for ω tend to provide poor results, while 25% to 33% seems to work well in this case. Although these results are specific to this dataset and do not necessarily

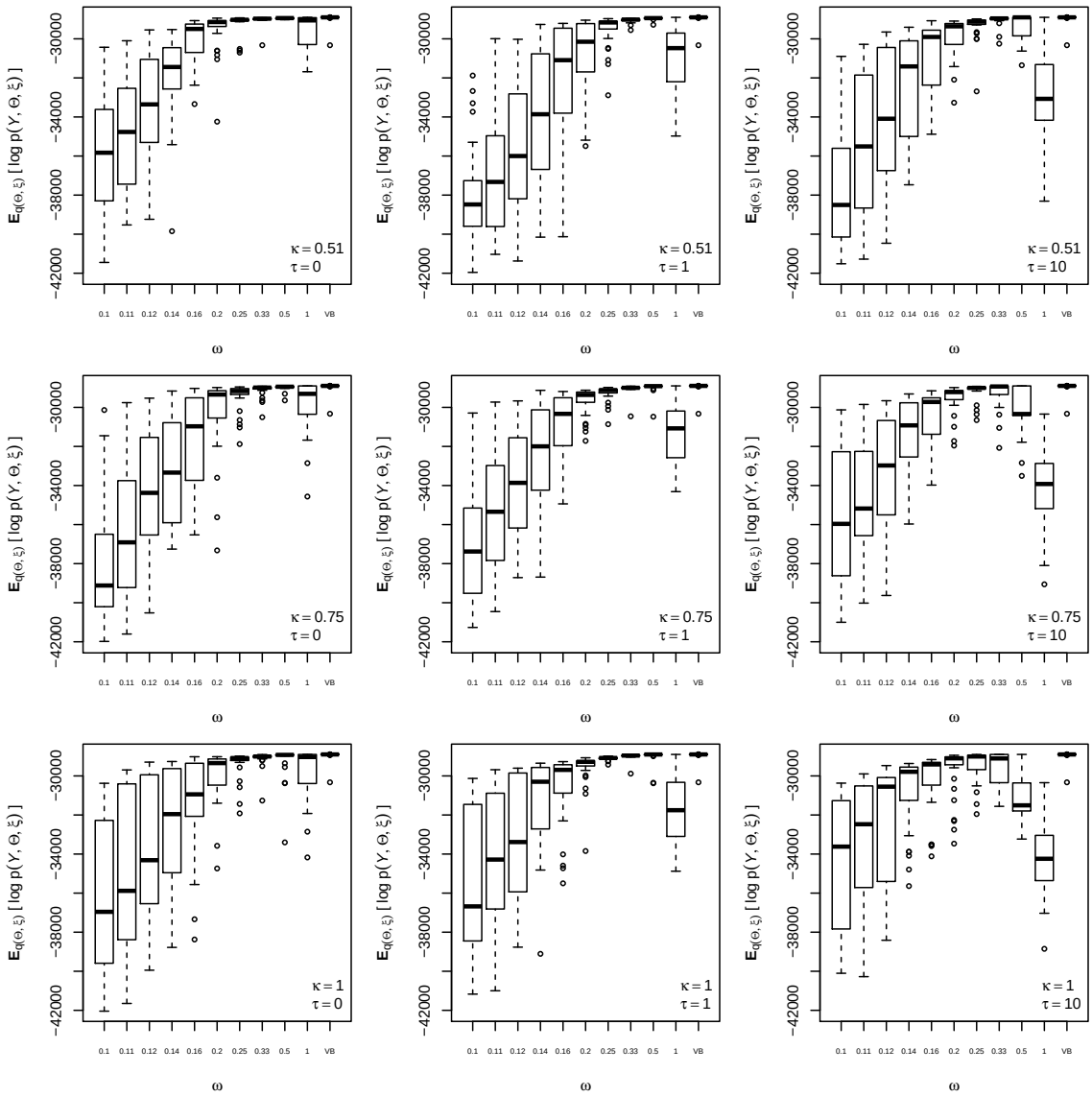

Figure 2.3: For distinct parameter configurations and values of $\omega = |S|/I$, a box plot summarizing the distribution of $F(q, \mathcal{Y}) - H[q(\cdot)]$ for 32 initial conditions is shown. For every initial condition, the standard variational Bayes algorithm is executed until convergence. Then, the corresponding stochastic variational algorithms are run for as much time as the variational algorithm.

generalize, Figure 2.3 supports the idea that the selection of the step sizes greatly influences the efficiency of the algorithm. The optimal choice of these parameters is problem specific, both the network size and the expected community sizes should be considered when selecting

these tuning parameters for the algorithm.

The second topic we address is the ability of the model of recovering the underlying community structure, comparing the performance of the two computational methods available. For this purpose we fit the model and obtain $100,000$ samples from the posterior distribution using the MCMC algorithm of Section 1.2, as well as 32 runs of the stochastic variational algorithm setting $\kappa = 0.6$, $\tau = 1$ and $|S| = \omega I$ with $\omega = 0.25$.

With respect to the execution time of these algorithms, an average run of a C implementation of the stochastic variational algorithm for this dataset takes 15 seconds on a standard laptop with 8GB of RAM and a 2.66GHz Intel Core i7 processor; in contrast, it takes approximately 4.5 hours for the MCMC under similar conditions. Although this is not intended as a formal algorithm efficiency comparison, it gives a rough idea of the significant difference in computational time between the two approaches.

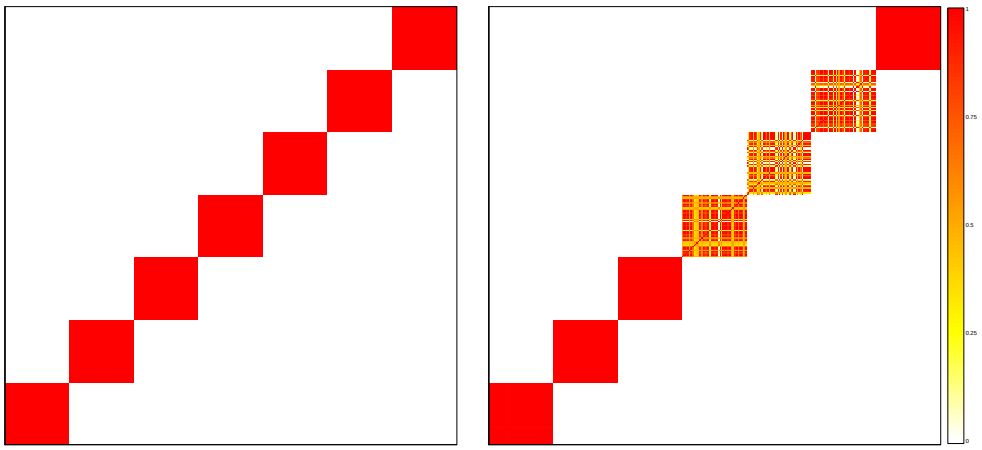

Figure 2.4: (Left) Monte Carlo estimates of pairwise posterior probabilities of same community. That is, for every pair (i, j), $Pr(\xi_i = \xi_j \mid \mathcal{Y})$ is shown $ARI = 1$. (Right) Variational approximation $q(\xi_i = \xi_j)$ $ARI = 0.9$.

The left panel of Figure 2.4 shows the resulting mean pairwise incidence matrix for the MCMC; that is, for every pair (i, j) of vertices in the network, this matrix shows the posterior probability of i and j belonging to the same community *i.e,* $p(\xi_i = \xi_j \mid \mathcal{Y})$. From this figure, it is clear that the model is capable of fully recovering the underlaying community structure in the network. In turn, the right panel of Figure 2.4 shows the inferred cluster structure from $q(\xi_i = \xi_j)$ for the stochastic variational approximation achieving the highest lower bound out of 32 runs. Here, it can be seen that the variational approximation recovers most of the underlying community structure in the network, although higher levels of uncertainty are observed; particularly in the fourth, fifth and sixth communities. In this case, given that true vertex partition is known, we are able to evaluate the resulting community structure using the adjusted Rand index (Hubert and Arabie, 1985). The ARI is a chance-corrected measure of similarity between two clusters based pairwise agreements. Although negative values are possible, under this metric a value of zero indicates that the is no more agreement than that expected by chance, while a ARI of one signifies that the two partitions are equivalent. For the results shown in Figure 2.4, the ARI between the true vertex partition and that obtained from applying the clustering procedure proposed by Lau and Green (2007) with relative error cost of 0.5 to the mean posterior co clustering probabilities is 1, while the corresponding ARI using the variational approximation takes the value 0.9.

Next, we test the stochastic variational algorithm is through its predictive performance. To this end, we carried a twenty-fold cross validation exercise where, for each validation subset, the *receiver operating characteristic* (ROC) curve, and its corresponding

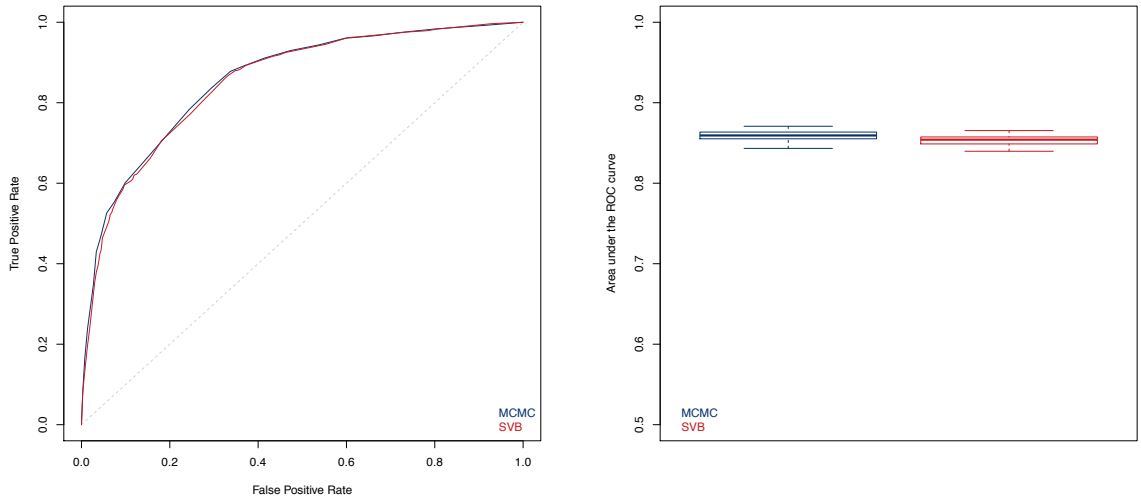

Figure 2.5: (Left) Receiver operating characteristic curves for a typical validation subset. (Right) Boxplots of the area under the ROC curve for MCMC and SVB algorithms.

area under the curve (AUC), for the SVB algorithm is compared to that of the MCMC. These results are shown in Figure 2.5, where it is possible to observe that, in this case, the stochastic variational algorithm performs reasonably close to the MCMC in terms of prediction. Furthermore, it is worthwhile mentioning that even in the case where the true underlying community structure is fully recovered, the ROC curve is not necessarily that of the perfect classifier; this is because in this dataset some of the community interaction probabilities are, in fact, close to 0.5.

Figure 2.6 compares the evolution of the variational and stochastic variational algorithms. It shows the bound calculated over the complete network as a function of elapsed execution time. From this figure it is possible to observe that, in this case, the stochastic variational optimization performs reasonably close to the MCMC algorithm, supporting the findings of Figures 2.4 and 2.5. More interestingly, Figure 2.6 suggests that even in a

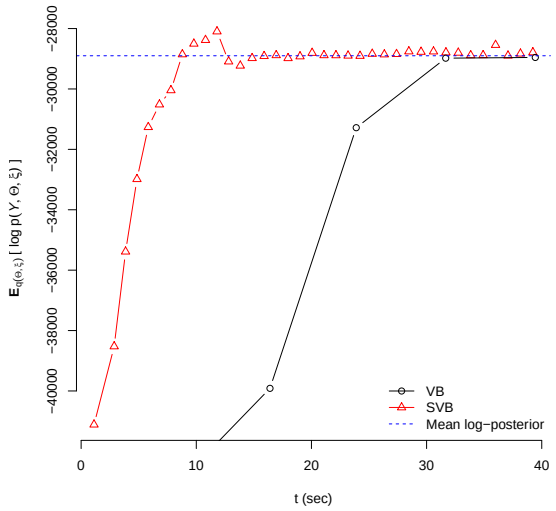

Figure 2.6: Evolution of $\mathbb{E}_{q(\Theta,\boldsymbol{\xi})}[\log P(\mathcal{Y},\Theta,\boldsymbol{\xi})]$ with respect to execution time in seconds. As before, $\omega = 0.25$, $\tau = .6$ and $\kappa = 1$ in the stochastic variational algorithm.

case where the variational algorithm converges closely to the posterior distribution, since the stochastic algorithm climbs faster in the early stages, if the algorithm cannot be run until convergence, the stochastic version may be preferable. The horizontal dashed line in this plots corresponds to the value obtained by averaging $\log P(\mathcal{Y},\Theta,\boldsymbol{\xi})$ evaluated at the posterior samples. This quantity, differs from $\mathbb{E}_{q(\Theta,\boldsymbol{\xi})}[\log P(\mathcal{Y},\Theta,\boldsymbol{\xi})]$, as the former averages over the posterior while the later takes the expectation with respect to the variational distribution. However, these two quantities are in a similar scale and, therefore, give an idea of how well the variational approximates the posterior distribution.

Now, we explore a second simulated dataset with a similar structure, *i.e.* $I = 350$ individuals evenly split among $K^\star = 7$ communities, but with interaction probabilities that are less clearly differentiated. Specifically, as can be seen in Figure 2.7, interaction probabilities for the sixth community have been have been selected near the average of

those corresponding to communities five and seven.

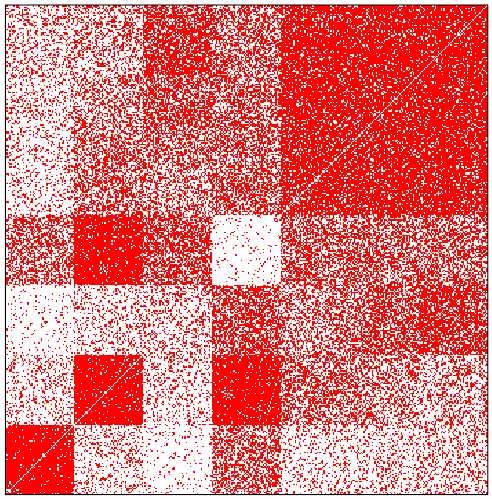

Figure 2.7: Adjacency matrix for second simulated dataset. $Y_{i,j} = 1$ is represented by a red dot, while a lack of interaction is shown in white.

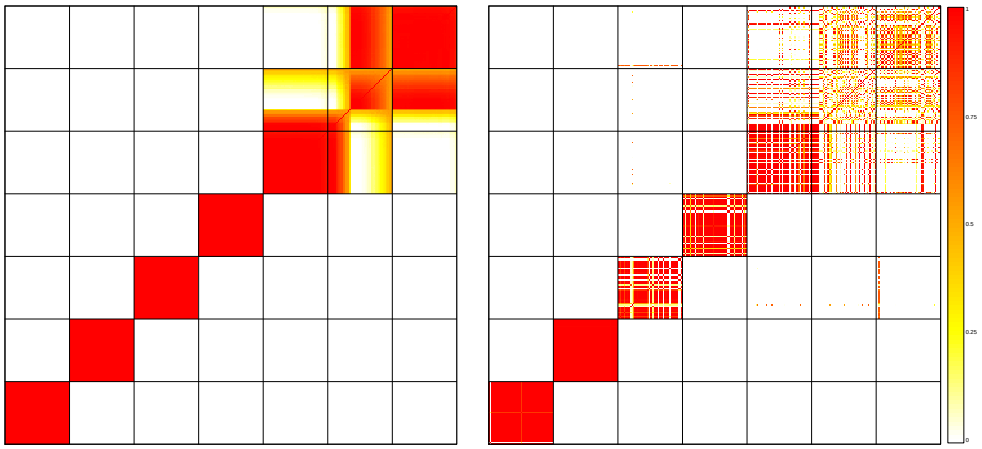

Figure 2.8: (Left) Monte Carlo estimates of pairwise posterior probabilities of same community. That is, for every pair (i, j), $Pr(\xi_i = \xi_j \mid \mathcal{Y})$ is shown $ARI = 0.81$ (Right) Variational approximation $q(\xi_i = \xi_j)$ $ARI = 0.67$.

Figure 2.8 shows the inferred community structures for this dataset under both

computational approaches. In this case solid black lines are added to facilitate the recognition of the true community structure. From this figure it can be seen that both algorithms have difficulties separating the individuals in the sixth community. While the MCMC places part of these vertices in community five and the rest on community seven with high probability, the stochastic variational algorithm tends to average over the modes mixing individuals from the three communities into a single cluster. The Adjusted Rand Index between these point estimates to the partition and the true community structure is 0.81 and 0.67 for the MCMC and variational approximation respectively.

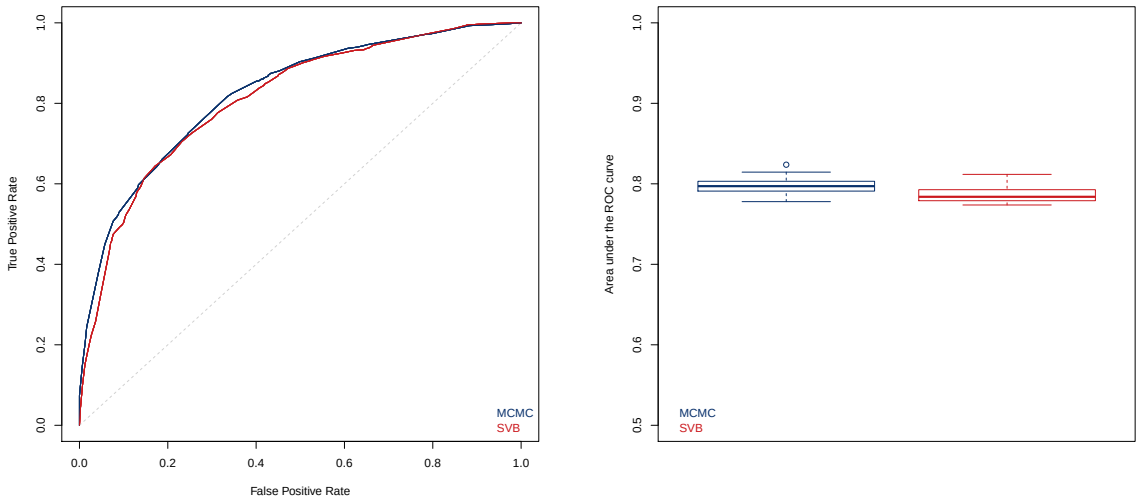

Figure 2.9: (Left) Receiver operating characteristic curves for a typical validation subset. (Right) Boxplots of the area under the ROC curve for MCMC and SVB algorithms.

In turn, from Figure 2.9 it can be seen that the predictive performance is affected similarly for both algorithms. In both cases there is a reduction in the AUC level, and a small increase in the variability is observed.

2.2.2 Coauthorship network

In this section we consider the coauthorship network by Newman (2006). In this network, vertices represent ($I = 379$) scientists working in the field of *Network Science* and an edge connecting two vertices indicates the existence of collaboration(s) among those scientists. This dataset includes a total of 914 publications up to early 2006, and is a subset of a larger network constructed from the bibliographies of the two reviews Newman (2003) and Boccaletti et al. (2006).

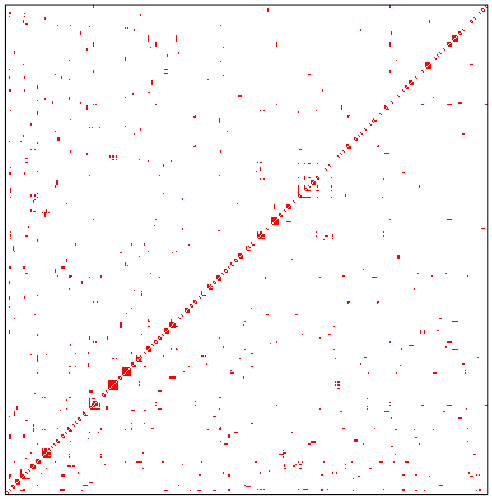

Figure 2.10: Raw data for the collaboration network of Newman (2006). In this network vertices represent authors of scientific papers in the field of network science, and an edge represents the existence of at least one collaboration between those authors.

In this case, since the number communities in the network $K^\star$ is unknown, we choose the maximum number of communities in the model as $K = 30$, trying to overestimate number of communities in the network while, simultaneously, keeping the model computationally tractable. As in the previous example, the hyperparameters in the model

are fixed to the values $\alpha = a = b = 1$, and in the stochastic variational algorithm the batch sizes are taken as $|S| = \omega I$ with $\omega = 0.25$, while the step sizes sequence is defined by taking $\kappa = 0.6$ and $\tau = 1$.

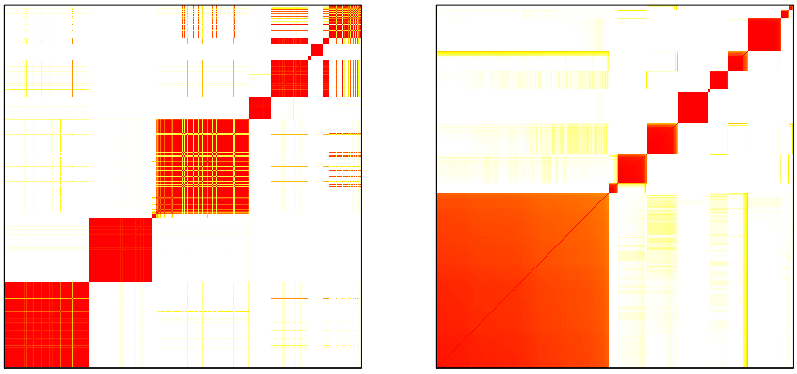

Figure 2.11: Pairwise incidence matrices under MCMC (left) and stochastic variational approximation (right).

Figure 2.11 shows the results generated by the MCMC and the stochastic variational algorithm. The left panel shows the mean pairwise incidence matrix resulting from $200,000$ samples of the posterior distribution taken after a burn in period of $300,000$ iterations. With a different ordering of the vertices, the right panel shows the inferred community structure using the best out of 32 runs of the stochastic variational approximation. Here, is possible to see that, in contrast to the simulated dataset, there is not a clear correspondence between the communities found by the two methods. Figure 2.12 shows the overlap between the partitions derived from both algorithms. This plot displays the pairwise incidence matrix of the stochastic variational approximation under the ordering obtained from the MCMC. Thus, confirming that the two approaches lead to qualitatively different solutions; particularly, white spaces in the anti-diagonal blocks correspond to MCMC

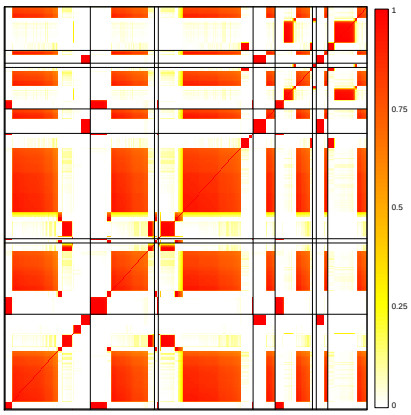

Figure 2.12: Overlap in community structure from the two methods. This figure plot the incidence matrix from the stochastic variational algorithm using the ordering from the MCMC.

communities being separated by the stochastic variational approximation, while red areas outside of the main anti diagonal blocks indicate grouping of individuals in distinct MCMC communities. Lastly, Figure 2.13 presents the adjacency matrix (raw data) permuted to

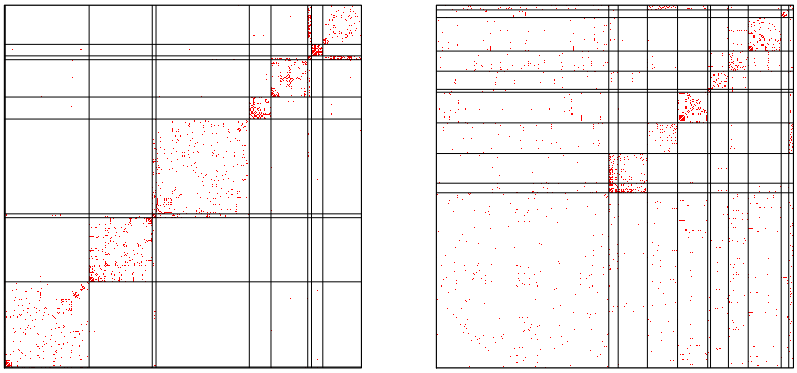

Figure 2.13: Adjacency matrix ordered with respect to MCMC (left) and stochastic variational approximation (right) communities.

show the corresponding elicited communities. The lines correspond to the clustering estimate that result from the method proposed by Lau and Green (2007) with relative error

cost of 0.5.

Next, we evaluate the out-of-sample performance of the model. A randomly selected subset consisting of 5% of the potentially observed links are treated as missing values and predicted in the same way as in Section 2.2.1. Figure 2.14 shows the ROC curves and the boxplots corresponding to the area under the ROC curves for the twenty test sets under both the MCMC and the SVB algorithm. From this figure it is clear that, for this dataset, the MCMC substantially outperforms the stochastic variational algorithm in terms of predictive performance.

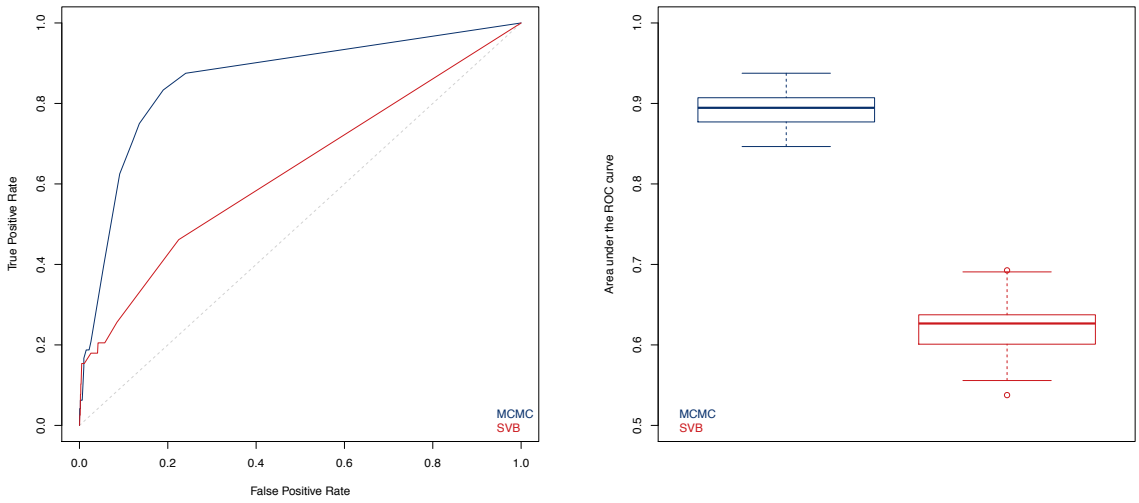

Figure 2.14: (Left) Receiver operating characteristic curves for a typical validation subset. (Right) Boxplots of the area under the ROC curve for MCMC and SVB algorithms.

2.2.3 Internet Movie Database network

To asses its scalability, we tested the stochastic variational algorithm using a network constructed from a subset of the *Internet Movie Database*. This network consists of $9,647$ vertices representing movies and $1,050,162$ edges indicating whether two movies

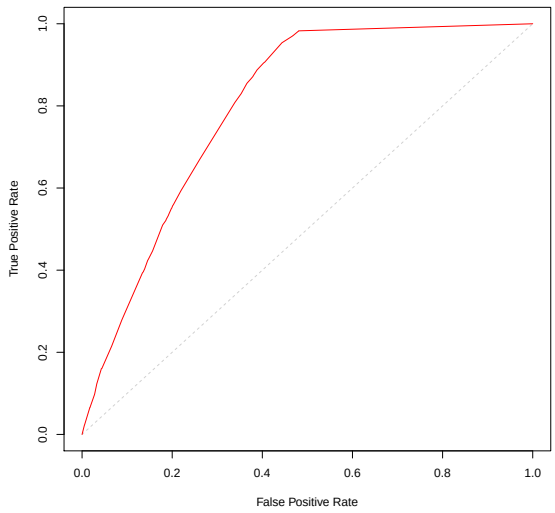

Figure 2.15: Receiver operating characteristic curve for a randomly selected validation subset with the IMDb dataset.

share at least one cast member. In this case the algorithm takes approximately eight hours to reach convergence in 60 iterations with $K = 40$, $\omega = 0.25$, $\kappa = 0.6$ and $\tau = 1$. The results suggest the existence of $K^\star = 25$ different communities. Interestingly, we observed an association between the resulting communities and the IMDb genre classification with, for example, a third of documentaries in the data being clustered together to form a single community. Additionally, we evaluate the predictive accuracy of the model under this dataset using cross validation. Figure 2.15 shows the corresponding receiver operating characteristic curve which, in this case, attains an area under the curve of 80%.

2.2.4 Simulated IMDb dataset

To further explore the performance of the stochastic variational algorithm for a dataset with dimension such as the one introduced in Section 2.2.3, we created a simulated

dataset with similar characteristics. Specifically, we take the recovered community structure and mean variational community parameters and set them as the ground truth. Then, using these values, we randomly generate a new set of edges among the same $9,647$ vertices. The resulting network consists of $1,051,101$ connections.

Setting $K = 40$, $\kappa = 0.6$ and $\tau = 1$, we test the algorithm for two different values of the proportion of subsampled vertices, $\omega = 0.15$ and $\omega = 0.25$. The ARI corresponding to the highest achieving lower bound out of 32 runs are shown in Table 2.1. There it can be seen that in both cases the algorithm recovers the underlying structure only partially, and that, in this example, a greater proportion of co-clustering is recovered when $\omega = 0.25$.

ARI	$\omega = .15$	$\omega = .25$
truth	0.495	0.7

Table 2.1: Adjusted Rand index comparing inferred community structure to the true partition used for data simulation.

Next, Figure 2.16 explores the quality of the stochastic variational approximation trough its predictive performance. The green ROC curve in this plot is calculated by predicting a randomly selected subset of $25,00$ interactions in the network using the true partition and true interaction probabilities in the data. In turn, the red and pink ROC curves are obtained using the inferred community structure and interaction probabilities from the stochastic variational algorithm with $\omega = 0.15$ and $\omega = 0.25$ respectively. Here it can be seen that the predictive performance is fairly close to prediction under the truth in both cases, with $\omega = 0.25$ slightly outperforming $\omega = 0.15$ in terms of the AUC.

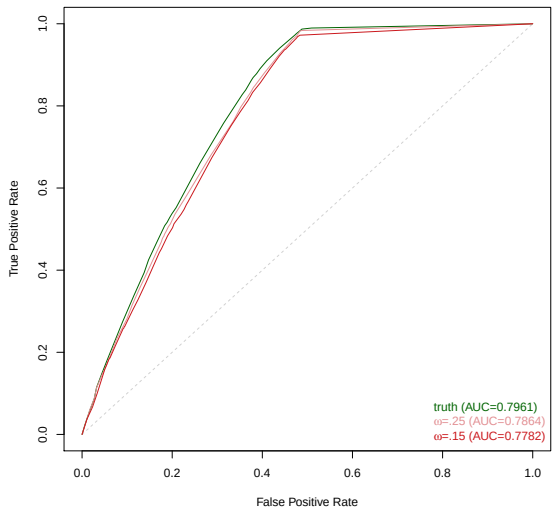

Figure 2.16: Receiver operating characteristic curves for a randomly selected validation subset with the simulated dataset.

2.3 Discussion

In this section we have introduced a stochastic variational algorithm for the stochastic blockmodel. In light of the well known computational efficiency of variational methods, here we investigate into the quality of the approximation to the posterior distribution that this algorithm is able to produce; specifically, we explore both the inferential and predictive accuracy of this algorithm.

As a first step we note that the results for the stochastic variational algorithm are highly dependent on the value of the tuning parameters τ, κ and ω. Providing general guidance for the selection of these parameters is complicated as the results will typically depend on true number and size of structurally equivalent factions in the data. Thus, every application of the algorithm requires careful consideration of this topic. Furthermore, in the illustrations explored in this section a subsample size of at least 25% is required to

obtain adequate results, therefore limiting the gains in computational efficiency. Also, it is worthwhile mentioning that, despite stochastic gradient optimization helping avoiding local modes when compared to standard variational Bayes, in a multimodal parameter space, such as the one of the stochastic blockmodel, multiple runs from different starting points are still required.

Another characteristic that we have observed from the stochastic variational algorithm is "smoothing over modes" which in the setting of community detection means that the algorithm will have a tendency to miss communities that are not very clearly differentiated. Most importantly, we highlight the fact that when there is substantial uncertainty in the cluster structure the stochastic variational algorithm might be significantly outperformed by the MCMC.

In practice, the stochastic variational approximation has the advantage of being able to fit the stochastic blockmodel to networks where MCMC is simply infeasible, such as the example of Section 2.2.3. In these cases it is important for the user to proceed cautiously, being aware of the drawbacks of the algorithm.

Chapter 3

Identifying hierarchical structures in network data

As described in the introduction, networks frequently exhibit multilevel community structure by which we mean that they can be partitioned into groups of structurally equivalent vertices and, in turn, communities that exhibit similar interaction patterns across the network might be further clustered into supercommunities. A couple efforts have been made in the literature to study the topic of hierarchical structures in network data. In each case, however, the definition of concepts such as supercommunity or hierarchical structure has been slightly different. For example, Clauset et al. (2007) defines hierarchical organization as rooted binary tree with leaves representing the network vertices. Denoting $D = \{D_1, D_2, \ldots, D_{n-1}\}$ the internal nodes of the tree, they model directly in the space of dendrogram by taking a likelihood of the form

$$\mathcal{P}(\mathcal{Y} \mid D, \theta) = \prod_{i=1}^{n-1} \theta_i^{L_i(\mathcal{Y})R_i(\mathcal{Y})}(1-\theta_i)^{E_i - L_i(\mathcal{Y})R_i(\mathcal{Y})} \tag{3.1}$$

where E_i is the number of edges in the network with lowest common ancestor D_i, and L_i and R_i are the number of vertices to left and right of D_i respectively. Then, the profile likelihood $P\left(\mathcal{Y} \mid D, \hat{\theta}(D, \mathcal{Y})\right)$ is combined with a uniform prior to generate a pseudo-posterior distribution, $P\left(D \mid \mathcal{Y}\right)$), which is explored using a Metropolis-Hastings algorithm. This algorithm uses simple switch moves into single internal nodes sampled uniformly at random, attempting to explore the space of dendrograms. The result of this algorithm is a consensus dendrogram which defines a hierarchical structure among the vertices in the network. However, it is not straightforward to know how or where to "cut" the tree in order to obtain communities or higher level group structures.

In turn, Ho et al. (2012) define a model where vertices are sequentially clustered c_i according to a nested Chinese restaurant process (Blei et al., 2010), while interaction probabilities are derived from denominated *community compatibility* matrices $\mathbf{B}$. Specifically, each vertex is assigned a multiscale membership vector $\boldsymbol{\theta}_i$ according to probabilities drawn from a stick-breaking process (Sethuraman, 1994); then, for each possible interaction, $y_{i,j}$, the interaction probability is taken to be a function $S(\mathbf{B}, z_{i,j}, z_{j,i}, c_i, c_j)$ with $z_{i,j}$ following a Multinomial($\boldsymbol{\theta}_i$) and

$$S(\mathbf{B}, z_{i,j}, z_{j,i}, c_i, c_j) = \begin{cases} B_{h,h'} & \text{if } h \text{ and } h' \text{ have the same parent} \\ 0 & \text{otherwise} \end{cases}$$

where $h = c_i[\min\{z_{i,j}, z_{j,j}\}]$ and $h' = c_j[\min\{z_{i,j}, z_{j,j}\}]$. Inference in this model is carried out trough MCMC and, again, a consensus hierarchical structure $\mathbf{c} = \{c_1, \ldots, c_n\}$ can be found. Under $\mathbf{c}$ communities and higher order structures are naturally defined; however, it is important to note that communities obtained in this setting will be fundamentally different

from those in a stochastic blockmodel as interaction probabilities will be inhomogeneous within communities; thus, leading to group of vertices that are not necessarily structurally equivalent across the network.

In this chapter we introduce a simple extension of the stochastic blockmodel that allows for multi-level community detection, we describe a Markov chain Monte Carlo sampler as well as a variational algorithm for approximate posterior inference in the model, and we explore the model using simulated and real datasets.

3.1 Multilevel stochastic blockmodel

The stochastic blockmodel can easily be generalized in a hierarchical fashion and, thus, it naturally lends itself for the problem at hand; namely, the discovery of multilevel structures in networks. For this purpose the logit Gaussian prior structure is favored over the Beta model, as it provides a more convenient parametrization of the second level mixture through its mean parameters.

For each element in $\mathcal{Y}$ let

$$p(y_{i,j} \mid \theta_{\xi_i,\xi_j}) = \frac{(\exp\{\theta_{\phi(\xi_i,\xi_j)}\})^{y_{i,j}}}{1+\exp\{\theta_{\phi(\xi_i,\xi_j)}\}}$$

where $\theta_{\xi_i,\xi_j} = \log\left(\frac{\lambda_{i,j}}{1-\lambda_{i,j}}\right)$, so that the likelihood function is given by

$$p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}) = \prod_{k=1}^{K}\prod_{l=k}^{K} \frac{(\exp\{\theta_{k,l}\})^{s_{k,l}}}{(1+\exp\{\theta_{k,l}\})^{n_{k,l}}}. \tag{3.2}$$

Now, we assume that the community parameters come from a mean mixture of Gaussians

$$\theta_{k,l} \,|\, \eta_{\zeta_k,\zeta_l}, \sigma^2 \sim \mathcal{N}(\eta_{\phi(\zeta_k,\zeta_l)}, \sigma^2), \tag{3.3}$$

where the location parameters, $\eta_{r,s}$, are assumed conditionally independent from a common distribution

$$\eta_{r,s} \,|\, \mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2), \tag{3.4}$$

and where, once again, symmetry restrictions allow to consider only the subset

$$H = \{\eta_{r,s} : 1 \leq r \leq s \leq R,\ r, s \in \mathbb{N}\}.$$

As in the single level model, σ^2 controls the variability in the propensity of interactions between communities in a supercommunity and μ governs the overall density of the network; while, in the second level, τ^2 controls the dispersion of the mean of the community parameters.

The community indicators remain as in Section 1.1; namely,

$$Pr(\xi_i = k \mid w_k) = w_k; \qquad i = 1, 2, \ldots, I, \tag{3.5}$$

with

$$\boldsymbol{w} \sim \mathcal{D}ir\left(\boldsymbol{\alpha_w}\right). \tag{3.6}$$

and $\boldsymbol{\alpha_w} = \left(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$. In turn, the supercommunity indicators $\boldsymbol{\zeta}$ mimic the structure of the first level indicators, taking a Categorical distribution in the set $\{1, 2, \ldots, R\}$

$$Pr(\zeta_k = r \,|\, v_r) = v_r \qquad k = 1, 2, \ldots, K, \tag{3.7}$$

with weight vector $\boldsymbol{v}$ such that

$$\boldsymbol{v} \sim \mathcal{D}ir\left(\boldsymbol{\alpha_v}\right), \tag{3.8}$$

where $\boldsymbol{\alpha_v} = \left(\frac{\beta}{R}, \frac{\beta}{R}, \ldots, \frac{\beta}{R}\right)$.

Analogous to the single level model we have that, as $K \to \infty$,

$$Pr(K^\star = k \mid \alpha) = S(I,k)\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + I)}$$

and further, when $R \to \infty$,

$$Pr(R^\star = r \mid K^\star, \beta) = S(K^\star, r)\beta^r \frac{\Gamma(\beta)}{\Gamma(\beta + K^\star)}$$

Then, conditionally, the expected number of supercommunities can be approximated as

$$\mathbb{E}[R^\star \mid \beta, K^\star] \approx \beta \log\left(\frac{\beta + K^\star}{\beta}\right)$$

and the expected number of supercommunites is found to satisfy

$$\mathbb{E}[R^\star \mid \alpha, \beta] = \mathbb{E}[\mathbb{E}[R^\star \mid \beta, K^\star] \mid \alpha] \approx \beta \log\left(\frac{\beta + \alpha \log\left(\frac{\alpha+I}{\alpha}\right)}{\beta}\right) - \frac{\beta\alpha \log\left(\frac{\alpha+I}{\alpha}\right)}{2\left(\beta + \alpha \log\left(\frac{\alpha+I}{\alpha}\right)\right)^2},$$

the details of this derivation can be found in Appendix C.

Figure 3.1 shows the effect of the hyperparameters α and β in the number of occupied communities and supercommunities implied by the prior. It can be seen that, as it is usual, larger values of the concentration parameters favor a larger number of components at both levels. Note also that the standard stochastic blockmodel can then be recover by either setting $R = 1$ or letting $\beta \to 0$.

Finally, all hyperparameters are assumed independent a priori, that is

$$\pi(\mu, \sigma^2, \tau^2, \alpha, \beta) = \pi(\mu)\pi(\sigma^2)\pi(\tau^2)\pi(\alpha)\pi(\beta).$$

with conditionally conjugate hyperpriors in the community parameter's side

$$\mu \sim \mathcal{N}(\mu_0, \sigma_\mu^2), \quad \tau^2 \sim \mathcal{IG}(\alpha_\tau, \beta_\tau) \quad \text{and} \quad \sigma^2 \sim \mathcal{IG}(\alpha_\sigma, \beta_\sigma) \tag{3.9}$$

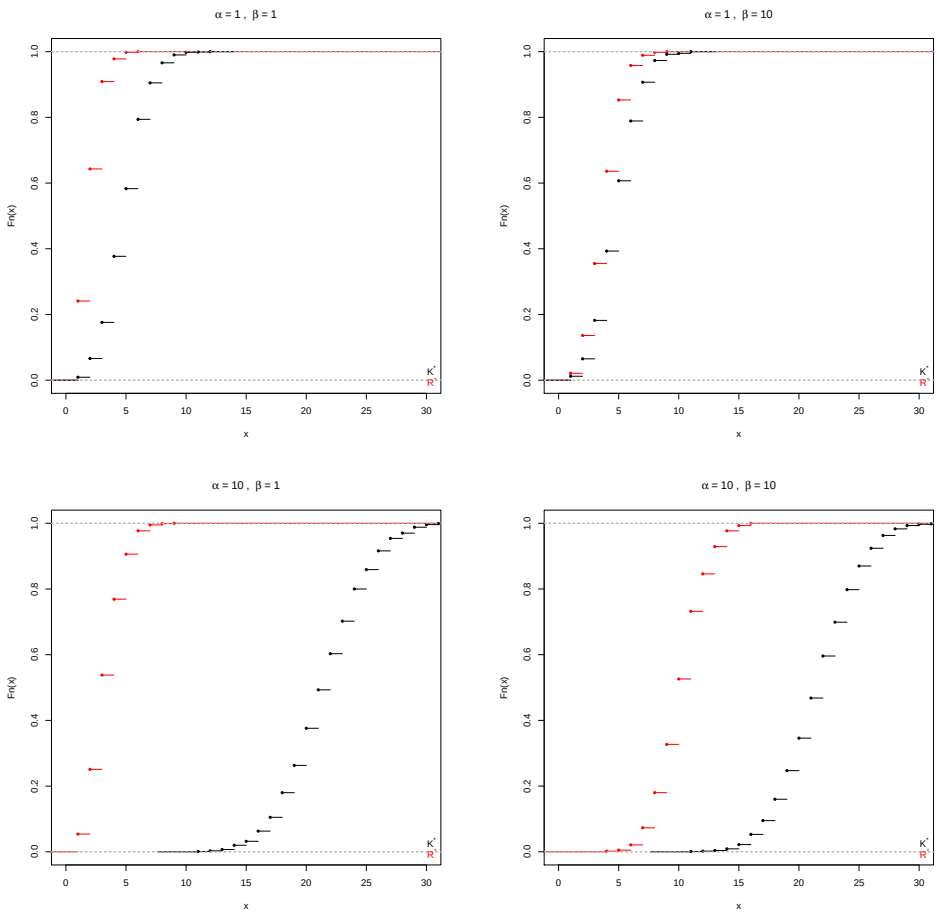

Figure 3.1: Prior CDF of effective number of communities $K^\star$ and supercommunities $R^\star$ under four different scenarios for the hyperparameters α and β.

and a Gamma hyperpriors for the concentration parameters

$$\alpha \sim \mathcal{G}(\alpha_\alpha, \beta_\alpha) \quad \text{and} \quad \beta \sim \mathcal{G}(\alpha_\beta, \beta_\beta). \tag{3.10}$$

3.2 Posterior inference using Markov chain Monte Carlo

Irrespective of the choice of the prior distributions the model described above does not lead to closed form posteriors and, thus, some form of approximation is required for inferential purposes. In this section a Markov chain Monte Carlo algorithm to generate

samples from the joint posterior distribution is derived for the model. As a first step notice that the form of the likelihood already suggests that the full conditionals for Θ are not members of a standard family of distributions. However, following Polson et al. (2013),

$$\frac{(\exp\{\theta_{k,l}\})^{s_{k,l}}}{(1+\exp\{\theta_{k,l}\})^{n_{k,l}}} = \exp\left\{\left(s_{k,l} - \frac{n_{k,l}}{2}\right)\theta_{k,l}\right\}\mathbb{E}\left[\exp\left\{-\frac{\theta_{k,l}^2}{2}\gamma_{k,l}\right\}\right]$$

where $\gamma_{k,l} \sim \mathcal{PG}(n_{k,l}, 0)$ is a Polya-Gamma random variable. Thus, augmenting the parameter space with $\Gamma = [\gamma_{k,l}]$ a matrix of a priori independent Polya-Gamma random variables the likelihood can be expressed as

$$p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}, \Gamma) \propto \exp\left\{\sum_{k=1}^{K}\sum_{l=k}^{K}\left(s_{k,l} - \frac{n_{k,l}}{2}\right)\theta_{k,l}\right\}\prod_{k=1}^{K}\prod_{l=k}^{K}\exp\left\{-\frac{\theta_{k,l}^2}{2}\gamma_{k,l}\right\}. \tag{3.11}$$

Denoting the set of all parameters in the model $\Upsilon = \left\{\Theta, \boldsymbol{\xi}, H, \boldsymbol{\zeta}, \mu, \sigma^2, \tau^2, \boldsymbol{w}, \alpha, \boldsymbol{v}, \beta, \right\}$, the augmented joint posterior satisfies

$$\begin{aligned}
p(\Upsilon, \Gamma \mid \mathcal{Y}) \propto{}& \exp\left\{\sum_{k=1}^{K}\sum_{l=k}^{K}\left(s_{k,l} - \frac{n_{k,l}}{2}\right)\theta_{k,l}\right\}\prod_{k=1}^{K}\prod_{l=k}^{K}\mathbb{E}\left[\exp\left\{-\frac{\theta_{k,l}^2}{2}\gamma_{k,l}\right\}\right] \\
&(\sigma^2)^{-\left(\frac{1}{4}K(K+1)+\alpha_\sigma+1\right)}\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{l=k}^{K}(\theta_{k,l}-\eta_{\phi(\zeta_k,\zeta_l)})^2 - \frac{\beta_\sigma}{\sigma^2}\right\}\exp\left\{-\frac{1}{2\sigma_\mu^2}(\mu-\mu_0)^2\right\} \\
&(\tau^2)^{-\left(\frac{1}{4}R(R+1)+\alpha_\tau+1\right)}\exp\left\{-\frac{1}{2\tau^2}\sum_{r=1}^{R}\sum_{s=r}^{R}(\eta_{r,s}-\mu)^2 - \frac{\beta_\tau}{\tau^2}\right\}\prod_{k=1}^{K} w_k^{\frac{\alpha}{K}+n_k-1}\prod_{r=1}^{R} v_r^{\frac{\beta}{R}+m_r-1} \\
&\frac{\Gamma(\alpha)}{\left[\Gamma\left(\frac{\alpha}{K}\right)\right]^K}\alpha^{\alpha_\alpha-1}\exp\left\{-\beta_\alpha\alpha\right\}\frac{\Gamma(\beta)}{\left[\Gamma\left(\frac{\beta}{R}\right)\right]^R}\beta^{\alpha_\beta-1}\exp\left\{-\beta_\beta\beta\right\}\pi(\Gamma)
\end{aligned} \tag{3.12}$$

where, for all $r \in \{1, 2, \ldots, R\}$, $m_r = \sum_{\mathcal{T}_r} 1$ with $\mathcal{T}_r = \{k : \zeta_k = r\}$.

From (C.4) we derive an MCMC algorithm that allows us to obtain sampling-based approximate inference for the model. The main ideas behind this algorithm are summarized in what follows, while the details can be found in Appendix A.

First, from the form of the augmented likelihood (3.11) it can be anticipated that the community parameters are conditionally conjugate given the auxiliary variables. Thus,

the elements of Θ are sample from their corresponding Gaussian full conditional distribution. Importantly, the full conditional distribution for the auxiliary parameters remains in the Polya-Gamma family and can, therefore, be sampled as described in Polson et al. (2013).

The indicators are sampled from their Categorical full conditional distributions both for the communities and the supercommunities. At the supercommunity level the clustering probabilities are affected from the data trough the term

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{l=k}^{K}(\theta_{k,l}-\eta_{\phi(\zeta_k,\zeta_l)})^2\right\}$$

which plays a role equivalent to that of the likelihood in the community level.

The full conditional for $\eta_{r,s}$ is also Gaussian with mean and precision parameters that can be expressed as linear combinations of the prior mean μ and the proportion of observed interactions among the vertices in supercommunities r and s. In turn, if conditionally conjugate priors are assumed, the full conditional distribution for the mean μ and variance σ^2 and τ^2 hyperparameters are straightforward. Namely, a Gaussian and Inverse Gamma distributions respectively.

Finally, the concentration parameters α and β can be sampled following ideas from Escobar and West (1995).

3.3 Posterior inference using variational Bayes

In this section we introduce a variational Bayes algorithm to approximate the posterior distribution of the multilevel stochastic blockmodel. Again, we discus the main ideas behind the algorithm here and leave the details to Appendix B.

Similar to its full conditional distribution, the variational distribution for the elements of Θ is not member of a standard family of distributions. In order to overcome this issue, a first alternative would be to introduce Polya-Gamma auxiliary variables in the form of (3.11); this approach conduces to a Gaussian update for the elements of Θ, but translates into non-standard updates for the auxiliary variables. In turn, these updates could be handled using, for example, ideas from non-conjugate variational message passing (Knowles and Minka, 2011). However, this approach introduces two additional sources of error to the variational approximation. Namely, it uses an approximate solution to the variational distribution for the auxiliary variables, and it gives an approximation for the posterior distribution of the extended set of parameters whose marginal does not necessarily minimize the Kullback-Leibler divergence to the original posterior distribution. Instead, we relax the evidence lower bound following the approach of Jaakkola and Jordan (2000), which yields an exact solution to the original variational problem. To this end, using a first order Taylor expansion around $\gamma_{\phi(\xi_i,\xi_j)} \in \Re$, it is easily seen that

$$p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}) \geq \tilde{p}(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}, \Gamma) \equiv \prod_{i=1}^{I-1} \prod_{j=i+1}^{I} \exp\left\{y_{i,j}\theta_{\phi(\xi_i,\xi_j)}\right\} S\left(\gamma_{\phi(\xi_i,\xi_j)}\right)$$
$$\times \exp\left\{\frac{-\left(\theta_{\phi(\xi_i,\xi_j)} + \gamma_{\phi(\xi_i,\xi_j)}\right)}{2} + \lambda\left(\gamma_{\phi(\xi_i,\xi_j)}\right)\left(\theta^2_{\phi(\xi_i,\xi_j)} - \gamma^2_{\phi(\xi_i,\xi_j)}\right)\right\},$$

where $\lambda(x) = \frac{1}{2x}\left(S(x) - \frac{1}{2}\right)$, $S(x) = (1 + \exp\{-x\})^{-1}$, and the equality holds whenever $\theta^2_{k,l} = \gamma^2_{k,l}$ for all $k \leq l$. Then, the relaxed lower bound is given by

$$\tilde{F}(q, \mathcal{Y}) \equiv \mathbb{E}_{q(\Theta,\boldsymbol{\xi})}\left[\log \tilde{p}(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}, \Gamma)\right] + \mathbb{E}_{q(\Upsilon)}\left[\log p(\Upsilon)\right] + H[q]$$
$$\leq \mathbb{E}_{q(\Theta,\boldsymbol{\xi})}\left[\log p(\mathcal{Y} \mid \Theta, \boldsymbol{\xi})\right] + \mathbb{E}_{q(\Upsilon)}\left[\log p(\Upsilon)\right] + H[q] = F(q, \mathcal{Y}).$$

Maximization of $\tilde{F}(q, \mathcal{Y})$ leads to an approximation of the variational distribution for the

community parameters with the form

$$\log \tilde{q}(\theta_{k,l}) = -\frac{1}{2}\left\{2\lambda(\gamma_{k,l})\mathbb{E}_{q(\boldsymbol{\xi})}\left[n_{k,l}\right] + \mathbb{E}_{q(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\right\}\theta_{k,l}^2 + \left\{\mathbb{E}_{q(\boldsymbol{\xi})}\left[s_{k,l}\right] - \frac{\mathbb{E}_{q(\boldsymbol{\xi})}\left[n_{k,l}\right]}{2} + \mathbb{E}_{q(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\mathbb{E}_{q(H,\boldsymbol{\zeta})}\left[\eta_{\phi(\zeta_k,\zeta_l)}\right]\right\}\theta_{k,l} + C, \quad (3.13)$$

a Gaussian kernel, while maximizing $\mathbb{E}_{q(\Theta,\boldsymbol{\xi})}\left[\log \tilde{p}(\mathcal{Y} \mid \Theta, \boldsymbol{\xi}, \Gamma)\right]$ gives that the optimal auxiliary parameters satisfy $\gamma_{k,l}^2 = \mathbb{E}_{q(\theta_{k,l})}\left[\theta_{k,l}^2\right]$.

The variational distributions for most of the parameters in the model take the same form as their full conditional counterpart. The expectations involved in the calculation of the variational parameters can then be found in closed form and a (iterative) variational Bayes algorithm can be implemented. The only notable exception is given by the variational distribution of the concentration parameters α and β, since the Gamma distribution is not conjugate in this case. However, using a first order approximation to $\log \Gamma(x)$ we are able to approximate the variational distributions $q^\star(\alpha)$ and $q^\star(\beta)$ with Gamma distributions closely related to those of Escobar and West (1995).

3.4 Evaluation

In this section we illustrate the model and compare the performance of the two algorithms described in sections 3.2 and 3.3.

3.4.1 Simulated data

As a first step, we make use of the simulated dataset shown in Figure 3.2, this network is constructed with $I = 140$ individuals evenly split into $K^\star = 7$ communities. In

turn, these communities are split into $R^\star = 2$ supercommunities formed by the first four and the last three communities respectively. Our goal is to explore the the performance of the model in recovering community structure in a case in which the ground truth is known.

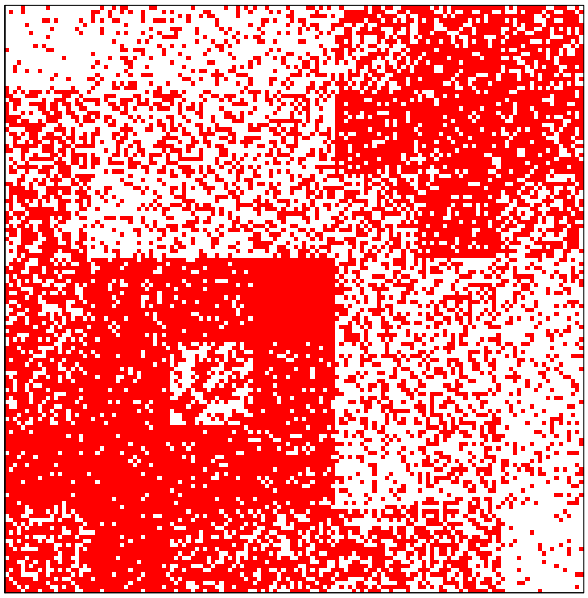

Figure 3.2: Image representation of the adjacency matrix. Here actors in the network are placed along the horizontal and vertical axis. $y_{i,j} = 1$ is represented by a red dot, while a lack of interaction is shown in white.

Choosing $K = 20$ and $R = 10$, and the rest of the hyperparameters as $\mu_0 = 1$, $\sigma_\mu^2 = 1$, $\alpha_\tau = \alpha_\sigma = 2$, $\beta_\tau = \beta_\sigma = 1$, $\alpha_\alpha = \beta_\alpha = \alpha_\beta = \beta_\beta = 1$, the top two panels of Figure 3.3 show the MCMC posterior pairwise co-membership probabilities for communities and supercommunities respectively. That is, for every pair (i, j), $Pr(\xi_i = \xi_j \mid \mathcal{Y})$ is shown in the left panel, while the right side shows $Pr(\zeta_{\xi_i} = \zeta_{\xi_j} \mid \mathcal{Y})$. From here it can be observed that the model is capable of recovering the underlying community structure with little uncertainty

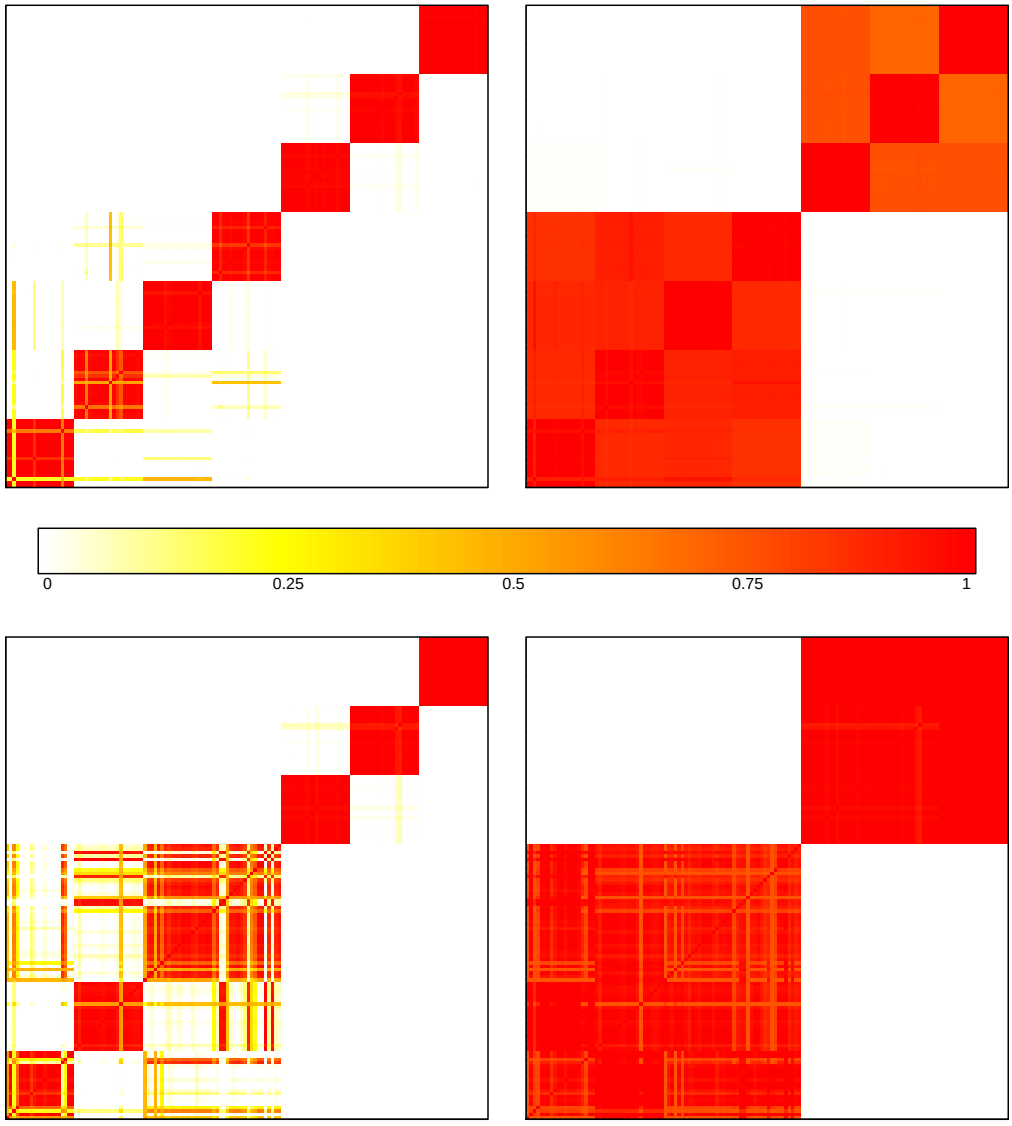

Figure 3.3: Community estimates for simulated data. **Top:** Monte Carlo estimates of pairwise posterior probabilities of same community, $Pr(\xi_i = \xi_j)$, ARI=0.99 (left), and supercommunity, $Pr(\zeta_{\xi_i} = \zeta_{\xi_j})$, ARI=1 (right). **Bottom** Variational approximations $q(\xi_i = \xi_j)$, ARI=0.75 (left), and $q(\zeta_{\xi_i} = \zeta_{\xi_j})$ ARI=1 (right).

in both levels, achieving an ARI of 0.99 and 1 respectively. In turn, the bottom panels of Figure 3.3 show the respective results for the solution achieving the highest lower bound out of 32 parallel runs of the variational approximation. In this case the algorithm is able

to learn most of the structure on the data, although it is worthwhile noticing the higher levels of uncertainty and the fact that, in this particular solution, communities three and four are not well discerned. In this case the ARI for the lower level is 0.75, while the ARI between the supercommunities and the true partition at this level is 1.

Figure 3.4 shows the evolution of the evidence lower bound as a function of execution time, and the mean posterior likelihood from the MCMC after $100,000$ posterior samples have been obtained in approximately $4,500$ seconds (75 minutes).

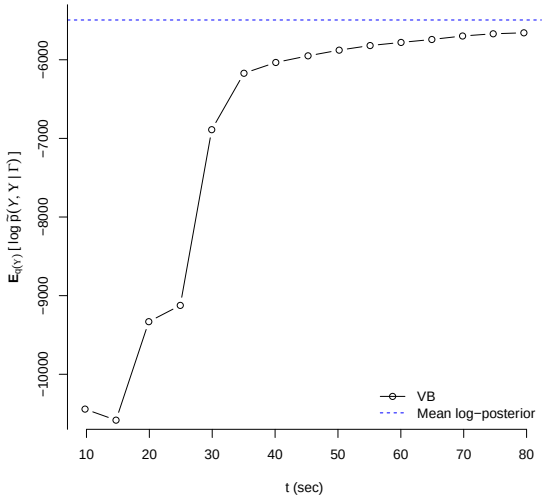

Figure 3.4: Evolution of the lower bound as a function of execution time.

Looking deeper into the results from the MCMC, Figure 3.5 shows the prior and posterior distributions associated with the parameters σ^2, α and β. These plots show reasonable agreement between the prior and posterior, although the data appears to pull towards a larger number of communities than those suggested by the prior. Also, in order to investigate the sensitivity to $\pi(\sigma^2)$, we tested taking $\alpha_\sigma = 2$ and $\beta_\sigma = 10$, as well as $\alpha_\sigma = 2$

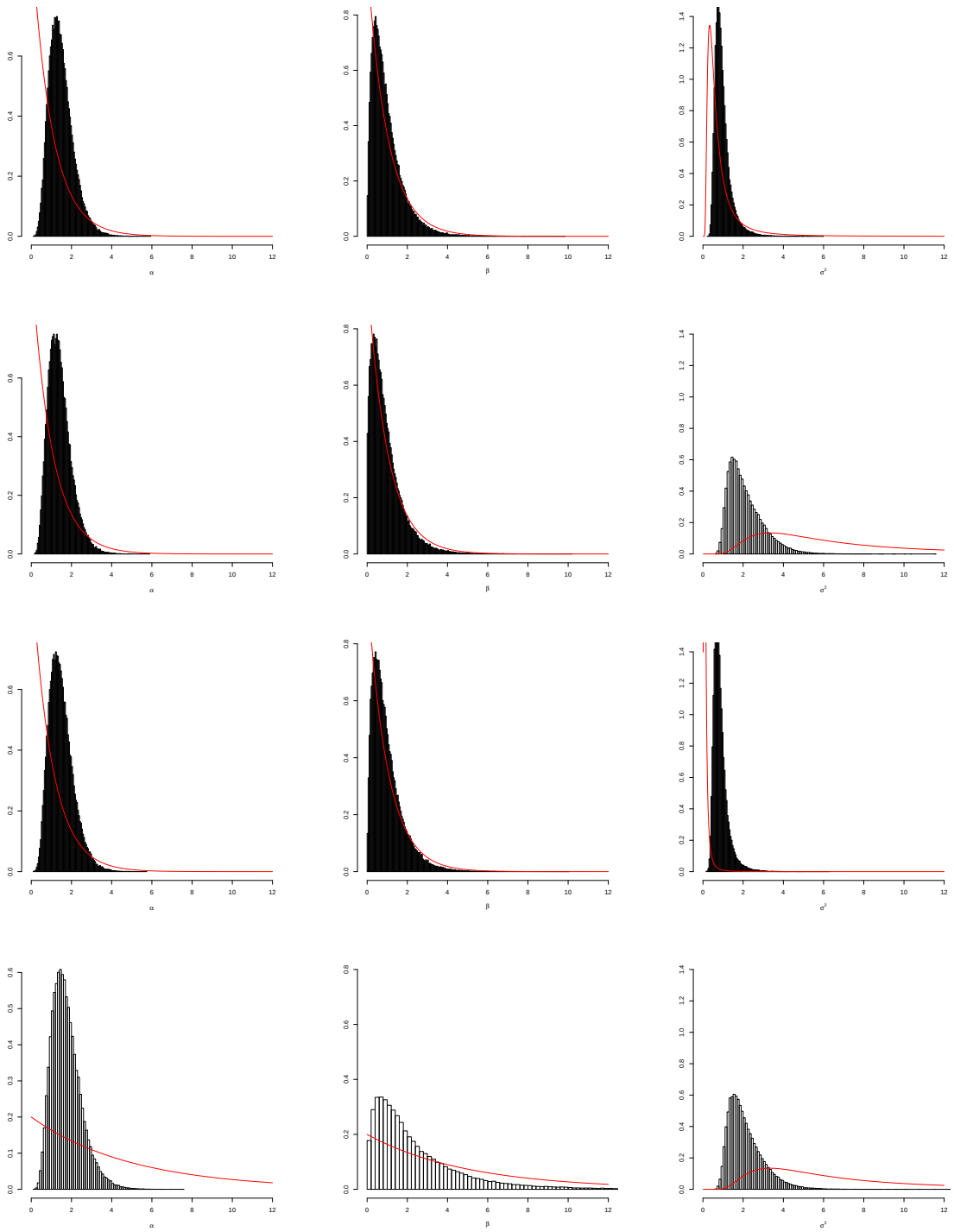

Figure 3.5: prior (line) and posterior (histogram) distributions for the hyperparameters α (left), β (center) and σ^2 (right) for the simulated dataset under (a) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,1)$ (b) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,10)$ (c) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,0.1)$ (d) $\alpha, \beta \sim Exp(5)$ and $\sigma^2 \sim \mathcal{IG}(2,10)$.

and $\beta_\sigma = 0.1$. The corresponding posteriors can be seen in the bottom three rows of Figure 3.5. Furthermore, we note that inference in the community structure (not shown) remained unaffected.

Lastly, we look at the results of clustering the vertices in the network using an agglomerative approach based on modularity maximization (Clauset et al., 2004). Figure 3.6 shows the resulting dendrogram with colors corresponding to the true community structure. From this figure it can be seen that the algorithm identifies subgroups in the community structure, but fails to identify whole communities. This is due, in part, to the fact that modularity maximization will tend to find groups of vertices with assortative mixing patterns, but will not uncover dissasortative structures such as the third community.

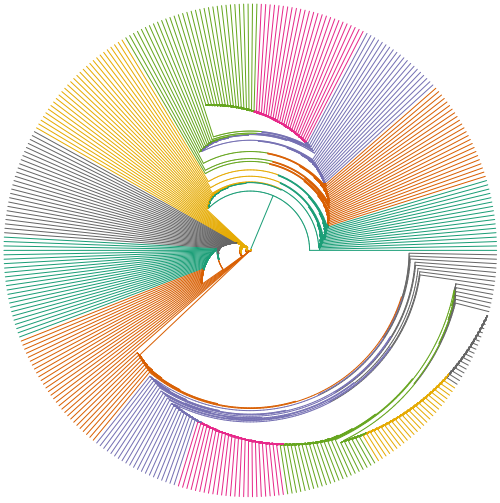

Figure 3.6: Hierarchical structure from agglomerative clustering for simulated dataset.

3.4.2 Coauthorship network

As a second illustration we consider again the co-autoship network by Newman (2006) introduced in Section 2.2.2. As a reminder, in this network the vertices represent ($I = 379$) authors of scientific papers in the field of network science and the existence of an edge between two vertices indicates that those authors have collaborated in at least in one of the 914 publications included in the network.

Figure 3.7 shows the structure recovered with $K = 100$, $R = 15$ and hyperparameters $\mu_0 = 1$, $\sigma_\mu^2 = 1$, $\alpha_\tau = \alpha_\sigma = 2$, $\beta_\tau = \beta_\sigma = 1$, $\alpha_\alpha = \beta_\alpha = \alpha_\beta = \beta_\beta = 1$, from the MCMC (top) and, with a different ordering of the vertices, for the variational Bayes (bottom) for communities (left) and supercommunities (right) respectively. From these plots it can be seen that, in this case, the inferred structure is significantly different between the two methods. In particular, the MCMC finds a much larger number of smaller communities, with most of the clustering occurring in the second level. This partition appears to be consistent with a structure where scientist form close-knit small research groups which, in turn, form four supercommunities with higher level of collaboration. Instead, the variational approximation finds five larger communities but does not capture any hierarchical structure. In order to asses the proximity of these two solutions, Figure 3.8 presents the overlap between the supercommunities obtained from the MCMC and the communities found by the variational algorithm. That is, the variational estimates for the the first level pairwise co-clustering probabilities are shown displayed under the optimal ordering obtained with the MCMC. From this figure it is interesting to note that although the communities from the variational algorithm are broken into different supercommunities in the MCMC, a good

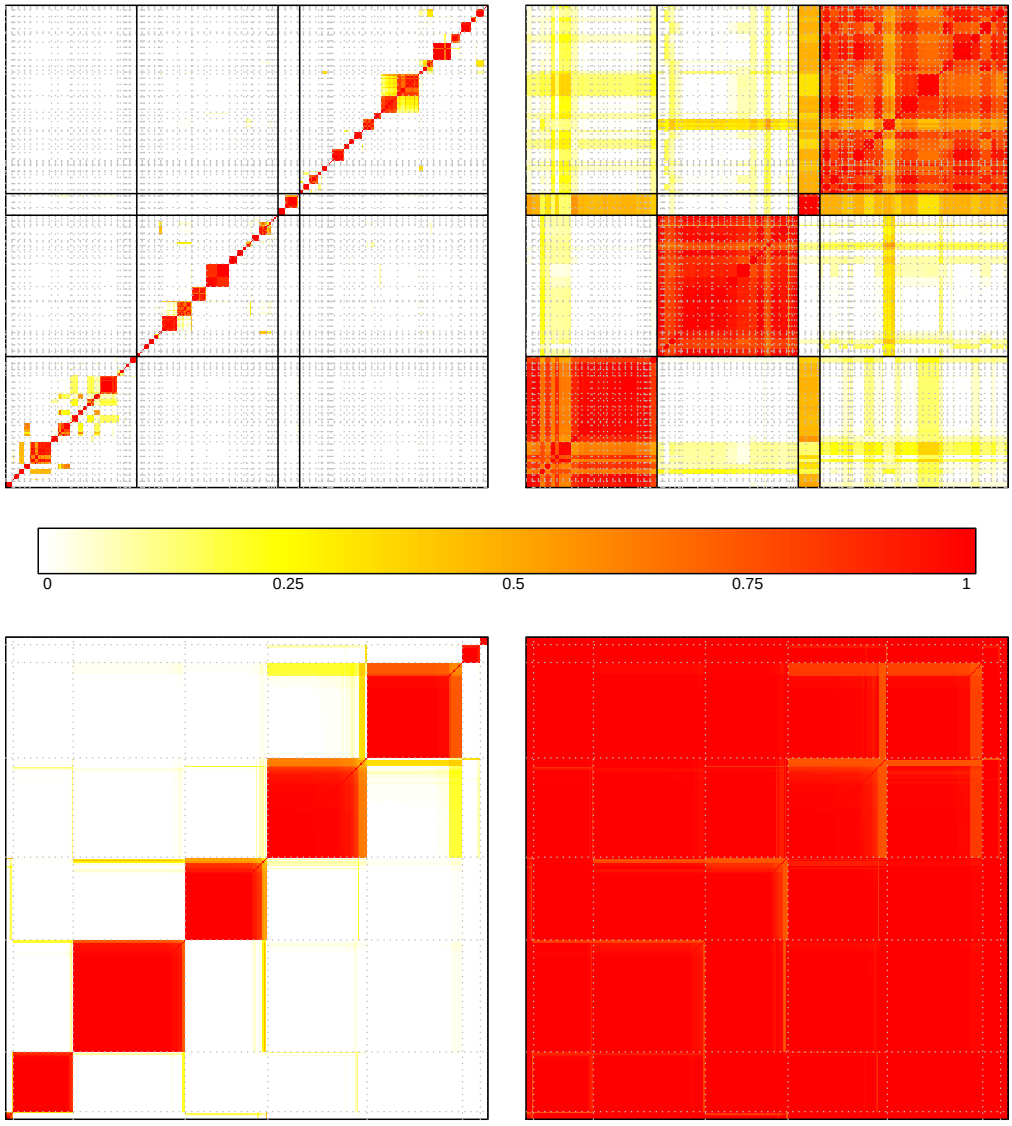

Figure 3.7: Community estimates for collaboration network. **Top:** Monte Carlo estimates of pairwise posterior probabilities of same community, $Pr(\xi_i = \xi_j \mid \mathcal{Y})$ (left), and supercommunity, $Pr(\zeta_{\xi_i} = \zeta_{\xi_j} \mid \mathcal{Y})$ (right). **Bottom** Variational approximations $q(\xi_i = \xi_j)$ (left), and $q(\zeta_{\xi_i} = \zeta_{\xi_j})$ (right).

proportion of the vertices remain together.

In turn, Figure 3.9 shows the the adjacency matrix permuted to show the corresponding community structure under the MCMC (left) and variational Bayes (right). From

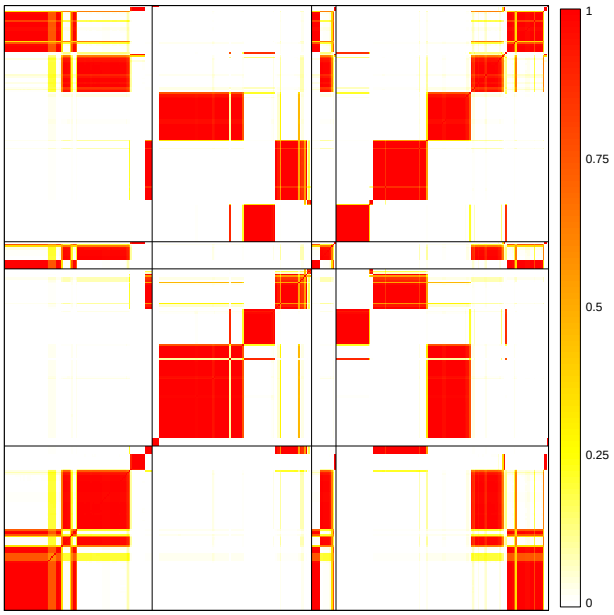

Figure 3.8: Overlap in community structure between obtained under the MCMC and variational algorithms. Colors correspond to the variational probabilities of same community, while the ordering is taken to represent the hierarchical community structure from the MCMC.

this figure it can be seen all four supercommunities recovered by the model trough the MCMC are highly assortative, while the multiple communities found within each supercommunity exhibit a slightly higher propensity of interaction. Instead, the larger communities found by the variational algorithm display a mixture of assortative and disassortative groups in the network

Figure 3.10 shows the evolution of the evolution of the ELBO with respect to execution time, along the mean posterior likelihood of the $100,000$ posterior realizations from the MCMC obtained in approximately $153,000$ seconds (42 hours). From this figure is interesting to observe that the ELBO improves rapidly in the early on, but then takes a large number of steps until convergence resulting in a total execution time of around 22

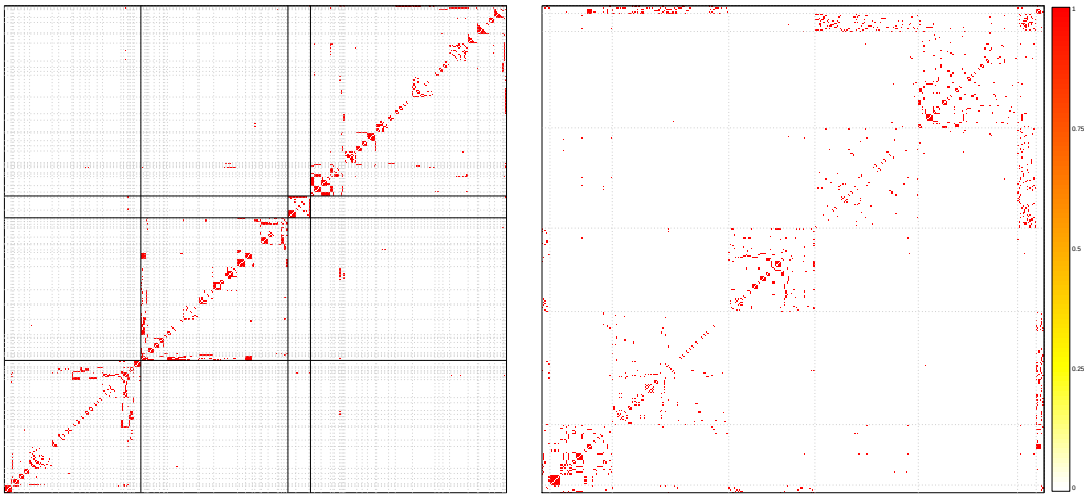

Figure 3.9: Adjacency matrix of the collaboration network ordered with respect to MCMC (left) and variational (right) community structure.

hours, which is not a dramatical improvement over the MCMC.

Figure 3.11 shows the prior and posterior distributions for α, β and σ^2 under different choices of prior for σ^2 and α and β. Here it can be seen that, the data pulls the first level concentration parameter significantly to the right, suggesting, as saw also in Figure 3.7, a much larger number of communities than those implied by the prior. Also, the posterior distribution for σ^2 concentrates around larger values, which is consistent with the communities observed in Figure 3.9 that exhibit either very low or very high proportion of interactions.

Finally, Figure 3.12 shows the result of applying agglomerative clustering based on modularity maximization to this dataset. Here, colors represent the supercommunity membership inferred from the MCMC. Again, the lack of structure in the ordering of the colors in this plot suggests that the agglomerative clustering method leads to solution that

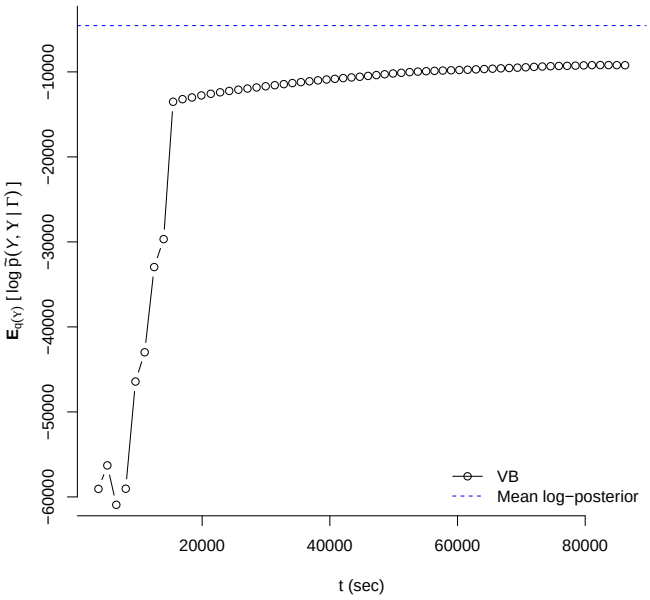

Figure 3.10: Evolution of the lower bound as a function of execution time.

is qualitatively different from that obtained under the stochastic blockmodel.

3.4.3 Food network

As a third illustration we use a network of 86 actors in a biological food web of parasitoids in *Tetramesa*, a genus of grass-infesting chalcid wasps (Clauset et al., 2008; Dawah et al., 1995). In this section we concentrate on the results from the MCMC only as, from the previous illustrations, we expect better behavior than the variational approximation. Figure 3.13 shows the results from fitting the MCMC using $K = 30$, $R = 5$, $\mu_0 = 1$, $\sigma_\mu^2 = 1$, $\alpha_\tau = \alpha_\sigma = 2$, $\beta_\tau = \beta_\sigma = 1$, and $\alpha_\alpha = \beta_\alpha = \alpha_\beta = \beta_\beta = 1$.

As it can be seen in this figure, the model has found no evidence of hierarchical structure in the network. This feature illustrates a significant difference with agglomerative

based methods where, once the dendrogram is cut at the second level, the algorithm will always return a partition. In contrast, the stochastic blockmodel has the possibility of selecting a single supercommunity.

Finally, Figure 3.14 show the corresponding sensitivity analysis under $\mathcal{IG}(2,1)$, $\mathcal{IG}(2,10)$ and $\mathcal{IG}(2,0.1)$ priors for σ^2, as well as a $\mathcal{E}xp(5)$ for α and β. Again, it can be observed that the posterior distributions for α and β are unaffected by the choice of prior for σ^2. In turn, the second choice of prior allows the distribution for the variance parameter to be slightly shifted to larger values.

3.5 Discussion

In this chapter we have introduced a hierarchical extension of the stochastic blockmodel with the goal of being able to identify multilevel community structure in networks. We have also presented a Markov chain Monte Carlo and a variational Bayes algorithm that allow to fit the model and obtain approximate posterior inference. Using simulated and real dataset we have observed that, in fact, the model is capable of identifying communities and supercommunities when these are present in the data. Furthermore, we note that the model is able to return a single supercommunity when there is no evidence of multilevel community structure in the data.

As expected from the case of the single level stochastic blockmodel we observe that the Markov chain Monte Carlo algorithm consistently outperforms its variational Bayes counterpart. For this reason, we recommend the use of MCMC whenever the network size makes it computationally feasible.

Although our results in terms of inference for the community structure have shown to be robust to the choice of prior for the scale parameters in the model, others have argued against the use of the Inverse gamma as a non-informative distribution (Gelman, 2006). For this reason, a possible direction for future research is testing an alternative prior distribution for these parameters, such as the scale Beta2 of Pérez et al. (2017).

Finally, we note that, for simplicity, throughout this chapter we have focused on the case of undirected and binary networks. However generalizations to undirected and/or count networks are straightforward, requiring removing the symmetry constraints in the community parameters and changing the form of the kernel respectively.

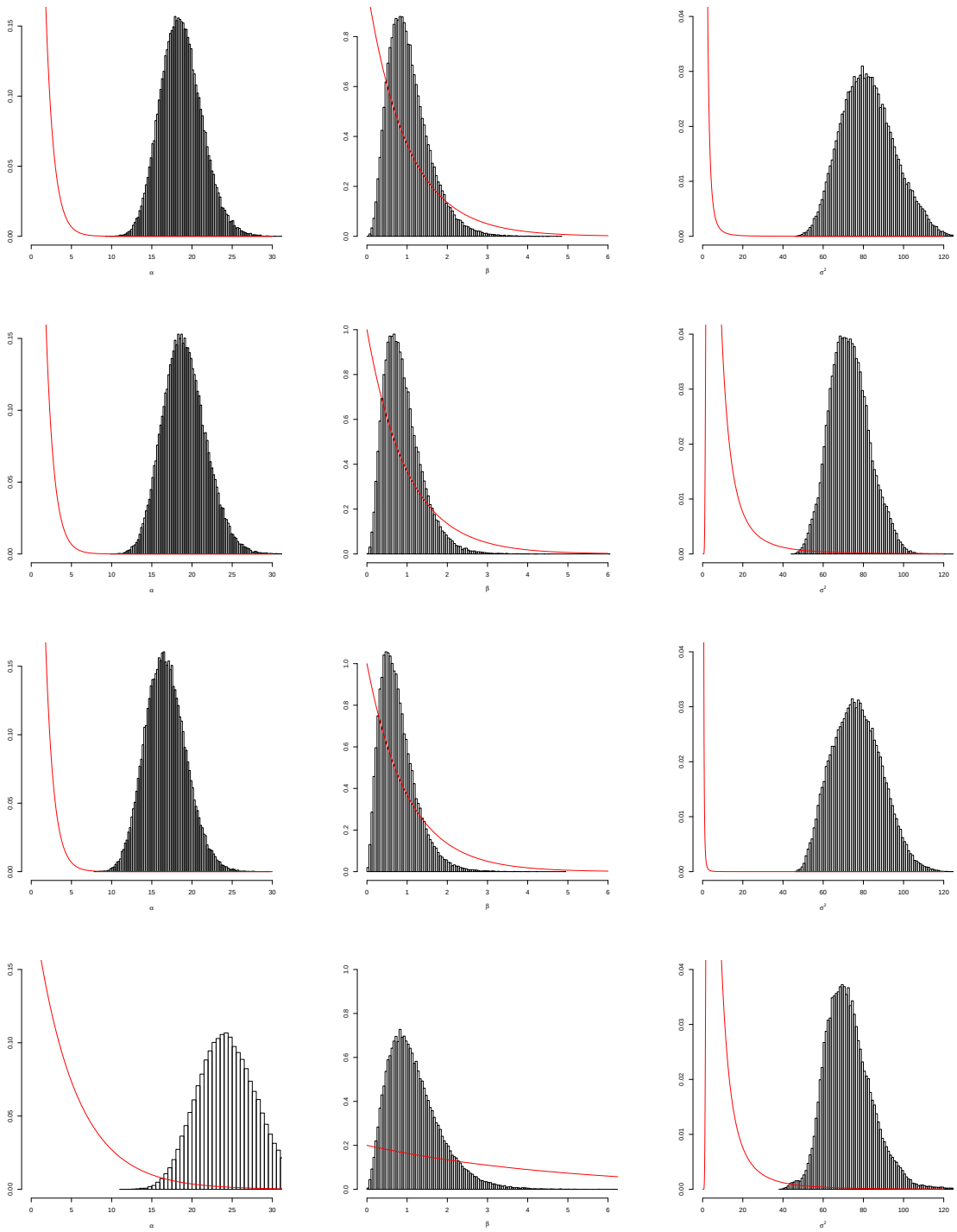

Figure 3.11: prior (line) and posterior (histogram) distributions for the hyperparameters α (left), β (center) and σ^2 (right) for the collaboration network dataset under (a) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,1)$ (b) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,10)$ (c) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,0.1)$ (d) $\alpha, \beta \sim Exp(5)$ and $\sigma^2 \sim \mathcal{IG}(2,10)$.

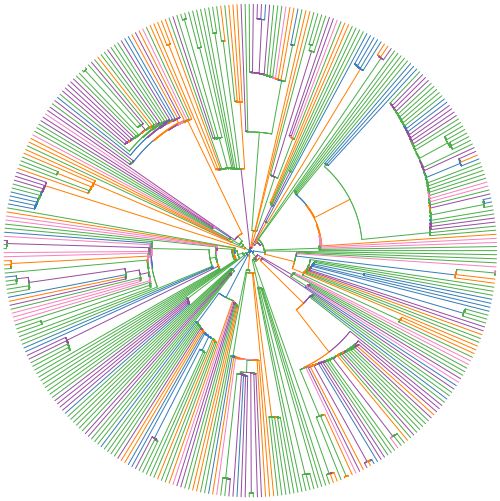

Figure 3.12: Hierarchical structure from agglomerative clustering for the collaboration network.

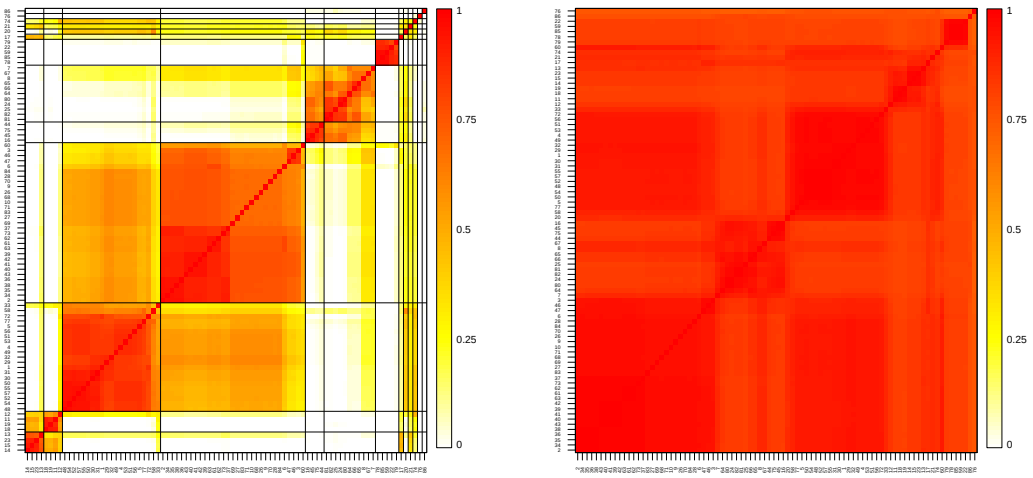

Figure 3.13: Inferred community and supercommunity structure for the food web of grassland species.

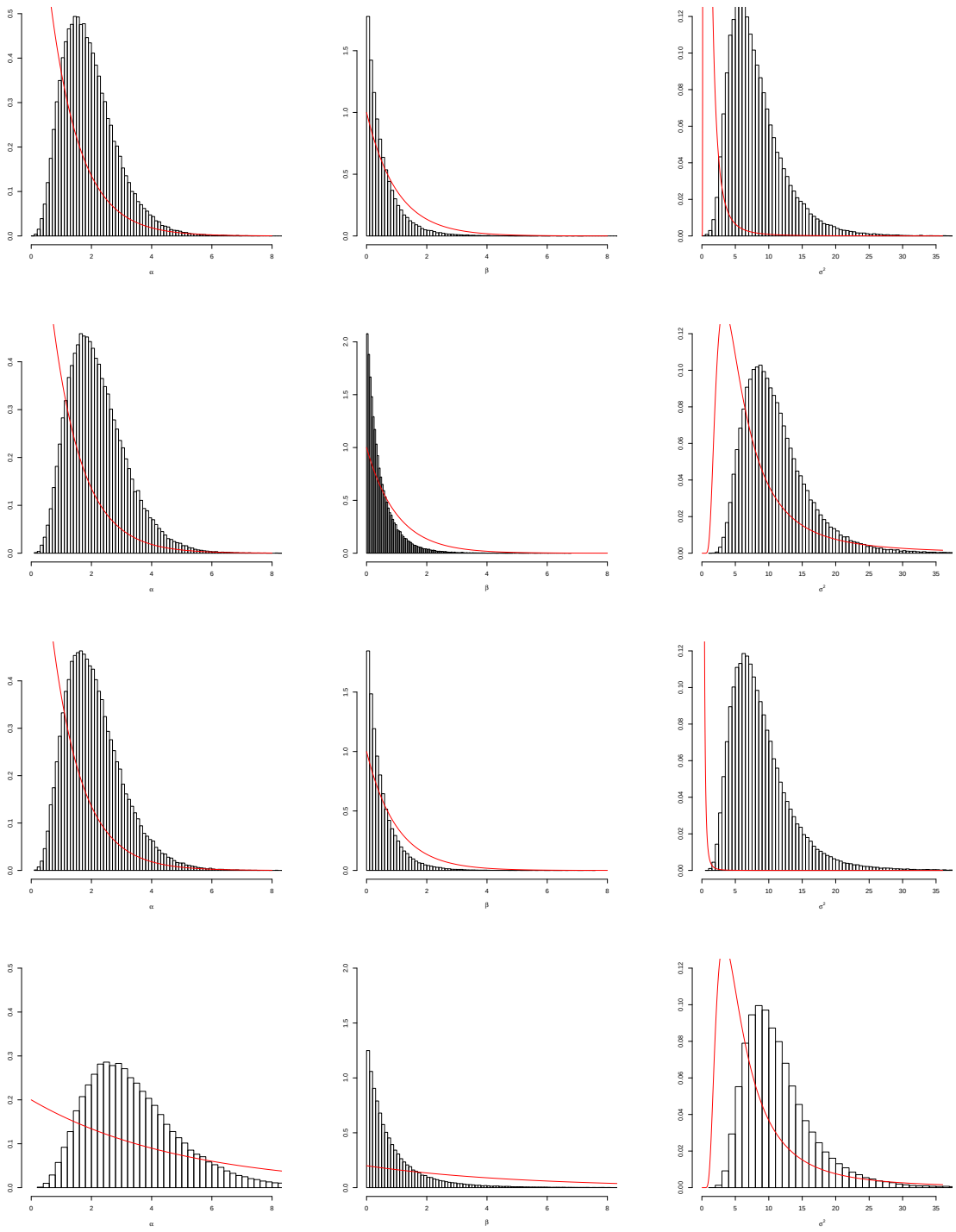

Figure 3.14: prior (line) and posterior (histogram) distributions for the hyperparameters α (left), β (center) and σ^2 (right) for the food web network under (a) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,1)$ (b) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,10)$ (c) $\alpha, \beta \sim Exp(1)$ and $\sigma^2 \sim \mathcal{IG}(2,0.1)$ (d) $\alpha, \beta \sim Exp(5)$ and $\sigma^2 \sim \mathcal{IG}(2,10)$.

Chapter 4

Dynamic evolution of communities in networks

Many real-world applications consist not of a simple network but of a collection a networks measured over time. We are particularly interested in situations in the longitudinal study of the relationship between a fixed set of network nodes. A popular approach to tackle dynamic modeling of networks has been to extend static models by introducing dynamic dependency across some set of parameters in the model. In this way, Sarkar and Moore (2005), Westveld and Hoff (2011), Durante and Dunson (2014) and Sewell and Chen (2015) have presented extensions of latent space models. The main idea behind these models is to express the transformed interaction probabilities as a quadratic combination

$$g(\lambda_{i,j,t}) = \mu(t) + v_i(t)'v_j(t)$$

with v_i and v_j time dependent latent variables. Notice that none of these models is focused on community detection and, therefore, there is no direct community structure obtained

from fitting the model. One possibility to explore dynamic community structures under this type of models would be to apply a clustering algorithm in the latent space for each time step independently.

Alternatively, Rodríguez (2012) and Betancourt et al. (2015) have looked at extending stochastic blockmodels using hidden Markov models. In this setting it is assumed that there are S distinct possible states for the network corresponding to a set of community parameters Θ_s and a vector of community parameters $\boldsymbol{\xi}_s$. Then, an interaction between vertex i and vertex j at time t occurs with probability $\theta_{\xi_{\zeta_{ti}},\xi_{\zeta_{tj}},\zeta_t}$ where $\zeta_t = s$ indicates that the network is at state s at time t and a dynamic can be introduced through $\boldsymbol{\zeta}$. Specifically, a first order Markov process is taken as

$$Pr(\zeta_t = s \mid \zeta_{t-1} = r, \boldsymbol{\pi}_r) = \pi_{r,s}$$

with symmetric Dirichlet priors for the transition probabilities. This models are particularly useful for change point identification in the network structure; however they can lead abrupt changes in the community structures.

Here we focus on community evolution. Specifically, there are two ways in which communities might evolve as the network changes across time: *migration* of individuals between communities, and changes in the propensity of interaction between individuals within or across communities. In this chapter we concentrate on modeling dynamic network data in a way that is capable of capturing the community structure in the network at each period and the evolution of these communities by "borrowing" information across time.

A first possibility to study network dynamics is to track specific features or summaries of the network, such as transitivity or degree, over time; this, however, implies an

intrinsic loss of information when modeling only these summaries as opposed to the whole data. A second option would be to try to put the data into the framework of the model introduced in Chapter 1 by projecting these networks into a single *consensus* network that can then be analyzed using the static stochastic blockmodel. However, this procedure will typically be inappropriate and lead to an artificial reduction of the uncertainty in the model. Furthermore, this approach does not allow to identify features that are common across time or possible structural changes in the network; both interesting questions from an application perspective. On the opposite end of the spectrum, a stochastic blockmodel could be fitted separately to each network, but this approach has the obvious drawback of ignoring any dependence structure in the data. Instead, in this chapter we propose the use of a dynamic extension of the stochastic blockmodel.

A major challenge in dynamically extending the stochastic blockmodel is defining an appropriate transition probability for the community indicators $p(\boldsymbol{\xi}_t \mid \boldsymbol{\xi}_{t-1})$. To this end, Yang et al. (2011) assume a Hidden Markov model for each sequence $\{\xi_{t,i}\}_{t=1}^{T}$ with a constant transition probability matrix, while Xing et al. (2010) extend the mixed membership stochastic blockmodel using an autoregressive structure on the hyperparameters. Here we propose the use of fragmentation-coagulation processes (Bertoin, 2006) described in the following section.

4.1 Random partitions

A generalization of the Chinese restaurant process prior used in the previous chapters is given as follows:

We say that $\boldsymbol{\xi}$ follows a *generalized Chinese restaurant process* (GCRP) with parameters α and β if its distribution corresponds to the EPPF of a two-parameter Poisson-Dirichlet (Pitman-Yor) process, *i.e.,*

$$p(\xi_1, \xi_2, \ldots, \xi_I) = \frac{\Gamma(\alpha+1)}{(\alpha+\beta N^I)\Gamma(\alpha+I)} \prod_{k=1}^{N^I} (\alpha+\beta k) \frac{\Gamma(n_k^I - \beta)}{\Gamma(1-\beta)}$$

for $0 \leq \beta < 1$ and $\alpha > -\beta$, and where $N^I = \max\{\boldsymbol{\xi}\}$. Notice that the GCRP reduces to the standard Chinese restaurant process when $\beta = 0$.

Now consider two partitions of $\{1, 2, ..., I\}$, $\boldsymbol{\zeta}$ and $\boldsymbol{\gamma}$. It is said that $\boldsymbol{\gamma}$ is a fragmentation of $\boldsymbol{\zeta}$ if and only if $\gamma_i = \gamma_j$ implies that $\zeta_i = \zeta_j$; alternatively, it can be said that $\boldsymbol{\zeta}$ is a coagulation of $\boldsymbol{\gamma}$. In particular, following Bertoin (2006), we say that a partition $\boldsymbol{\xi}$ is a *random GCRP coagulation* of the partition $\boldsymbol{\gamma}$ if $\xi_i = \varsigma_{\gamma_i}$ where $\boldsymbol{\varsigma}$ follows a GCRP prior. Similarly, we say that a partition $\boldsymbol{\gamma}$ is a *random GCRP fragmentation* of $\boldsymbol{\zeta}$ if it is generated by sequentially splitting each cluster in $\boldsymbol{\xi}$ according to a GCRP.

Figure 4.1 presents a graphical example of a fragmentation-coagulation process in which, in a fist stage, each of the first two clusters fragments in two, while the third remains unaffected. In a second stage, clusters one, four and five combine into a single group and, similarly, clusters two and three coagulate.

An appealing feature of random fragmentation-coagulation processes is given by the fact that, if $\boldsymbol{\zeta} \mid \alpha \sim \text{CRP}(\alpha)$, and $\boldsymbol{\xi}$ is generated by letting $\boldsymbol{\gamma} \mid \boldsymbol{\zeta}, \beta \sim \text{GCRPFrag}(\boldsymbol{\zeta}, 0, \beta)$ and $\boldsymbol{\xi} \mid \boldsymbol{\gamma}, \alpha, \beta \sim \text{GCRPCoag}(\boldsymbol{\gamma}, \alpha/\beta, 0)$, then the marginal distribution of $\boldsymbol{\xi}$ is also a $\text{CRP}(\alpha)$. Thus, fragmentation-coagulation processes provide a natural way to define transition kernels that lead to a stationary process on the community indicators for a dynamic extension of stochastic blockmodels. Having a stationary process on these parameters is

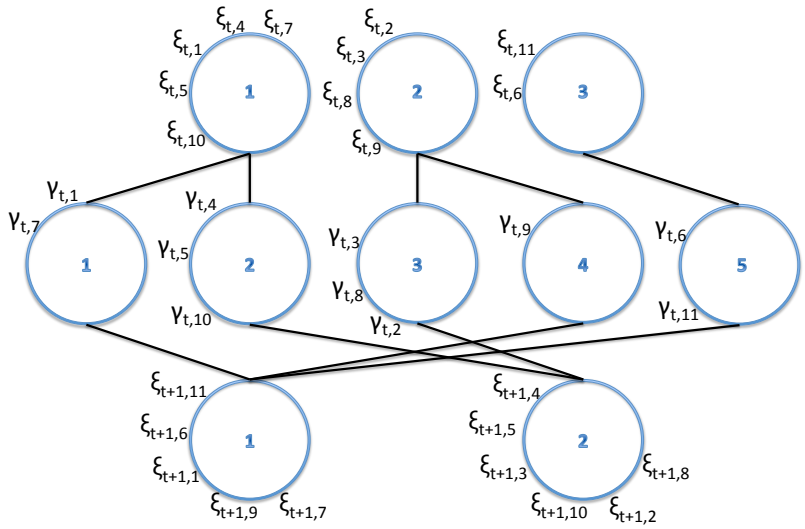

Figure 4.1: Graphical example of fragmentation-coagultion processes.

consistent with the fact that, a priori, we have no information that allows us to distinguish the partition across time. Furthermore, it also leads to a prior structure that is computationally convenient as, conditionally, at any point in time, the prior for the community indicators reduces to that of the static model.

4.2 A model for dynamic networks

In this section we propose a dynamic extension of the stochastic blockmodel. Turning attention to directed networks, for $i, j = 1, 2, \ldots, I$ and $i \neq j$ interactions are assumed conditionally independent with

$$y_{i,j,t} \mid \Theta_t, \boldsymbol{\xi}_t \sim \mathcal{B}er\left(\lambda_{i,j,t}\right), \quad \text{for } t = 1, 2, \ldots, T, \text{ and } i \neq j. \tag{4.1}$$

where

$$\lambda_{i,j,t} = \text{expit}(\theta_{\xi_{ti},\xi_{tj},t}) \equiv \frac{\exp\{\theta_{\xi_{ti},\xi_{tj},t}\}}{1+\exp\{\theta_{\xi_{ti},\xi_{tj},t}\}}.$$

Using the result from Section 4.1, a joint prior for $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T\}$ can be constructed by placing a CRP(α) on $\boldsymbol{\xi}_1$, and sequentially assuming a common fragmentation-coagulation process $\boldsymbol{\gamma_t} \mid \boldsymbol{\xi}_t, \beta \sim \text{GCRPFrag}(\boldsymbol{\xi}_t, 0, \beta)$ and $\boldsymbol{\xi}_{t+1} \mid \boldsymbol{\gamma}_t, \alpha, \beta \sim \text{GCRPCoag}(\boldsymbol{\gamma}_t, \alpha/\beta, 0)$. This leads to a strongly stationary process in which, for any t, we have that $\boldsymbol{\xi}_t \sim CRP(\alpha)$, which means that marginally the model retains the properties of the static stochastic blockmodel. In this case the parameter $\beta \in (0,1)$ controls the dependency in the community indicators across time. As $\beta \to 0$ the partition becomes more stable leading to $\boldsymbol{\xi}_t = \boldsymbol{\gamma}_t = \boldsymbol{\xi}_{t+1}$ for all t. In turn, $\beta \to 1$ yields a total fragmentation in which $\boldsymbol{\gamma}_t$ is comprised of I singleton clusters; therefore implying independence between $\boldsymbol{\xi}_t$ and $\boldsymbol{\xi}_{t+1}$. As before, the parameter α controls the expected number of communities a priori, as well as the prior co-clustering probabilities

$$Pr(\xi_{ti} = \xi_{tj} \mid \alpha) = \frac{1}{1+\alpha} \quad \text{for any } t = 1, 2, \ldots, T.$$

The hyperparameters α and β are assigned a $\mathcal{G}(a_\alpha, b_\alpha)$ and $\mathcal{B}eta(a_\beta, b_\beta)$ prior respectively, and learned from the data.

A second difficulty in extending the stochastic blockmodel is the specification of a joint prior on the community parameters $\{\Theta_1, \Theta_2, \ldots, \Theta_T\}$. Assuming time independence in the interaction probabilities may lead to simple but unrealistic models, as we expect certain features of the communities to remain fairly stable. To the best of the authors' knowledge, dynamical modeling of the community parameters remains an open research question. Here, we propose the use of autoregressive priors that account for the structure of the communities.

This kind of prior allows for smooth changes in the interaction probabilities. Intuitively, our choice of mean function can be thought as letting, the community parameters at time t be (in mean) an average of the values of the community parameters at $t-1$ weighted by number of current possible interactions. Formally, for any $t \geq 2$ the elements of Θ_t are assumed to be conditionally independent from

$$\theta_{k,l,t} \mid \boldsymbol{\xi}_t, \Theta_{t-1}, \boldsymbol{\xi}_{t-1} \sim \mathcal{N}\left(\mathbf{m}_{k,l}(\boldsymbol{\xi}_t, \Theta_{t-1}, \boldsymbol{\xi}_{t-1}), \mathbf{v}_{k,l}(\boldsymbol{\xi}_t)\right) \tag{4.2}$$

with the mean function given by

$$\mathbf{m}_{k,l}(\boldsymbol{\xi}_t, \Theta_{t-1}, \boldsymbol{\xi}_{t-1}) = \begin{cases} \frac{1}{n_{k,l,t}} \sum_{\mathcal{S}_{k,l,t}} \theta_{\xi_{t_i}, \xi_{t_j}, t-1} \text{ if } n_{k,l,t} > 0 \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{4.3}$$

and variance function

$$\mathbf{v}_{k,l}(\boldsymbol{\xi}_t) = \begin{cases} \frac{\sigma^2}{n_{k,l,t}} \text{ if } n_{k,l,t} > 0 \\ \sigma^2 \quad \text{otherwise} \end{cases} \tag{4.4}$$

where $n_{k,l,t} = \sum_{\mathcal{S}_{k,l,t}} 1$ and the sum is taken over $\mathcal{S}_{k,l,t} = \left\{(i,j) : i \neq j, \xi_{t_i} = k, \xi_{t_j} = l\right\}$. Interestingly, notice that for a model with constant partition in the indicators, this prior reduces to independent random walk processes. For Θ_1 we maintain the prior structure as

$$\theta_{k,l,1} \mid \boldsymbol{\xi}_1, \sigma^2 \sim \mathcal{N}\left(0, \sigma^2\right) \tag{4.5}$$

for $k, l = 1, 2, \ldots, K_1$, with $K_t = \max\{\boldsymbol{\xi}_t\}$. The hyperparameter σ^2 can then be fixed, or an extra layer can be added to de model by assigning conditionally conjugate prior

$$\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma).$$

4.3 Posterior sampling

As before, we introduce Polya-Gamma auxiliary variables (Polson et al., 2013) in order to have a closed form full conditional distribution for the community parameters. In this case each term in the likelihood can be re-expressed as

$$p(Y_t \mid \Theta_t, \boldsymbol{\xi}_t) \propto \exp\left\{\sum_{k=1}^{K_t}\sum_{l=1}^{K_t}\left(s_{k,l,t} - \frac{n_{k,l,t}}{2}\right)\theta_{k,l,t}\right\}\prod_{k=1}^{K_t}\prod_{l=1}^{K_t}\mathbb{E}\left[\exp\left\{-\frac{\theta_{k,l,t}^2}{2}\omega_{k,l,t}\right\}\right] \quad (4.6)$$

where $\omega_{k,l,t} \sim \mathcal{PG}(n_{k,l,t}, 0)$ is a Polya-Gamma random variable and $s_{k,l,t} = \sum_{\mathcal{S}_{k,l,t}} y_{i,j,t}$.

Denoting $\Upsilon = \left\{\{\Theta_t\}_{t=1}^T, \{\boldsymbol{\xi}_t\}_{t=1}^T, \{\boldsymbol{\gamma}_t\}_{t=1}^{T-1}, \{\Omega_t\}_{t=1}^T, \alpha, \beta\right\}$ the posterior can then be written as

$$p\left(\Upsilon \mid \{Y_t\}_{t=1}^T\right) \propto \prod_{t=1}^T p(Y_t \mid \Theta_t, \boldsymbol{\xi}_t, \Omega_t) p(\Theta_1 \mid \boldsymbol{\xi}_1) \prod_{t=2}^T p(\Theta_t \mid \boldsymbol{\xi}_t, \Theta_{t-1}, \boldsymbol{\xi}_{t-1}) \prod_{t=1}^T p(\Omega_t)$$
$$p(\boldsymbol{\xi}_1) \prod_{t=2}^T p(\boldsymbol{\xi}_t \mid \boldsymbol{\gamma}_{t-1}, \alpha, \beta) p(\boldsymbol{\gamma}_{t-1} \mid \boldsymbol{\xi}_{t-1}, \beta) p(\alpha, \beta) \quad (4.7)$$

from where a Markov Chain Monte Carlo algorithm can be envisioned. Specifically, we implement a collapsed algorithm similar to algorithm 4 in Neal (2000) for nonconjugate Dirichlet process mixture models. This algorithm can be broken down into three main steps. In the fist stage, the indicators and interaction probabilities are sampled jointly. Next, the algorithm resamples the interaction probabilities given the indicators to improve mixing, as is customary for collapsed samplers. Lastly we sample the hyperparameters controlling the fragmentation-coagulation process α and β.

For the first step, following Rodríguez (2014), the elements of the parameters associated with the community indicators, $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_{T-1}\}$, are sampled jointly in pairs $(\boldsymbol{\gamma}_{t-1}, \boldsymbol{\xi}_t)$ and $(\boldsymbol{\xi}_t, \boldsymbol{\gamma}_{t+1})$, in order to improve the mixing of the algorithm. Specifically, for $t = 2, \ldots, T-1$ the full conditional probabilities for the update of $(\gamma_{t-1_i}, \xi_{t_i})$ are

obtained from

$$p(\gamma_{t-1_i} = l, \xi_{t_i} = k, \boldsymbol{\theta}_t^\star \mid \Upsilon_{-(\gamma_{t-1_i}, \xi_{t_i})}, \{Y_t\}) \propto p(Y_t \mid \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,k)}, \Omega_t) p(\Theta_t, \boldsymbol{\theta}_t^\star \mid \boldsymbol{\xi}_t^{(i,k)}, \Theta_{t-1}, \boldsymbol{\xi}_{t-1})$$
$$p(\Theta_{t+1} \mid \boldsymbol{\xi}_{t+1}, \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,k)}) p(\boldsymbol{\gamma}_t \mid \boldsymbol{\xi}_t^{(i,k)}, \beta) p(\boldsymbol{\xi}_t^{(i,k)} \mid \boldsymbol{\gamma}_{t-1}^{(i,l)}, \alpha, \beta) p(\boldsymbol{\gamma}_{t-1}^{(i,l)} \mid \boldsymbol{\xi}_{t-1}, \beta) \tag{4.8}$$

where $\boldsymbol{\xi}_t^{(i,k)} = (\xi_{t1}, \ldots, \xi_{ti-1}, k, \xi_{ti+1} \ldots, \xi_{tI})$, $\boldsymbol{\gamma}_t^{(i,l)} = (\gamma_{t1}, \ldots, \gamma_{ti-1}, l, \gamma_{ti+1} \ldots, \gamma_{tI})$ and $\boldsymbol{\theta}_t^\star$ is the vector of interaction probabilities for the case in which a new community is opened. Denote $\boldsymbol{\xi}_{t-(i)} = (\xi_{t1}, \ldots, \xi_{ti-1}, \xi_{ti+1} \ldots, \xi_{tI})$, $\boldsymbol{\gamma}_{t-(i)} = (\gamma_{t1}, \ldots, \gamma_{ti-1}, \gamma_{ti+1} \ldots, \gamma_{tI})$ and K_t^{-i}, R_{t-1}^{-i} the number of clusters associated with $\boldsymbol{\xi}_{t-(i)}$ and $\boldsymbol{\gamma}_{t-(i)}$ respectively; notice that, in principle, there are $(R_{t-1}^{-i} + 1) \times (K_t^{-i} + 1)$ distinct possible ways of allocating $(\gamma_{t-1_i}, \xi_{t_i})$ corresponding to either assigning actor i to an existing cluster or opening a new one at both levels. However, many of these combinations have zero posterior probability as they lead to partitions that do not correspond to a fragmentation-coagulation process. From a computational perspective, this observation is crucial, as it leads to an efficient implementation of the algorithm, in which only combinations that are consistent with a coagulation-fragmentation process are explored, effectively reducing the amount of computation required.

Implied from the fragmentation-coagulation process, the priors satisfy

$$p(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{\gamma}_t, \alpha, \beta) = \frac{\Gamma(\alpha/\beta)}{\Gamma(R_t + \alpha/\beta)} \left(\frac{\alpha}{\beta}\right)^{K_{t+1}} \prod_{k=1}^{K_{t+1}} \Gamma(g_{t+1,k}) \tag{4.9}$$

where R_t the number of clusters induced by $\boldsymbol{\gamma}_t$ and $g_{t+1,k}$ the number of clusters from $\boldsymbol{\gamma}_t$ that have been grouped together to form the k-th cluster of $\boldsymbol{\xi}_{t+1}$, and

$$p(\boldsymbol{\gamma}_t \mid \boldsymbol{\xi}_t, \beta) = \prod_{k=1}^{K_t} \frac{\beta^{Q_{t_k}-1} \Gamma(Q_{t_k})}{\Gamma(n_{t_k})} \prod_{s=1}^{Q_{t_k}} \frac{\Gamma(q_{t_{k,s}} - \beta)}{\Gamma(1-\beta)} \tag{4.10}$$

with n_{t_k} the size of the k-th community associated with $\boldsymbol{\xi}_t$, Q_{t_k} is the number of subclusters into which the k-th cluster of $\boldsymbol{\xi}_t$ is fragmented, and $q_{t_{k,s}}$ is the size of the s-th subcluster in $\boldsymbol{\gamma_t}$ associated with the k-th cluster in $\boldsymbol{\xi}_t$.

For the case in which a new community is opened, sampling from (4.8) is handled in two steps; first the new interaction probabilities are obtained from its full conditional

$$p(\boldsymbol{\theta}_t^\star \mid \gamma_{t-1_i} = l, \xi_{t_i} = K_t^{-1} + 1, \Upsilon_{-(\gamma_{t-1_i}, \xi_{t_i})}, \{Y_t\}) \propto p(Y_t \mid \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,K_t^{-1}+1)}, \Omega_t)$$
$$p(\Theta_{t+1} \mid \boldsymbol{\xi}_{t+1}, \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,K_t^{-1}+1)}) p(\boldsymbol{\theta}_t^\star \mid \boldsymbol{\xi}_t^{(i,K_t^{-1}+1)}, \Theta_{t-1}, \boldsymbol{\xi}_{t-1}) \quad (4.11)$$

and then, the community indicators are then sampled with probabilities proportional to

$$p(\gamma_{t-1_i} = l, \xi_{t_i} = k \mid \boldsymbol{\theta}_t^\star, \Upsilon_{-(\gamma_{t-1_i}, \xi_{t_i})}, \{Y_t\}) \propto p(Y_t \mid \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,k)}, \Omega_t) p(\Theta_t, \boldsymbol{\theta}_t^\star \mid \boldsymbol{\xi}_t^{(i,k)}, \Theta_{t-1}, \boldsymbol{\xi}_{t-1})$$
$$p(\Theta_{t+1} \mid \boldsymbol{\xi}_{t+1}, \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,k)}) p(\boldsymbol{\gamma}_t \mid \boldsymbol{\xi}_t^{(i,k)}, \beta) p(\boldsymbol{\xi}_t^{(i,k)} \mid \boldsymbol{\gamma}_{t-1}^{(i,l)}, \alpha, \beta) p(\boldsymbol{\gamma}_{t-1}^{(i,l)} \mid \boldsymbol{\xi}_{t-1}, \beta). \quad (4.12)$$

The case for $t = T$ is slightly different as there is no contribution from the term of future interaction probabilities, in this case the full conditional distribution is given by

$$p(\gamma_{T-1_i} = l, \xi_{T_i} = k, \boldsymbol{\theta}_T^\star \mid \Upsilon_{-(\gamma_{T-1_i}, \xi_{T_i})}, \{Y_t\}) \propto p(Y_T \mid \Theta_T, \boldsymbol{\theta}_T^\star, \boldsymbol{\xi}_T^{(i,k)}, \Omega_t)$$
$$p(\Theta_T, \boldsymbol{\theta}_T^\star \mid \boldsymbol{\xi}_T^{(i,k)}, \Theta_{T-1}, \boldsymbol{\xi}_{T-1}) p(\boldsymbol{\gamma}_T \mid \boldsymbol{\xi}_T^{(i,k)}, \beta) p(\boldsymbol{\xi}_T^{(i,k)} \mid \boldsymbol{\gamma}_{T-1}^{(i,l)}, \alpha, \beta) p(\boldsymbol{\gamma}_{T-1}^{(i,l)} \mid \boldsymbol{\xi}_{T-1}, \beta) \quad (4.13)$$

An analogous approach is used to jointly sample $(\xi_{t_i}, \gamma_{t_i})$. The full conditional distribution for $t = 2, \ldots, T-1$, is

$$p(\xi_{t_i} = k, \gamma_{t_i} = l, \boldsymbol{\theta}_t^\star \mid \Upsilon_{-(\xi_{t_i}, \gamma_{t_i})}, \{Y_t\}) \propto p(Y_t \mid \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,k)}, \Omega_t) p(\Theta_t, \boldsymbol{\theta}_t^\star \mid \boldsymbol{\xi}_t^{(i,k)}, \Theta_{t-1}, \boldsymbol{\xi}_{t-1})$$
$$p(\Theta_{t+1} \mid \boldsymbol{\xi}_{t+1}, \Theta_t, \boldsymbol{\theta}_t^\star, \boldsymbol{\xi}_t^{(i,k)}) p(\boldsymbol{\xi}_{t+1} \mid \boldsymbol{\gamma}_t^{(i,l)}, \alpha, \beta) p(\boldsymbol{\gamma}_t^{(i,l)} \mid \boldsymbol{\xi}_t^{(i,k)}, \beta) p(\boldsymbol{\xi}_t^{(i,k)} \mid \boldsymbol{\gamma}_{t-1}, \alpha, \beta) \quad (4.14)$$

and, in particular, when $t = 1$

$$p(\xi_{1_i} = k, \gamma_{1_i} = l, \boldsymbol{\theta}_1^\star \mid \Upsilon_{-(\xi_{t_1}, \gamma_{1_i})}, \{Y_t\}) \propto p(Y_1 \mid \Theta_1, \boldsymbol{\theta}_1^\star, \boldsymbol{\xi}_1^{(i,k)}, \Omega_1) p(\Theta_1, \boldsymbol{\theta}_1^\star \mid \boldsymbol{\xi}_1^{(i,k)})$$
$$p(\Theta_2 \mid \boldsymbol{\xi}_2, \Theta_1, \boldsymbol{\theta}_1^\star, \boldsymbol{\xi}_1^{(i,k)}) p(\boldsymbol{\xi}_2 \mid \boldsymbol{\gamma}_1^{(i,l)}, \alpha, \beta) p(\boldsymbol{\gamma}_1^{(i,l)} \mid \boldsymbol{\xi}_1^{(i,k)}, \beta) p(\boldsymbol{\xi}_1^{(i,k)} \mid \alpha) \quad (4.15)$$

with

$$p(\boldsymbol{\xi}_1 \mid \alpha) = \frac{\Gamma(\alpha)\alpha^{K_t}}{\Gamma(I + \alpha)} \prod_{k=1}^{K_t} \Gamma(n_{1_k}). \quad (4.16)$$

To implement the second second step of our algorithm, we exploit the fact that, by introducing the auxiliary variable $\{\Omega_t\}$, the community parameters are conditionally conjugate and can be sample directly from the corresponding full conditional

$$p\left(\Theta_{k,l,t} \mid \Upsilon_{-(\Theta_{k,l,t})}, \{Y_t\}_{t=1}^T\right) \propto \prod_{\mathcal{S}_{k,l,t}} p(Y_{i,j,t} \mid \Theta_{k,l,t}, \boldsymbol{\xi}_t, \Omega_t) p(\Theta_{t,k,l,} \mid \boldsymbol{\xi}_t, \Theta_{t-1}, \boldsymbol{\xi}_{t-1})$$
$$p(\Theta_{t+1} \mid \boldsymbol{\xi}_{t+1}, \Theta_t, \boldsymbol{\xi}_t) \quad (4.17)$$

which is identified as a Gaussian distribution with mean and variance that can be found exploiting the linearity of $\mathbf{m}_{k,l}(\boldsymbol{\xi}_t, \Theta_{t-1}, \boldsymbol{\xi}_{t-1})$.

Now, as in the case of the static stochastic blockmodel, the full conditional for the auxiliary variables remains in the Polya-Gamma family, which can be sample following the approach described in Polson et al. (2013).

Finally, the hyperparameters governing the fragmentation coagulation process (α, β) are jointly sampled from their full conditional distribution using a random-walk Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) with α and β proposed from a bivariate Gaussian distribution in the log and logit scale respectively, and where the parameters in the covariance matrix are tuned to achieve an approximate acceptance rate of 40%.

4.4 Evaluation

4.4.1 Simulated data

In this section we illustrate the performance of the dynamic extension of the stochastic blockmodel in terms of the quality of the inferred community structure in a case in which we know the true structure of the partition. To this end we make use a dataset simulated from the model itself with $T = 10$ observations over time of a network consisting of $I = 30$ individuals. In particular, the true mean for the marginal distribution of the interaction probabilities at $t = 1$ is taken to be -2 which leads to an overall sparse network, and a relatively large variance, $\sigma^2 = 4$, such that initial blocks are clearly distinguishable. In the community indicators side, the true parameters governing the fragmentation-coagulation process are set to $\alpha = 1$ and $\beta = 0.15$, implying a number of expected clusters in the order of four, and a gradual community migration of vertices. The left column of Figure 4.2 shows the adjacency matrices matrices at three different point in time; note that the vertices have been ordered in a slightly different way in each row in order to simplify the representation of the true community structure. From these figures it can be observed that the model is flexible enough to allow for significant variation both in the number of communities in the network, as well as propensity of interaction between these communities.

The center column of Figure 4.2 show the posterior mean for the interaction probabilities at $t = 1$, $t = 3$ and $t = 8$ obtained from $20,000$ MCMC samples obtained after a burn-in period of $20,000$. In turn, the right column of this figure shows the corresponding posterior co-clustrering probabilities $p(\xi_{t_i} = \xi_{t_j} \mid \{Y_t\})$ for every pair of vertices at the same three points in time. From these plots we note that the model reasonably recovers the

underlying community structure in the first two cases; in contrast, for $t = 8$ we see that the small number of interactions in the dataset make it hard for the model to distinguish between the fist two communities.

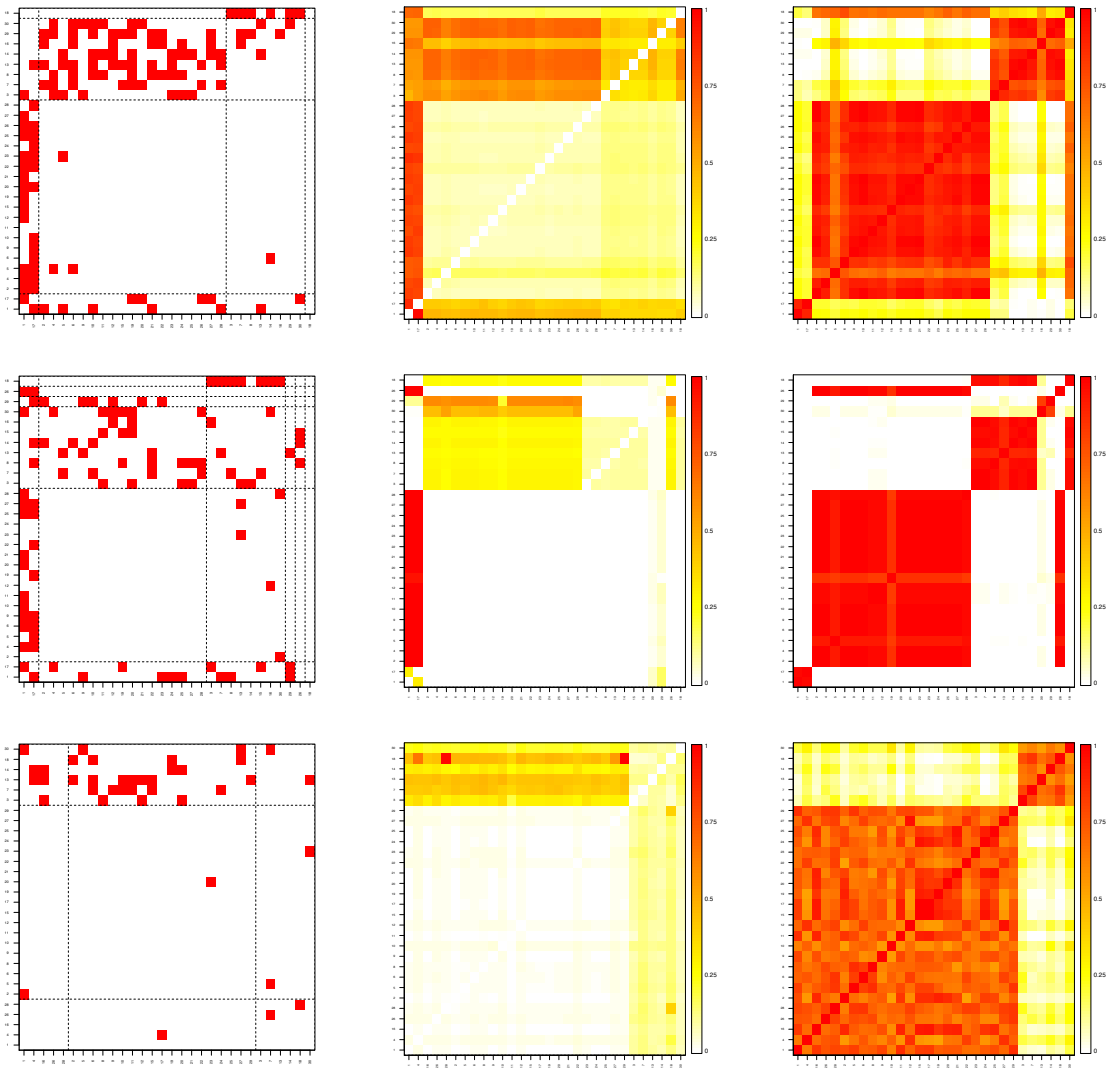

Figure 4.2: Adjacency matrix (left), matrix of posterior mean interaction probabilities $\hat{\lambda}_{i,j}$ (center) and posterior co-clustering probabilities (right) for three snapshots of the simulated dataset at $t = 1$ (top), $t = 3$ (middle), $t = 8$ (bottom).

4.4.2 Financial trading network

For a second illustration we make use of a dataset consisting of financial trading networks from the natural gas futures market in the New York Mercantile Exchange (NYMEX). This data, which was first analyzed by Betancourt et al. (2015), consists of $T = 201$ weeks of trading between January 2005 and December 2008. In this network the $I = 71$ vertices represent traders, and the presence of an edge ($y_{t,i,j} = 1$) signifies that trader i sold to trader j at least once during week t. From the characteristics of the market, we expect to observe traders with similar transaction patterns and, thus, groups of structurally equivalent vertices, generating community structure. Also from background knowledge of the dataset, we expect to observe a change in the structure of the network after September 2006, when electronic trading was introduced into the market.

The algorithm described in Section 4.3 was used to obtain $20,000$ samples after a burn-in period of equal length. Figure 4.3 shows the resulting community structure for four of the 201 time points, using different ordering of the vertices. From this Figure it is clear that the model does recover community structure from the data and that this structure, in fact, evolves with time.

Next, in order to investigate the stability of the community structure, for each pair of the resulting optimal partitions we compute the adjusted Rand index (Hubert and Arabie, 1985). This measure of similarity between two partitions is a chance-corrected proportion of pairwise agreements and, thus, higher values represent higher levels of agreement with an upper limit value of one, representing perfect agreement. The resulting matrix of ARI values is shown in Figure 4.4 where it can be seen that there is a clear disruption in the

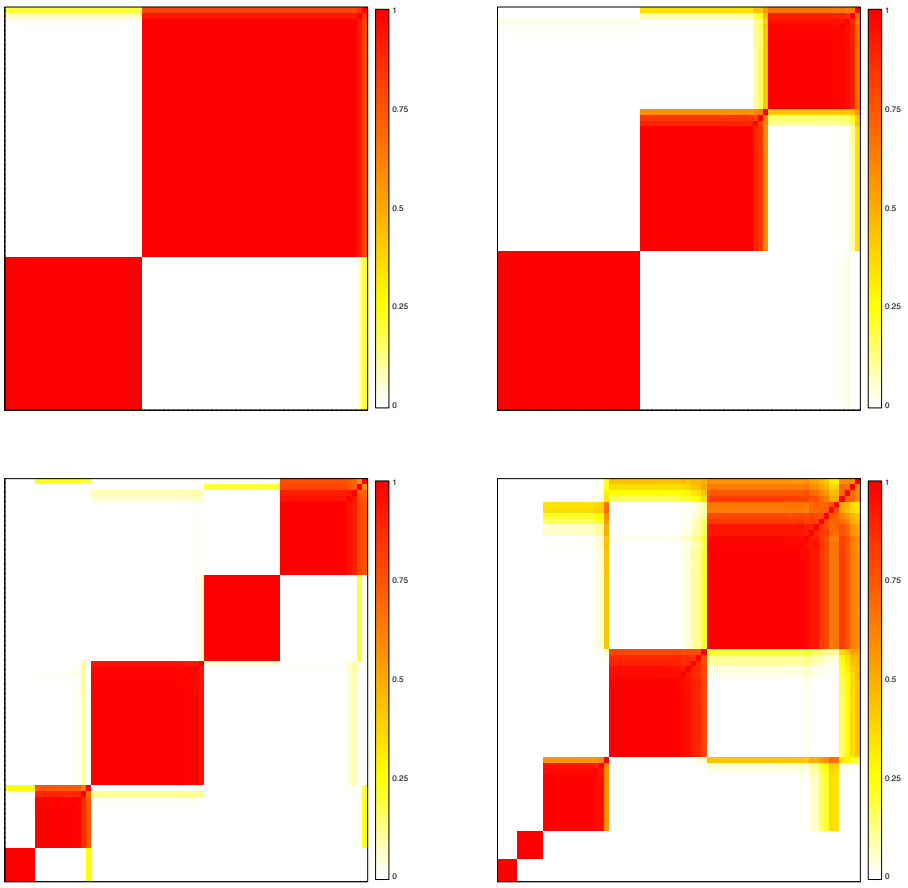

Figure 4.3: Mean posterior co-clustering probabilities for four out of the 201 observations of the network, $t = 16$ (top left), $t = 37$ (top right), $t = 98$ (bottom left), and $t = 171$ (bottom right).

community structure that coincides with the introduction of electronic trading. Prior to this event the community structure appears fairly stable, while some similarity is also present in the communities for the last two years of data.

Finally, we evaluate the predictive accuracy of the model by performing out-of-sample cross validation for the last ten observations. That is, for each $T = 191, 192, \ldots, 200$, the one-step-ahead prediction of Y_{T+1} is produce using the information contained in the set $\{Y_1, Y_2, \ldots, Y_T\}$ only. For each case the receiver operating characteristic (ROC) is computed

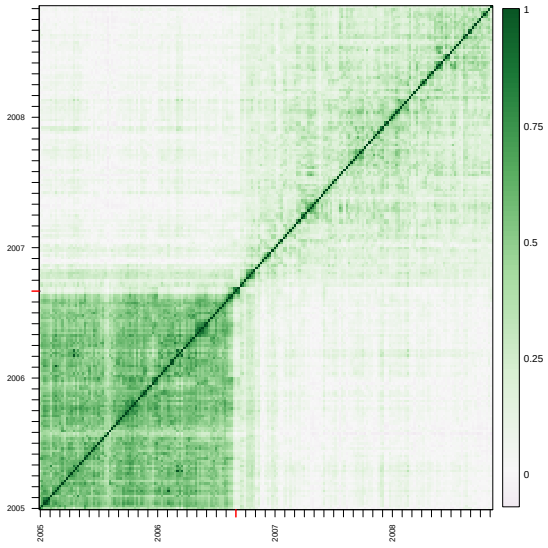

Figure 4.4: Matrix of pairwise adjusted Rand index. That is, for each observation the optimal community structure is obtained by fitting the model, and then for each pair (t, s) the ARI is computed from the inferred partitions.

along its corresponding area under the curve (AUC). These results are shown in Figure 4.5.

The predictive accuracy of the model averages 80% across the ten left-out weeks.

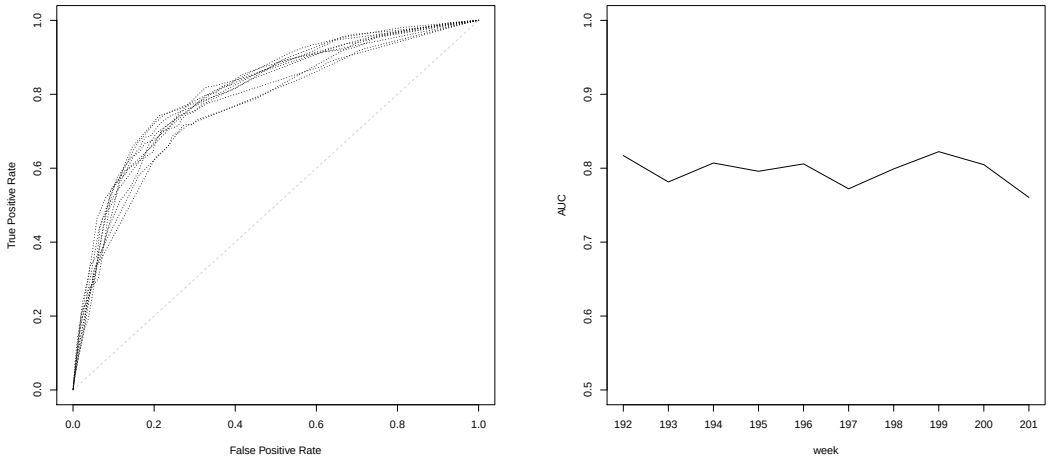

Figure 4.5: ROC (left) and corresponding AUC (right) for each of the ten cross validation points using the NYMEX dataset.

4.5 Discussion

In this section we presented an extension of the stochastic blockmodel that introduces dynamics both in the community indicators $\boldsymbol{\xi}_t$ and the community parameters Θ_t. We also described a Markov chain Monte Carlo algorithm to obtain approximate posterior inference from the model. Using simulated data, we observe that the model successfully recovers the community structure in a case in which the ground truth is known. Also, we have applied the model to a real financial trading network, where we see results that are in line with our knowledge of the network external to the dataset. Furthermore, even though is not the main focus in this setting we have observe that the model appears to retain an acceptable level of predictive accuracy.

Importantly, we note that the dynamics introduced in the community parameters do not lead to a stationary process, which would be a desirable property since, a priori, there is no reason to assume a different form in the interaction probabilities across time. Modifying the autoregressive structure to force a stationary process is one direction we would like to explore in future work. Other interesting generalizations of the model would be allowing α and β to change over time, as well as consider adding or dropping vertices in-between observations as a plausible change in a network.

Chapter 5

Conclusions

Throughout this work we have studied the problem of network community detection using stochastic blockmodels in different settings. Being able to partition a network into smaller groups of vertices –denominated communities– helps us uncover structural equivalences between the individuals in the network and, therefore, gain a better understanding of their structure and the system that the network represent.

In the fist part we explore a stochastic variational algorithm as an alternative to MCMC, the usual method for fitting stochastic blockmodels in a Bayesian framework. Here we note that, in fact, the reduction in computational time can be staggering; however, it might come at a price of significant loss int he accuracy of the solution.

Next we introduce an extension of the stochastic blockmodel that, using simple ideas from Bayesian hierarchical modeling, enables the recovery of multilevel community structures. A key feature of this model is its capability of simultaneously recovering assortative and disassortative mixing both at the community and the supercommunity level.

Finally we turn our attention to the problem of community detection in a dynamic setting. We introduce an extension of the stochastic blockmodel that allows for smooth transitions in the community structure by including a temporal dependency that accounts for changes in the community membership and, simultaneously, a dynamic structure in the community parameters that reflects on the evolution of the propensities of interaction.

Bibliography

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2009). Mixed membership stochastic blockmodels. In *Advances in Neural Information Processing Systems*, pages 33–40.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307.

Bertoin, J. (2006). *Random fragmentation and coagulation processes*, volume 102. Cambridge University Press.

Betancourt, B., Boyd, N., and Rodríguez, A. (2015). Modeling and prediction of financial trading networks: An application to the nymex natural gas futures market. Unpublished manuscript.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308.

Clauset, A., Moore, C., and Newman, M. E. (2007). Structural inference of hierarchies in networks. In *Statistical network analysis: models, issues, and new directions*, pages 1–13. Springer.

Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):066111.

Dawah, H. A., Hawkins, B. A., and Claridge, M. F. (1995). Structure of the parasitoid communities of grass-feeding chalcid wasps. *Journal of animal ecology*, pages 708–720.

Devroye, L. (2009). On exact simulation algorithms for some distributions related to jacobi theta functions. *Statistics & Probability Letters*, 79(21):2251–2259.

Durante, D. and Dunson, D. B. (2014). Nonparametric bayes dynamic modelling of relational data. *Biometrika*, 101(4):883–898.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, 6:290–297.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the american Statistical association*, 81(395):832–842.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6(6):721–741.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233.

Gopalan, P. K., Wang, C., and Blei, D. (2013). Modeling overlapping communities with node popularities. In *Advances in neural information processing systems*, pages 2850–2858.

Grassberger, P. (1983). On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157–172.

Guo, F., Hanneke, S., Fu, W., and Xing, E. P. (2007). Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th international conference on Machine learning*, pages 321–328. ACM.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

Hanneke, S., Fu, W., Xing, E. P., et al. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Ho, Q., Parikh, A. P., and Xing, E. P. (2012). A multiscale community blockmodel for network exploration. *Journal of the American Statistical Association*, 107(499):916–934.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Ishwaran, H. and Zarepour, M. (2000). Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5.

Knowles, D. A. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, pages 1701–1709.

Kurihara, K., Welling, M., and Teh, Y. W. (2007). Collapsed variational dirichlet process mixture models. In *IJCAI*, volume 7, pages 2796–2801.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Luce, R. D., Macy Jr, J., and Tagiuri, R. (1955). A statistical model for relational analysis. *Psychometrika*, 20(4):319–327.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104.

Newman, M. E. (2010). *Networks: an introduction*. Oxford University Press.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200.

Pérez, M.-E., Pericchi, L. R., and Ramírez, I. C. (2017). The scaled beta2 distribution as a robust prior for scales. *Bayesian Analysis*.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models

using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56(9):1082–1097.

Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rodríguez, A. (2012). Modeling the dynamics of social networks using bayesian hierarchical blockmodels. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(3):218–234.

Rodríguez, A. (2014). A bayesian nonparametric model for exchangeable multinetwork data based on fragmentation and coagulation processes. Unpublished manuscript.

Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4(61):76.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27–64.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657.

Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Westveld, A. H. and Hoff, P. D. (2011). A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, pages 843–872.

Xing, E. P., Fu, W., Song, L., et al. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566.

Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks–a bayesian approach. *Machine learning*, 82(2):157–189.

Appendix A

Details for MCMC algorithm for the multilevel stochastic blockmodel

The full conditionals distributions can be derived from (C.4). As a first step, let $\Upsilon_{-\theta_{k,l}}$ be the set of all parameters in the model with the exception of $\theta_{k,l}$. Then, for $1 \leq k \leq l \leq K$,

$$p(\theta_{k,l} \mid \Upsilon_{-\theta_{k,l}}, \Gamma, \mathcal{Y}) \propto \exp\left\{-\frac{1}{2}\left(\gamma_{k,l} + \frac{1}{\sigma^2}\right)\theta_{k,l}^2 + \left(s_{k,l} - \frac{n_{k,l}}{2} + \frac{\eta_{\phi(\zeta_k,\zeta_l)}}{\sigma^2}\right)\theta_{k,l}\right\} \quad \text{(A.1)}$$

which can be identified as a Gaussian kernel $\mathcal{N}(\mu^{\star}_{\theta_{k,l}}, \sigma^{2^{\star}}_{\theta_{k,l}})$ with mean parameter given by $\mu^{\star}_{\theta_{k,l}} = \left(\gamma_{k,l} + \frac{1}{\sigma^2}\right)^{-1}\left(s_{k,l} - \frac{n_{k,l}}{2} + \frac{\eta_{\phi(\zeta_k,\zeta_l)}}{\sigma^2}\right)$ and variance $\sigma^{2^{\star}}_{\theta_{k,l}} = \left(\gamma_{k,l} + \frac{1}{\sigma^2}\right)^{-1}$.

Now, for the auxiliary variables,

$$p(\gamma_{k,l} \mid \Upsilon, \Gamma_{-(k,l)}, \mathcal{Y}) \propto \exp\left\{-\frac{\theta_{k,l}^2}{2}\gamma_{k,l}\right\}\pi(\gamma_{k,l}) \quad \text{(A.2)}$$

that remains in the Polya-Gamma family, specifically, $\gamma_{k,l} \mid \Upsilon, \Gamma_{-(k,l)}, \mathcal{Y} \sim \mathcal{PG}(n_{k,l}, \theta_{k,l})$, and, thus, can be sampled using the approach proposed by Polson et al. (2013) that builds on Devroye (2009).

In the case of the community indicators,

$$Pr(\xi_i = k \mid \Upsilon_{-\xi_i}, \Gamma, \mathcal{Y}) \propto w_k \prod_{\substack{j=1 \\ j \neq i}}^{I} \frac{\left(\exp\{\theta_{\phi(\xi_j,k)}\}\right)^{y_{\phi(\xi_j,k)}}}{1 + \exp\{\theta_{\phi(\xi_j,k)}\}}, \tag{A.3}$$

which is a Categorical distribution.

For the first level variance parameter σ^2,

$$p(\sigma^2 \mid, \Upsilon_{-\sigma^2}, \Gamma, \mathcal{Y}) \propto (\sigma^2)^{-\left(\frac{1}{4}K(K+1)+\alpha_\sigma+1\right)} \exp\left\{ \frac{-\left[\frac{1}{2}\sum_{k=1}^{K}\sum_{l=k}^{K}(\theta_{k,l} - \eta_{\phi(\zeta_k,\zeta_l)})^2 + \beta_\sigma\right]}{\sigma^2} \right\} \tag{A.4}$$

which is an $\mathcal{IG}(\alpha_\sigma^\star, \beta_\sigma^\star)$ with $\alpha_\sigma^\star = \alpha_\sigma + \frac{1}{2}K(K+1)$ and $\beta_\sigma^\star = \frac{1}{2}\sum_{k=1}^{K}\sum_{l=k}^{K}(\theta_{k,l} - \eta_{\phi(\zeta_k,\zeta_l)})^2 + \beta_\sigma$. Notice, however, that it may be the case that some communities have no subjects assigned to them. Therefore, mixing in this sampling scheme can be improved by defining

$$x_{k,l} = \begin{cases} 1 & \text{if } n_{k,l} \geq 1 \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq k \leq l \leq K$ and $N = \sum_{k=1}^{K}\sum_{l=k}^{K} x_{k,l}$, and sample σ^2 from an $\mathcal{IG}(\alpha^{\star\star}, \beta^{\star\star})$ with parameters $\alpha_\sigma^{\star\star} = \frac{N+2\alpha_\sigma}{2}$ and $\beta_\sigma^{\star\star} = \frac{1}{2}\sum_{k=1}^{K}\sum_{l=k}^{K}(\theta_{k,l} - \eta_{\phi(\zeta_k,\zeta_l)})^2 x_{k,l} + \beta_\sigma$.

For the elements of H,

$$p(\eta_{r,s} \mid \Upsilon_{-\eta_{r,s}}, \Gamma, \mathcal{Y}) \propto \exp\left\{ -\frac{1}{2}\left(\frac{m_{r,s}}{\sigma^2} + \frac{1}{\tau^2}\right)\eta_{r,s}^2 + \left(\frac{t_{r,s}}{\sigma^2} + \frac{\mu}{\tau^2}\right)\eta_{r,s} \right\} \tag{A.5}$$

where $m_{r,s}$ and $t_{r,s}$ play the role of $n_{k,l}$ and $s_{k,l}$ respectively in the second level of the hierarchy. That is, $t_{r,s} = \sum_{\mathcal{T}_{r,s}} \theta_{k,l}$ while $m_{r,s} = \sum_{\mathcal{T}_{r,s}} 1$, with the sum taken over the set

$\mathcal{T}_{r,s} = \{(k,l) :\ k \leq l,\ (r,s) = \phi(\zeta_k, \zeta_l)\}$. Thus, $\eta_{r,s}$ can be sampled from of a $\mathcal{N}(\mu^{\star}_{\eta_{r,s}}, \tau^{2^{\star}}_{\eta_{r,s}})$ with mean $\mu^{\star}_{\eta_{r,s}} = \left(\frac{m_{r,s}}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{t_{r,s}}{\sigma^2} + \frac{\mu}{\tau^2}\right)$ and variance $\tau^{2^{\star}}_{\eta_{r,s}} = \left(\frac{m_{r,s}}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}$.

Similar to the case of $\boldsymbol{\xi}$, the full conditional for $\boldsymbol{\zeta}$ satisfies that

$$p(\boldsymbol{\zeta} \mid, \Upsilon_{-\boldsymbol{\zeta}}, \Gamma, \mathcal{Y}) \propto p(\boldsymbol{\zeta} \mid \boldsymbol{v}) p(\Theta \mid H, \boldsymbol{\zeta}, \sigma^2).$$

That is, for every $r = 1, 2, \ldots, R$, and every $k = 1, 2, \ldots, K$,

$$Pr(\zeta_k = r \mid \Upsilon_{-\zeta_k}, \Gamma, \mathcal{Y}) \propto v_r \exp\left\{-\frac{1}{2\sigma^2}\sum_{l=1}^{K}(\theta_{\phi(\xi_j,k)} - \eta_{\phi(r,\zeta_l)})^2\right\} \tag{A.6}$$

The full conditional for μ is given by

$$p(\mu \mid, \Upsilon_{-\mu}, \Gamma, \mathcal{Y}) \propto \exp\left\{-\frac{1}{2}\left(\frac{\frac{1}{2}R(R+1)}{\tau^2} + \frac{1}{\sigma^2_{\mu}}\right)\mu^2 + \mu\left(\frac{\sum_{r=1}^{R}\sum_{s=r}^{R}\eta_{r,s}}{\tau^2} + \frac{\mu_{\mu}}{\sigma^2_{\mu}}\right)\right\} \tag{A.7}$$

which is then sampled from a Gaussian $\mathcal{N}(\mu^{\star\star}_{\mu}, \sigma^{2^{\star\star}}_{\mu})$ with variance $\sigma^{2^{\star\star}}_{\mu} = \left(\frac{M}{\tau^2} + \frac{1}{\sigma^2_{\mu}}\right)^{-1}$, and mean $\mu^{\star\star}_{\mu} = \left(\frac{M}{\tau^2} + \frac{1}{\sigma^2_{\mu}}\right)^{-1}\left(\frac{\sum_{r=1}^{R}\sum_{s=r}^{R}\eta_{r,s}z_{r,s}}{\tau^2} + \frac{\mu_{\mu}}{\sigma^2_{\mu}}\right)$, and where, analogously to the first level, $M = \sum_{r=1}^{R}\sum_{s=r}^{R} z_{r,s}$ and, for $1 \leq r \leq s \leq R$

$$z_{r,s} = \begin{cases} 1 & \text{if } m_{r,s} \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

As in the case of σ^2, for τ^2

$$p(\tau^2 \mid, \Upsilon_{-\tau^2}, \Gamma, \mathcal{Y}) \propto (\tau^2)^{-\left(\frac{1}{4}R(R+1)+\alpha_{\tau}+1\right)} \exp\left\{\frac{-\left[\frac{1}{2}\sum_{r=1}^{R}\sum_{s=r}^{R}(\eta_{r,s}-\mu)^2 + \beta_{\tau}\right]}{\tau^2}\right\} \tag{A.8}$$

which is sampled from $\mathcal{IG}(\alpha^{\star\star}_{\tau}, \beta^{\star\star}_{\tau})$ with $\alpha^{\star\star}_{\tau} = \frac{M+2\alpha_{\tau}}{2}$ and $\beta^{\star\star}_{\tau} = \frac{1}{2}\sum_{r=1}^{R}\sum_{s=r}^{R}(\eta_{r,s} - \mu)^2 z_{r,s} + \beta_{\tau}$.

Now, the weights can be sampled from

$$p(\boldsymbol{w} \mid \Upsilon_{-\boldsymbol{w}}, \Gamma, \mathcal{Y}) \propto \prod_{k=1}^{K} w_k^{\frac{\alpha}{K}+n_k-1} \tag{A.9}$$

a Dirichlet with parameter vector $\left(\frac{\alpha}{K}+n_1, \frac{\alpha}{K}+n_2, \ldots, \frac{\alpha}{K}+n_K\right)$ and, similarly,

$$\boldsymbol{v} \mid \Upsilon_{-\boldsymbol{v}}, \Gamma, \mathcal{Y} \sim \mathcal{D}ir\left(\frac{\beta}{R}+m_1, \frac{\beta}{R}+m_2, \ldots, \frac{\beta}{R}+m_R\right).$$

Finally, for the concentration parameters,

$$p(\alpha \mid \Upsilon_{-\alpha}, \Gamma, \mathcal{Y}) \propto \frac{\Gamma(\alpha)}{\left[\Gamma\left(\frac{\alpha}{K}\right)\right]^K} \alpha^{\alpha_\alpha - 1} \exp\{-\beta_\alpha \alpha\} \tag{A.10}$$

that does not lead to a closed form. Thus, α could be sampled using a Metropolis-Hastings step with, for example, a random walk on the log scale. Alternatively, by recognizing that for large enough K this distribution approximates the full conditional for the concentration parameters of a Dirichlet process, α can be sampled from a mixture of Gammas borrowing from Escobar and West (1995). The case of β is analogous for large enough R.

Appendix B

Details for the variational Bayes algorithm for the multilevel stochastic blockmodel

The optimal variational distribution for the community indicators is Categorical satisfying:

$$\log q^{\star}(\xi_i = k) = \mathbb{E}_{q(w_k)}\left[\log w_k\right] + \sum_{j \neq i} y_{\phi(i,j)} \sum_{l=1}^{K} \mathbb{E}_{q(\theta_{\phi(k,l)})}\left[\theta_{\phi(k,l)}\right] q^{\star}(\xi_j = l) + \sum_{j \neq i} \sum_{l=1}^{K} \mathbb{E}_{q(\theta_{\phi(k,l)})}\left[\log\left(1 + \exp\left\{\theta_{\phi(k,l)}\right\}\right)\right] q^{\star}(\xi_j = l) + C. \quad \text{(B.1)}$$

For the first level variance,

$$\log q^{\star}(\sigma^2) = -\left(\frac{1}{4}K(K+1) + \alpha_{\sigma} + 1\right)\log \sigma^2 - \frac{\frac{1}{2}\sum_{k=1}^{K}\sum_{l=k}^{K}\mathbb{E}_{q(\Theta,H,\zeta)}\left[(\theta_{k,l} - \eta_{\phi(\zeta_k,\zeta_l)})^2\right] + \beta_{\sigma}}{\sigma^2} + C \quad \text{(B.2)}$$

which is readily identified as an Inverse Gamma distribution.

The distribution for the location parameters H is also approximated by a Gaussian, in this case

$$\log q^{\star}(\eta_{r,s}) = -\frac{1}{2}\left\{\mathbb{E}_{q(\sigma)}\left[\frac{1}{\sigma^2}\right]\mathbb{E}_{q(\boldsymbol{\zeta})}\left[m_{r,s}\right] + \mathbb{E}_{q(\tau)}\left[\frac{1}{\tau^2}\right]\right\}\eta_{r,s}^2 + \left\{\mathbb{E}_{q(\sigma)}\left[\frac{1}{\sigma^2}\right]\mathbb{E}_{q(\boldsymbol{\zeta})}\left[t_{r,s}\right] + \mathbb{E}_{q(\tau)}\left[\frac{1}{\tau^2}\right]\mathbb{E}_{q(\mu)}\left[\mu\right]\right\}\eta_{r,s} + C. \quad \text{(B.3)}$$

Similar to the case of $\boldsymbol{\xi}$, the supercommunity indicators follow a Categorical distribution with probabilities given by

$$\log q^{\star}(\zeta_k = r) = \mathbb{E}_{q(v_r)}\left[\log v_r\right] - \frac{1}{2}\mathbb{E}_{q(\sigma)}\left[\frac{1}{\sigma^2}\right]\sum_{l\neq k}\mathbb{E}_{q(\Theta,H,\boldsymbol{\zeta})}\left[\left(\theta_{\phi(k,l)} - \eta_{\phi(\zeta_k,\zeta_l)}\right)^2\right] + C. \quad \text{(B.4)}$$

Now, for the second level mean parameter,

$$\log q^{\star}(\mu) = -\frac{1}{2}\left\{\frac{1}{2}R(R+1)\mathbb{E}_{q(\tau)}\left[\frac{1}{\tau^2}\right] + \frac{1}{\sigma_{\mu}^2}\right\}\mu^2 + \left\{\mathbb{E}_{q(\tau)}\left[\frac{1}{\tau^2}\right]\sum_{r=1}^{R}\sum_{s=r}^{R}\mathbb{E}_{q(H)}\left[\eta_{r,s}\right] + \frac{\mu_{\mu}}{\sigma_{\mu}^2}\right\}\mu + C, \quad \text{(B.5)}$$

again, a Gaussian distribution, while the variational distribution for the variance parameter is an Inverse Gamma such that

$$\log q^{\star}(\tau^2) = -\left(\frac{1}{4}R(R+1) + \alpha_{\tau} + 1\right)\log\tau^2 - \frac{\frac{1}{2}\sum_{r=1}^{R}\sum_{s=r}^{R}\mathbb{E}_{q(H,\mu)}\left[(\eta_{r,s}-\mu)^2\right] + \beta_{\tau}}{\tau^2} + C. \quad \text{(B.6)}$$

With respect to the weight vectors, it can be observed that

$$\log q^{\star}(\mathbf{w}) = \sum_{k=1}^{K}\left(\mathbb{E}_{q(\boldsymbol{\xi})}\left[n_k\right] + \frac{1}{K}\mathbb{E}_{q(\alpha)}\left[\alpha\right] - 1\right)\log w_k + C \quad \text{(B.7)}$$

and

$$\log q^{\star}(\mathbf{v}) = \sum_{r=1}^{R}\left(\mathbb{E}_{q(\boldsymbol{\zeta})}\left[m_r\right] + \frac{1}{R}\mathbb{E}_{q(\beta)}\left[\beta\right] - 1\right)\log v_r + C \quad \text{(B.8)}$$

which remain in the Dirichlet distribution.

Finally, for the first level concentration parameter α,

$$\log q^{\star}(\alpha) = \log\Gamma(\alpha) - K\log\Gamma\left(\frac{\alpha}{K}\right) + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{q(w_k)}\left[\log w_k\right] + (\alpha_\alpha - 1)\log\alpha - \beta_\alpha\alpha + C \quad \text{(B.9)}$$

which, using that $\log\Gamma(x) = (x-1)\log x + \mathcal{O}(\log x)$, can be approximated by

$$\log\tilde{q}(\alpha) = \log\alpha\,(\alpha_\alpha + K - 1 - 1) - \alpha\left(\beta_\alpha - \frac{1}{K}\sum_{k=1}^{K}\left\{\mathbb{E}_{q(w_k)}\left[\log w_k\right] - \log\frac{1}{K}\right\}\right) + C \quad \text{(B.10)}$$

and, analogously,

$$\log\tilde{q}(\beta) = \log\beta\,(\alpha_\beta + R - 1 - 1) - \beta\left(\beta_\beta - \frac{1}{R}\sum_{r=1}^{R}\left\{\mathbb{E}_{q(v_r)}\left[\log v_r\right] - \log\frac{1}{R}\right\}\right) + C. \quad \text{(B.11)}$$

From equations (3.13) to (B.11) is possible to derive the complete algorithm with the respective optimal variational parameters as shown below

$$\tilde{q}(\theta_{k,l}) = \mathcal{N}(m_{k,l}, \nu_{k,l}) \quad \text{(B.12)}$$

with $m_{k,l} = \nu_{k,l}\left[\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} y_{i,j}\epsilon_{k,l}^{i,j} - \frac{1}{2}\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\epsilon_{k,l}^{i,j} + \frac{o}{p}\sum_{r=1}^{R}\sum_{s=r}^{R} g_{r,s}\delta_{r,s}^{k,l}\right]$ and

$\nu_{k,l} = \left[\frac{1}{2}\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\epsilon_{k,l}^{i,j}\left(\frac{1-\exp\{-\gamma_{k,l}\}}{\gamma_{k,l}(1+\exp\{\gamma_{k,l}\})}\right) + \frac{o}{p}\right]^{-1}$, while $\gamma_{k,l}^2 = \nu_{k,l} + m_{k,l}^2$ where

$$\epsilon_{k,l}^{i,j} = \begin{cases} q(\xi_i = k)q(\xi_j = l) + q(\xi_i = l)q(\xi_j = k) & \text{if } k \neq l \\ q(\xi_i = k)q(\xi_j = l) & \text{if } k = l. \end{cases}$$

and analogously,

$$\delta_{r,s}^{k,l} = \begin{cases} q(\zeta_k = r)q(\zeta_l = s) + q(\zeta_k = s)q(\zeta_l = r) & \text{if } r \neq s \\ q(\zeta_k = l)q(\zeta_l = s) & \text{if } r = s. \end{cases}$$

While, for the community indicators

$$q^{\star}(\xi_i = k) = \varpi_{i,k} \tag{B.13}$$

with

$$\log \varpi_{i,k} = \boldsymbol{\Psi}\left(\psi_k\right) + \sum_{j \neq i}\left[y_{\phi(i,j)} \sum_{l=1}^{K} m_{\phi(k,l)} \varpi_{j,l} - \sum_{l=1}^{K}\left(\log\left(1 + \exp\left\{m_{k,l} + \frac{1}{2}\nu_{k,l}\right\}\right) - \frac{\left(\exp\{\nu_{k,l}\} - 1\right)\exp\{2m_{k,l} + \nu_{k,l}\}}{2\left(1 + \exp\left\{m_{k,l} + \frac{1}{2}\nu_{k,l}\right\}\right)^2}\right)\varpi_{j,l}\right] + C$$

and where it has been made use of the fact that, from the second order Delta method, $\mathbb{E}_{q(\theta_{k,l})}\left[\log(1+\exp\theta_{k,l})\right] \approx \log\left(1 + \exp\left\{m_{k,l} + \frac{1}{2}\nu_{k,l}\right\}\right) - \frac{\left(\exp\{\nu_{k,l}\}-1\right)\exp\{2m_{k,l}+\nu_{k,l}\}}{2\left(1+\exp\left\{m_{k,l}+\frac{1}{2}\nu_{k,l}\right\}\right)^2}$.

In the case of the first level variance parameter,

$$q^{\star}(\sigma^2) = \mathcal{IG}(o, p) \tag{B.14}$$

with $p = \beta_{\sigma} + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=k}^{K}\left(\nu_{k,l} + m_{k,l}^2\right) - \sum_{r=1}^{R}\sum_{s=r}^{R} g_{r,s}\sum_{k=1}^{K}\sum_{l=k}^{K} m_{k,l}\delta_{r,s}^{k,l} + \frac{1}{2}\sum_{r=1}^{R}\sum_{s=r}^{R}\left(g_{r,s}^2 + h_{r,s}^2\right)\sum_{k=1}^{K}\sum_{l=k}^{K}\delta_{r,s}^{k,l}$ and $o = \alpha_{\sigma} + \frac{1}{4}K(K+1)$.

For the upper level, the location parameters satisfy

$$q^{\star}(\eta_{r,s}) = \mathcal{N}(g_{r,s}, h_{r,s}^2) \tag{B.15}$$

where $g_{r,s} = h_{r,s}^2\left[\frac{o}{p}\sum_{k=1}^{K}\sum_{l=k}^{K} m_{k,l}\delta_{r,s}^{k,l} + \frac{a}{b}c\right]$ and $h_{r,s}^2 = \left[\frac{o}{p}\sum_{k=1}^{K}\sum_{l=k}^{K}\delta_{r,s}^{k,l} + \frac{a}{b}\right]^{-1}$, and the supercommunity indicators probabilities are such that

$$q^{\star}(\zeta_k = r) = \varrho_{k,r} \tag{B.16}$$

with

$$\log \varrho_{k,r} = \boldsymbol{\Psi}\left(\varphi_r\right) - \frac{1}{2}\left(\frac{o}{p}\right)\sum_{l=1}^{K}\left[\nu_{\phi(k,l)} + m_{\phi(k,l)}^2 - 2m_{\phi(k,l)}\sum_{s=1}^{R} g_{\phi(r,s)}\varrho_{l,s} + \sum_{s=1}^{R}\left(h_{\phi(r,s)}^2 + g_{\phi(r,s)}^2\right)\right] + C.$$

In the case of the hyperparameters, the optimal variational distribution for the mean is

$$q^{\star}(\mu) = \mathcal{N}(c, d^2) \tag{B.17}$$

where $c = d^2 \left[\frac{a}{b} \sum_{r=1}^{R} \sum_{s=r}^{R} g_{r,s} + \frac{\mu_\mu}{\sigma_\mu^2} \right]$ and $d^2 = \left[\left(\frac{a}{b} \right) \frac{1}{2} R(R+1) + \frac{1}{\sigma_\mu^2} \right]^{-1}$, while for the variance

$$q^\star(\tau^2) = \mathcal{IG}(a, b) \tag{B.18}$$

with $b = \beta_\tau + \frac{1}{2} \sum_{r=1}^{R} \sum_{s=r}^{R} \left(h_{r,s}^2 + g_{r,s}^2 \right) - c \sum_{r=1}^{R} \sum_{s=r}^{R} g_{r,s} + \frac{1}{4} R(R+1) \left(d^2 + c^2 \right)$
and $a = \alpha_\tau + \frac{1}{4} R(R+1)$.

Lastly, for the weight parameters on the indicators side,

$$q^\star(\mathbf{w}) = \mathcal{D}ir(\boldsymbol{\psi}) \tag{B.19}$$

with $\psi_k = \frac{1}{K} \left(\frac{a_\alpha}{b_\alpha} \right) + \sum_{i=1}^{I} \varpi_{i,k}$, and

$$q^\star(\mathbf{v}) = \mathcal{D}ir(\boldsymbol{\varphi}) \tag{B.20}$$

with $\varphi_r = \frac{1}{R} \left(\frac{a_\beta}{b_\beta} \right) + \sum_{k=1}^{K} \varrho_{k,r}$,

while their respective hyperparameters satisfy

$$\tilde{q}(\alpha) = \mathcal{G}(a_\alpha, b_\alpha) \tag{B.21}$$

with parameters $a_\alpha = \alpha_\alpha + K - 1$ and $b_\alpha = \beta_\alpha - \frac{1}{K} \sum_{k=1}^{K} \left[\boldsymbol{\Psi}(\psi_k) - \boldsymbol{\Psi} \left(\sum_{l=1}^{K} \psi_l \right) - \log \left(\frac{1}{K} \right) \right]$,

and

$$\tilde{q}(\beta) = \mathcal{G}(a_\beta, b_\beta) \tag{B.22}$$

where $a_\beta = \alpha_\beta + R - 1$ and $b_\beta = \beta_\beta - \frac{1}{R} \sum_{r=1}^{R} \left[\boldsymbol{\Psi}(\varphi_r) - \boldsymbol{\Psi} \left(\sum_{s=1}^{R} \varphi_s \right) - \log \left(\frac{1}{R} \right) \right]$.

Finally, it is found that the ELBO satisfies

$$\tilde{F}(q,\mathcal{Y}) \approx \sum_{i<j}\Bigg[y_{i,j}\sum_{k\leq l} m_{k,l}\epsilon_{k,l}^{i,j} - \sum_{k\leq l}\log\left(1+\exp\left\{-\gamma_{k,l}\right\}\right)\epsilon_{k,l}^{i,j} - \frac{1}{2}\left(m_{k,l}+\gamma_{k,l}\right)\epsilon_{k,l}^{i,j}$$
$$-\frac{1-\exp\left\{-\gamma_{k,l}\right\}}{4\gamma_{k,l}\left(1+\exp\left\{-\gamma_{k,l}\right\}\right)}\left(m_{k,l}^2+v_{k,l}-\gamma_{k,l}^2\right)\epsilon_{k,l}^{i,j}\Bigg] - \frac{K(K+1)}{4}\left(\log p - \boldsymbol{\Psi}(o)\right)$$
$$-\frac{o}{2p}\sum_{k\leq l}\left[\left(m_{k,l}^2+v_{k,l}\right) - 2m_{k,l}\sum_{r\leq s} g_{r,s}\delta_{r,s}^{k,l} + \sum_{r\leq s}\left(g_{r,s}^2+h_{r,s}^2\right)\delta_{r,s}^{k,l}\right] + \sum_{k=1}^{K}\left[\boldsymbol{\Psi}(\psi_k)\sum_{i=1}^{I}\varpi_{i,k}\right]$$
$$-I\boldsymbol{\Psi}\left(\sum_{k=1}^{K}\psi_k\right) + \alpha_\sigma\log\beta_\sigma - \log\Gamma(\alpha_\sigma) - (\alpha_\sigma-1)(\log p - \boldsymbol{\Psi}(o)) - \beta_\sigma\frac{o}{p} - \frac{R(R+1)}{4}\left(\log b - \boldsymbol{\Psi}(a)\right)$$
$$-\frac{a}{2b}\left[\sum_{r\leq s}\left(g_{r,s}^2+h_{r,s}^2\right) - 2c\sum_{r\leq s} g_{r,s} + \frac{1}{2}R\left(R+1\right)\left(c^2+d^2\right)\right] + \sum_{r=1}^{R}\left[\boldsymbol{\Psi}(\varphi_r)\sum_{k=1}^{K}\varrho_{k,r}\right]$$
$$-K\boldsymbol{\Psi}\left(\sum_{r=1}^{R}\varphi_r\right) - \frac{1}{2}\log(\sigma_\mu^2) - \frac{c^2+d^2-2c\mu_\mu+\mu_\mu^2}{\sigma_\mu^2} + \alpha_\tau\log\beta_\tau - \log\Gamma(\alpha_\tau) - (\alpha_\tau-1)(\log b - \boldsymbol{\Psi}(a))$$
$$-\beta_\tau\frac{a}{b} + \log\Gamma\left(\frac{a_\beta}{b_\beta}\right) + \frac{1}{2}\boldsymbol{\Psi}_1\left(\frac{a_\beta}{b_\beta}\right)\frac{a_\beta}{b_\beta^2} - R\left[\log\Gamma\left(\frac{a_\beta}{Rb_\beta}\right) + \frac{1}{2}\boldsymbol{\Psi}_1\left(\frac{a_\beta}{Rb_\beta}\right)\frac{a_\beta}{R^2b_\beta^2}\right]$$
$$+\sum_{r=1}^{R}\left(\frac{a_\beta}{Rb_\beta}-1\right)\left(\boldsymbol{\Psi}(\varphi_r) - \boldsymbol{\Psi}\left(\sum_{s=1}^{R}\varphi_s\right)\right) + \log\Gamma\left(\frac{a_\alpha}{b_\alpha}\right) + \frac{1}{2}\boldsymbol{\Psi}_1\left(\frac{a_\alpha}{b_\alpha}\right)\frac{a_\alpha}{b_\alpha^2}$$
$$-K\left[\log\Gamma\left(\frac{a_\alpha}{Kb_\alpha}\right) + \frac{1}{2}\boldsymbol{\Psi}_1\left(\frac{a_\alpha}{Kb_\alpha}\right)\frac{a_\alpha}{K^2b_\alpha^2}\right] + \sum_{k=1}^{K}\left(\frac{a_\alpha}{Kb_\alpha}-1\right)\left(\boldsymbol{\Psi}(\psi_k) - \boldsymbol{\Psi}\left(\sum_{l=1}^{K}\psi_l\right)\right)$$
$$+\alpha_\alpha\log\beta_\alpha - \log\Gamma(\alpha_\alpha) + (\alpha_\alpha-1)(\boldsymbol{\Psi}(a_\alpha) - \log b_\alpha) - \beta_\alpha\frac{a_\alpha}{b_\alpha} + \alpha_\beta\log\beta_\beta - \log\Gamma(\alpha_\beta)$$
$$+(\alpha_\beta-1)(\boldsymbol{\Psi}(a_\beta) - \log b_\beta) - \beta_\beta\frac{a_\beta}{b_\beta} + \frac{K(K+1)}{4} + \frac{1}{2}\sum_{k\leq l}\log(v_{k,l}) - \sum_{i=1}^{I}\sum_{k=1}^{K}\varpi_{i,k}\log\varpi_{i,k}$$
$$+o+\log p + \log\Gamma(o) - (1+o)\boldsymbol{\Psi}(o) + \frac{R(R+1)}{4} + \frac{1}{2}\sum_{r\leq s}\log(h_{r,s}^2) - \sum_{k=1}^{K}\sum_{r=1}^{R}\varrho_{k,r}\log\varrho_{k,r}$$
$$+\frac{1}{2}\left(1+\log d^2\right) + a + \log b + \log\Gamma(a) - (1+a)\boldsymbol{\Psi}(a) + \sum_{k=1}^{K}\log\Gamma(\psi_k) - \log\Gamma\left(\sum_{k=1}^{K}\psi_k\right)$$
$$-\left(K-\sum_{k=1}^{K}\psi_k\right)\boldsymbol{\Psi}\left(\sum_{k=1}^{K}\psi_k\right) - \sum_{k=1}^{K}\left(\psi_k-1\right)\boldsymbol{\Psi}(\psi_k) + \sum_{r=1}^{R}\log\Gamma(\varphi_r) - \log\Gamma\left(\sum_{r=1}^{R}\varphi_r\right)$$
$$-\left(R-\sum_{r=1}^{R}\varphi_r\right)\boldsymbol{\Psi}\left(\sum_{r=1}^{R}\varphi_r\right) - \sum_{r=1}^{R}\left(\varphi_r-1\right)\boldsymbol{\Psi}(\varphi_r) + a_\alpha - \log b_\alpha + \log\Gamma(a_\alpha) + (1-a_\alpha)\boldsymbol{\Psi}(a_\alpha)$$
$$+a_\beta - \log b_\beta + \log\Gamma(a_\beta) + (1-a_\beta)\boldsymbol{\Psi}(a_\beta)$$

which is used to monitor for convergence of the algorithm.

Appendix C

Details for derivation of expected number of clusters in the multilevel stochastic blockmodel

Direct calculation from the distribution of $K^\star$ yields that

$$\mathbb{E}\left[K^\star\right] = \alpha\left\{\Psi(\alpha + I) - \Psi(\alpha)\right\} \approx \alpha \log\left(\frac{\alpha + I}{\alpha}\right), \tag{C.1}$$

and

$$\mathbb{V}\left[K^\star\right] = \alpha\left\{\Psi(\alpha + I) - \Psi(\alpha)\right\} + \alpha^2\left\{\Psi'(\alpha + I) - \Psi'(\alpha)\right\} \approx \alpha \log\left(\frac{\alpha + I}{\alpha}\right). \tag{C.2}$$

Therefore, using the law of iterated expectations,

$$\mathbb{E}\left[R^\star\right] = \mathbb{E}\left[\mathbb{E}\left[R^\star \mid K^\star\right]\right] \approx \mathbb{E}\left[\beta \log\left(\frac{\beta + K^\star}{\beta}\right)\right] = \beta\mathbb{E}\left[\log\left(\frac{\beta + K^\star}{\beta}\right)\right] \tag{C.3}$$

And applying a second order Taylor approximation yields the desired result:

$$\mathbb{E}\left[R^{\star}\right] \approx \beta \left\{ \log\left(\mathbb{E}\left[\frac{\beta + K^{\star}}{\beta} \right] \right) - \frac{1}{2} \frac{\mathbb{V}\left[\frac{\beta + K^{\star}}{\beta} \right]}{\mathbb{E}^{2}\left[\frac{\beta + K^{\star}}{\beta} \right]} \right\} \tag{C.4}$$

$$= \beta \left\{ \log\left(\frac{\beta + \mathbb{E}\left[K^{\star}\right]}{\beta} \right) - \frac{1}{2} \frac{\frac{1}{\beta^{2}} \mathbb{V}\left[K^{\star}\right]}{\left(\frac{\beta + \mathbb{E}[K^{\star}]}{\beta} \right)^{2}} \right\} \tag{C.5}$$

$$\approx \beta \left\{ \log\left(\frac{\beta + \alpha \log\left(\frac{\alpha + I}{\alpha} \right)}{\beta} \right) - \frac{1}{2} \frac{\alpha \log\left(\frac{\alpha + I}{\alpha} \right)}{\left(\beta + \alpha \log\left(\frac{\alpha + I}{\alpha} \right) \right)^{2}} \right\}. \tag{C.6}$$