Bayesian Analysis (2004) **1**, Number 1, pp. 1–23

Stochastic Block Models of Mixed Membership

Edoardo M. Airoldi
School of Computer Science
Carnegie Mellon University

David M. Blei
Department of Computer Science
Princeton University

Stephen E. Fienberg
Department of Statistics, and
School of Computer Science
Carnegie Mellon University

Eric P. Xing
School of Computer Science
Carnegie Mellon University

Abstract. We consider the statistical analysis of a collection of unipartite graphs, i.e., multiple matrices of relations among objects of a single type. Such data arise, for example, in biological settings, collections of author-recipient email, and social networks. In such applications, typical analyses aim at: (i) clustering the objects of study or situating them in a low dimensional space, e.g., a simplex; and (ii) estimating relational structures among the clusters themselves. For example, in biological applications we are interested in estimating how stable protein complexes (i.e., clusters of proteins) interact. To support such integrated data analyses, we develop the family of *stochastic block models of mixed membership*. Our models combine features of mixed-membership models (Erosheva and Fienberg 2005) and block models for relational data (Holland et al. 1983) in a hierarchical Bayesian framework. We develop a *nested* variational inference scheme, which is necessary to successfully perform fast approximate posterior inference in our models of relational data. We present evidence to support our claims, using both real and synthetic data.

Keywords: hierarchical Bayesian models of mixed membership, mean-field approximation, naïve variational inference, nested variational inference, statistical network analysis, social and biological networks

1 Introduction

In many applications we wish to analyze observations about attribute measurements corresponding to pairs of objects of the same type, e.g., the presence of an interaction between a pair of proteins, or the number of papers where an author cites another author. In this case it is common to call the attributes "relations". Relations can be symmetric (e.g., two proteins reciprocate an interaction) or not (e.g., an author unilaterally decides to cite another author), depending on the semantics of the specific application we consider. We distinguish such applications from those where we wish to analyze observations about attribute measurements corresponding to individual objects of the same type. Observations are not paired in these applications, that is, measurements about the set attributes refer to a single object. For example, in the application to document analysis we measure the number of times each word is used by individual authors, independently of the others.

In such applications, typical analyses aim at: (i) clustering the objects of study or

situating them in a low dimensional space, e.g., a simplex; and (ii) estimating relational structures among the clusters themselves. For example, in biological applications we are interested in performing two tasks: identifying stable protein complexes, i.e., clusters of proteins, and estimating how such complexes interact with one another. In social network analysis the two tasks above translate in to identifying groups of people, and estimating how groups themselves communicate, from observations about email communications. This latter piece of information may reveal, for example, the informal structure of an organization.

To support such integrated data analyses, we introduce the family of *stochastic block models of mixed membership*. Models in this family combine features of mixed-membership models (Erosheva 2003; Erosheva and Fienberg 2005) and block models for relational data (Holland et al. 1983; Anderson et al. 1992; Nowicki and Snijders 2001) in a hierarchical Bayesian framework.

In this paper we make the following contributions: (a) we introduce stochastic block models of mixed membership, a subset of latent variable models for analyzing relational data, i.e., directed unipartite graphs with possibly weighted edges; (b) we subsume possible full model specifications into a general formulation, which is amenable to theoretical analysis; and (c) we develop a *nested* variational inference scheme, which is necessary to successfully perform fast approximate posterior inference in our models of relational data, which does not depend on the support of the data, and which scales to large problems. We explore theoretical and computational issues associated with these models via simulations and a biological case study.

2 The Scientific Problem

Here we describe the relational data we are interested in modeling, and highlight the differences with non-relational data. We describe the goals of the analysis, and we conclude with a general formulation of the problem.

2.1 Relational versus non-Relational Data

Relational data are directed, unipartite graphs. A unipartite graph, is a graph whose constituent nodes are of a single type, e.g., proteins; as opposed to bipartite and multi-partite graphs, whose constituent nodes are of two and of multiple types, respectively, e.g., authors and words, or employees, tasks and resources.

The typical independence and exchangeability assumptions adopted for non-relational data may not be appropriate for relational data. For example, attributes measurements at each object may be treated as independent or exchangeable, but it seems that a different set of assumptions is needed to model relations among pairs of objects. Figure 1 provides a brief illustration of this point. Furthermore, note that arguments about the appropriateness of assumptions do not depend on a specific goal of the analysis.

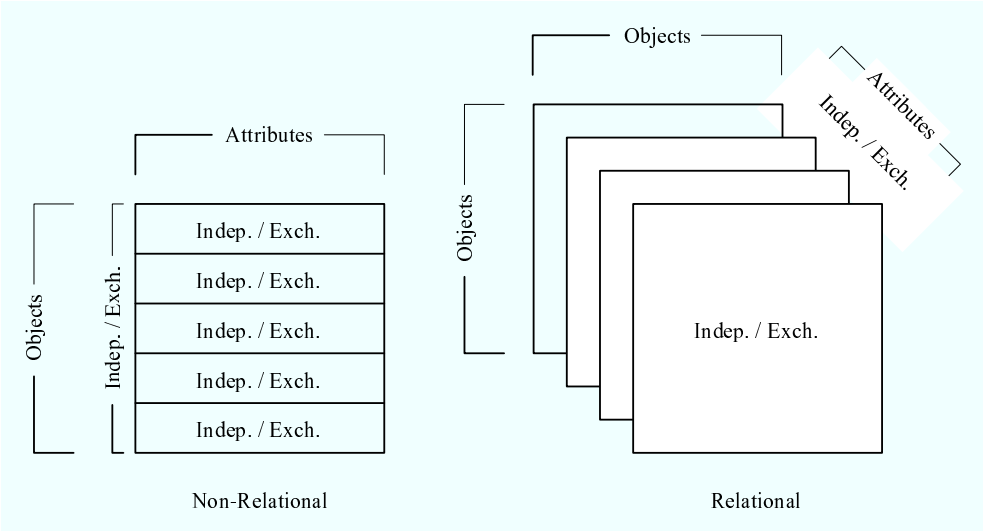

Figure 1: A bipartite graph (left panel) and a four unipartite graphs (right panel). In a unipartite graph setting there is one more dimension to the data, which is given by the observations on pairs of objects for each attribute; attributes measurements at each pairs may be treated as independent, but a set of extra assumptions is needed to model relations among pairs of objects. This fact suggests that statistical models and assumptions for non-relational data may not be appropriate for relational data.

2.2 Four Examples

Here we describe the essential characteristics of four example applications, which outline the different kinds of data we can analyze with our models.

Example 1. Consider the set of hand-curated protein interactions produced by the Munich Institute for Protein Sequencing (Mewes et al. 2004). A single set of interactions between proteins has been experimentally verified. Information about this unique, symmetric relation can be stored in one square table, whose entries are random variables with support $\{0, 1\}$ that encode presence or absence of an interaction between corresponding pairs of proteins.

Example 2. Consider the output of a battery of microarray experiments, on the same set of genes, $\mathcal{N}$, under different, R, experimental conditions, in Yeast (Krogan et al. 2006). Without entering into biological details[1], we wish to analyze probabilities of interactions between pairs of proteins[2], which are induced from correlations found in the gene expression experiments (Bhardwaj and Lu 2005). Information about this

[1]However, it is not to be excluded that such biological details may suggest alternative statistical analyses, on data with different degrees of pre-processing, as more desirable.

[2]Proteins are uniquely identified by genes in the microarray experiments, so far as the "one gene – one protein" dogma of molecular biology holds, in Yeast.

unique, symmetric relation can be stored in a collection of R square tables, one for each experimental condition, whose entries are random variables with support $[0, 1]$ that encode the probability of an interaction between corresponding pairs of proteins.

Example 3. Consider a collection of email communications within a company, say, Enron. Our observations consists of weekly summaries about how many emails each pair of employees exchange (Priebe et al. 2005). Information about this unique, asymmetric relation can be stored in a collection of R square tables, one for each week, whose entries are random variables with integer, non-negative support that encode the number of emails sent-received by the corresponding pair of employees.

Example 4. Consider a collection of sociometric relations among a group of monks (Sampson 1968). We observe responses to questions about J distinct social relations between pairs of monks, e.g., "Do you like X?" or "Do you trust X?", and the questionnaire is repeated at R_j epochs for each social relation, where $j = 1, \ldots, J$. The relations are asymmetric by design. Information about these repeated, asymmetric relations can be stored in a collection of $\sum_j R_j$ square tables, R_j replicates for each distinct social relation, j, whose entries are random variables with support $\{0, 1\}$ that encode binary responses of a monk regarding another monks.

From a modeling perspective, it is useful to give an abstract representation of the data we plan to analyze. Say we observe a collection of unipartite graphs, whose edge encode measurements on pair of nodes according to different, J, response variables, and we observe multiple, R_j, replicates of each graph,

$$\mathcal{G} = \{\ G_{jr} : j = 1, \ldots, J, \text{ and } r = 1, \ldots, R_j\ \}$$

where each graph $G_{jr} = (Y_{jr}, \mathcal{N})$, is defined over a common set of nodes, $\mathcal{N}$. The random quantities that encode the edge weights, e.g., Y_{jrnm}, where (n, m) is a pair of nodes in $\mathcal{N}$, have support in a separable, metric space. It is possible that each of the J response variables has support in a different spaces. The collection contains $\sum_j R_j$ graphs in total. Such a collection may contain missing values.

2.3 The Goals of the Analysis

There are three main goals: (1) identifying clusters of nodes; (2) determining the number of clusters; (3) estimating the probabilities of interaction among clusters.

Let us consider the first protein example above. We analyze the set of protein-protein interactions with the goal of identifying stable protein complexes, i.e., clusters of proteins, since they have been shown to be important for carrying out cellular processes (Krogan et al. 2006). Further, we want to know how many protein complexes are needed to explain the collection of protein interactions. Last, we want to estimate the probabilities according to which pairs of such protein complexes interact with one another.

3 Stochastic Block Models of Mixed Membership

The challenge is then to posit a rich class of models that is instrumental for thinking about the scientific problems we outlined above. On the other hand, it is desirable for such a formulation to be amenable to theoretical analysis, and for it to capture specificities of the relational data that are not shared by non-relational data. The class of stochastic block models of mixed-membership satisfies our desiderata.

In order to define the general specifications of models in this class we combine elements of models of mixed membership (Pritchard et al. 2000; Erosheva 2002; Rosenberg et al. 2002; Blei et al. 2003; Xing et al. 2003a, 2004; Erosheva et al. 2004; Airoldi et al. 2005; Blei and Lafferty 2006; Xing et al. 2006) with elements of block models for networks (Wasserman 1980; Wasserman and Anderson 1987; Wasserman and Faust 1994; Wasserman and Pattison 1996; Fienberg et al. 1985; Frank and Strauss 1986; Nowicki and Snijders 2001; Hoff et al. 2002; Airoldi et al. 2006a). We combine such elements in a hierarchical Bayes framework, where (i) latent variables encode semantic elements, e.g., protein-specific latent variables encode their functions, and (ii) a specific structure among observable and latent random elements is posited.

3.1 Model Formulation

We characterize the stochastic block models of mixed-membership in terms of assumptions at four levels.[3]

A1–Population Level. Assume that there are K classes or sub-populations in the population of interest. We denote by $f(y_{jnm}|\eta_{gh})$ the probability distribution of the j-th response graph at the pair of nodes (n, m), where the n-th node is in the h-th sub-population, the m-th node is in the k-th sub-population, and η_{gh} contains the relevant parameters. The indices n, m run in $\mathcal{N}$, and the indices g, h run in $[1, K]$. Within sub-population pairs, the observed paired responses are assumed independent.

A2–Node Level. The components of the membership vector $\theta_n = (\theta_{n1}, \ldots, \theta_{nK})'$ encodes the mixed-membership of the n-th node to the various sub-populations. The distribution of the observed response y_{jnm} given the relevant, node-specific membership scores, (θ_n, θ_m), is then

$$Pr\ (y_{jnm}|\theta_n, \theta_m, \eta) = \sum_{g,h=1}^{K} \theta_{ng} f(y_{jnm}|\eta_{gh})\theta_{mh}. \tag{1}$$

Conditional on the mixed-membership scores, the response edges y_{jnm} are independent of one another, both across distinct graphs and pairs of nodes.

[3]In the remainder of this paper, we index random quantities with up to five sub-script. Typically, the first subscript refers to the response variable, $j = 1, \ldots, J$, the second subscript refers to the replicate, $r = 1, \ldots, R_j$. The next (next two) subscript refers to a node (pair of nodes), $n, m \in \mathcal{N}$. The fifth subscript refers to the number of latent clusters, $k = 1, \ldots, K$, in those few cases where a vector has to be indexed by $jrnm$; see Equation 9 for an example. At times we will need two subscripts to index pairs of latent clusters, $g, h = 1, \ldots, K$

A3–Latent Variable Level. Assume that the vectors θ_n, i.e., the mixed-membership scores of the *n-th* subject, are realizations of a latent variable with distribution D_α, parameterized by vector α. The probability of observing $_{jnm}$, given the parameters, is then

$$Pr\ (y_{jnm}|\alpha,\eta) = \int \left(\sum_{g,h=1}^{K} \theta_{ng} f(y_{jnm}|\eta_{gh})\theta_{mh} \right) D_\alpha(d\theta). \tag{2}$$

A4–Sampling Scheme Level. Assume that the R independent replications of the J distinct response graphs are independent of one another. The probability of observing the whole collection of graphs, $\{y_{jrnm}\}$, given the parameters, is then given by the following equation.

$$Pr\ (\ \{y_{jrnm}\}\ |\ \alpha,\eta\ ) = \int \left(\prod_{j=1}^{J}\prod_{r=1}^{R}\prod_{n,m=1}^{N}\sum_{g,h=1}^{K} \theta_{ng} f(y_{jrnm}|\eta_{gh})\theta_{mh} \right) D_\alpha(d\theta). \tag{3}$$

The number of replications is not necessarily the same across different response graphs, i.e., $R = R_j$. Likewise, the block model can be response specific, i.e., $\eta = \eta_j$. More variations along these lines are possible.

A graphical representation of models in this family is given in Figure 2.

Full model specifications immediately adapt to the different kinds of data, e.g., multiple data types through the choice of f, or parametric or semi-parametric specifications of the prior on the number of clusters through the choice of D_α.

3.2 Admixture of Latent Blocks

Airoldi et al. (2006a) introduced the *Admixture of Latent Blocks* model to analyze a collection of protein-protein interactions. This model is defined by the simplest set of model specifications for a stochastic block model of mixed membership, and it was used to analyze the most basic kind of relational data. Given a single undirected unipartite graph with binary edges, the Admixture of Latent Blocks model recovers membership of nodes to clusters (i.e., the mixed membership vectors $\theta_{1:N}$) and cluster-to-cluster interaction probabilities (i.e., the block model η), under the assumption that K non-observable clusters exist.

Using this model on protein-protein interaction data: sub-populations correspond to non-observable "stable protein complexes", indexed by k; nodes correspond to "proteins", indexed by n; there is only one response variable that encodes whether a pair of proteins interacts or not, so that j is omitted; there is only one replicate, since the interactions have been measured with an experimental procedure such as "Yeast Two Hybrid" under a single experimental condition. The model assumes that each interaction in the collection is either present or absent given the memberships to specific protein complexes of the pair of single proteins involved. That is, each protein participates in the various interactions as a member of possibly different protein complexes. In order

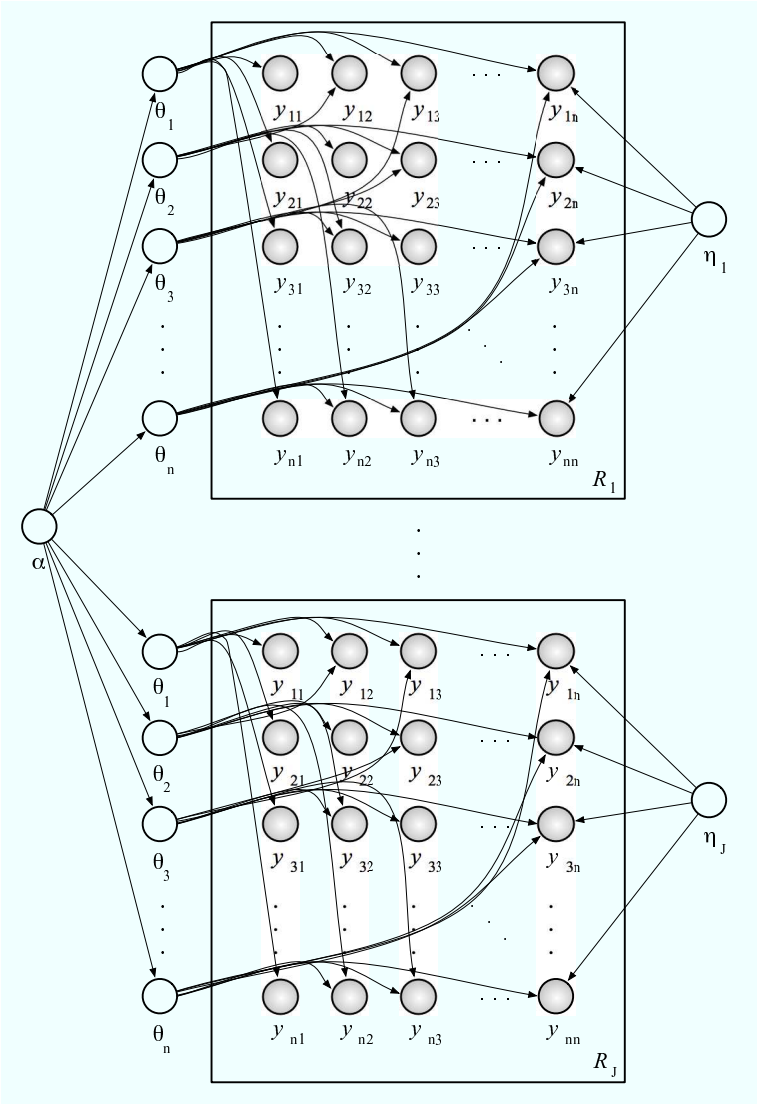

Figure 2: The graphical representation of stochastic block models of mixed membership using plates. Notes: (i) the mixed-membership vectors, $\theta_{1:N}$, are sampled once for all relations $j = 1, \ldots, J$, however, we plotted two sets of them, i.e. $j = 1$ and $j = J$, for clarity; (ii) we did not draw all the arrows out of the block models $\eta_{1:J}$, for clarity, however, all the interactions y_{jrnm} depend on the corresponding block model.

to simplify the inference, an explicit pair of indicator variables $(z^{\rightarrow}_{nm}, z^{\leftarrow}_{nm})$ is introduced for each interaction in the observed collection, which indicates the protein complexes that the two proteins are members of as they interact. The function $f(y_{nm}|\eta_{gh}) =$

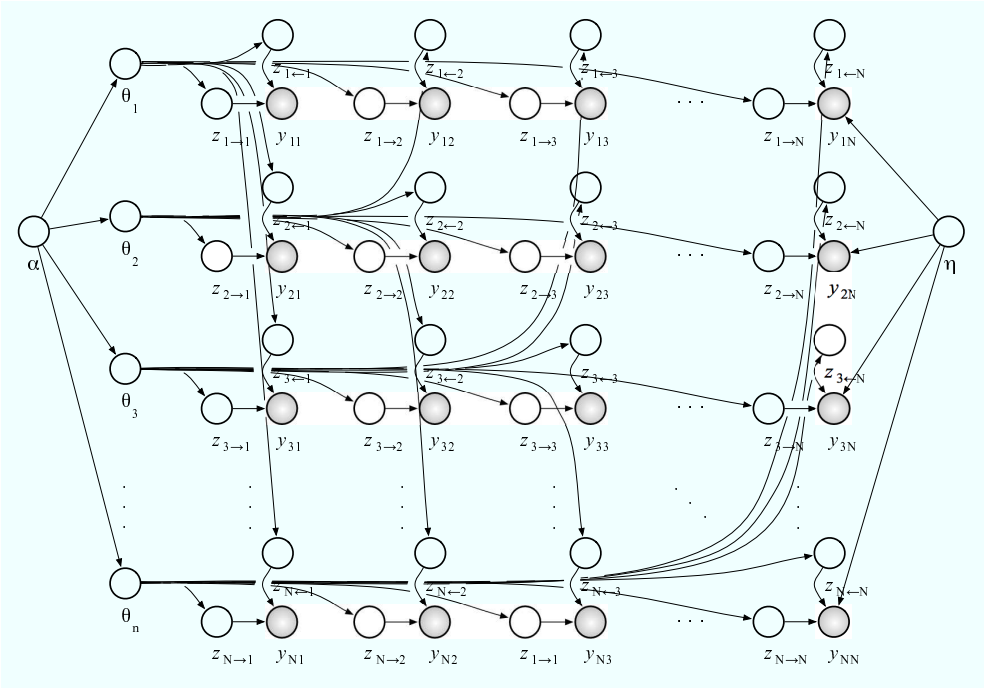

Figure 3: The graphical representation of the admixture of latent blocks introduced by Airoldi et al. (2006a) using plates. Note that we did not draw all the arrows out of the block model η, for clarity, however all the interactions y_{nm} depend on it.

$Pr\ (y_{nm} = 1 | z^{\rightarrow}_{nm} = g, z^{\leftarrow}_{nm} = h) = Bernoulli\ (\eta_{gh})$, where η_{gh} is the probability that a protein in complex g interacts with a protein in complex h. A mixed-membership vectors $\theta_{1:N}$ encode the expected protein complex proportions. They are distributed according to D_α, i.e., a Dirichlet distribution. We obtain equation 1 integrating out the protein complex indicator variables $(z^{\rightarrow}_{nm}, z^{\leftarrow}_{nm})$ at the interactions level—the latent indicators $z^{\rightarrow}_{nm}$ are distributed according to a $Multinomial\ (1, \theta_n)$, whereas the latent indicators $z^{\leftarrow}_{nm}$ are distributed according to a $Multinomial\ (1, \theta_m)$.

A graphical representation of this specific model is given in Figure 3.

4 Inference and Parameter Estimation

In order to learn the hyper-parameters, (α, η), and infer the mixed-membership vectors, $\theta_{1:N}$, we need to be able to evaluate the likelihood, which involves the non-tractable integral in Equation 3. Given the large amount of data available in the applications we are concerned with, we focus on variational methods, which present a computationally cheaper alternative to Monte Carlo Markov chain methods. Using variational methods, we find a tractable lower bound for the likelihood that can be used as a surrogate for

our inference purposes. This leads to approximate MLEs for the hyper-parameters and approximate posterior distributions for the mixed-membership vectors.

4.1 Variational Expectation-Maximization

Variational methods prescribe the use of a mean-field approximation to the posterior distribution of the latent variables given data and hyper-parameters (Jordan et al. 1999; Xing et al. 2003b). The mean-field approximation is obtained by positing a fully-factorized joint distributions over the latent variables,[4]

$$\begin{aligned} & q\left(\left\{\theta_n, z^{\rightarrow}_{jrnm}, z^{\leftarrow}_{jrnm}\right\} \mid \left\{\gamma_n, \phi^{\rightarrow}_{jrnm}, \phi^{\leftarrow}_{jrnm}\right\}\right) = \\ & = \prod_{n\in\mathcal{N}} q\left(\theta_n|\gamma_n\right) \prod_{j=1}^{J} \prod_{r=1}^{r_j} \prod_{n,m\in\mathcal{N}} \left(q\left(z^{\rightarrow}_{jrnm} \mid \phi^{\rightarrow}_{jrnm}\right) q\left(z^{\leftarrow}_{jrnm} \mid \phi^{\leftarrow}_{jrnm}\right)\right), \quad (4) \end{aligned}$$

which depends on a set of free parameters, $\{\gamma_n, \phi^{\rightarrow}_{jrnm}, \phi^{\leftarrow}_{jrnm}\}$. The mean-field approximation is then given by the following posterior distribution,

$$\tilde{p}\left(\left\{\theta_n, z^{\rightarrow}_{jrnm}, z^{\leftarrow}_{jrnm}\right\} \mid \left\{\gamma_n, \phi^{\rightarrow}_{jrnm}, \phi^{\leftarrow}_{jrnm}\right\}, \alpha, \eta_{1:J}\right) \quad (5)$$

where the conditioning on the data is now obtained indirectly, trough the *optimal values* of the free parameters,

$$\begin{aligned} \tilde{\gamma}_n &= \tilde{\gamma}_n\left(\left\{Y_{jr} : 0 \leq j \leq J, 0 \leq r \leq R_j\right\}\right), \\ \tilde{\phi}^{\rightarrow}_{jrnm} &= \tilde{\phi}^{\rightarrow}_{jrnm}\left(Y_{jr}\right), \\ \tilde{\phi}^{\leftarrow}_{jrnm} &= \tilde{\phi}^{\leftarrow}_{jrnm}\left(Y_{jr}\right). \end{aligned}$$

The fully factorized distribution q in Equation 4 leads to a lower bound for the likelihood. In fact, it is possible to find a closed form solution to the integral in Equation 3 by applying Jensen's inequality, and then integrating the latent variables out with respect q. The mean-field approximate posterior, $\tilde{p}$ in Equation 5, is obtained by substituting the lower bound for the likelihood in the calculations, as appropriate. Specifically, the mean-field approximation corresponds to the values of the free parameters, $\{\tilde{\gamma}_n, \tilde{\phi}^{\rightarrow}_{jrnm}, \tilde{\phi}^{\leftarrow}_{jrnm}\}$, that minimizes the Kullback-Leibler (KL) divergence between the true and the approximate posteriors.

The *variational EM* algorithm we develop for performing posterior inference is then an approximate EM algorithm. During the E step, we tighten the lower bound for the likelihood by minimizing the KL divergence between the true and the approximate posteriors over the free parameters, $\{\gamma_n, \phi^{\rightarrow}_{jrnm}, \phi^{\leftarrow}_{jrnm}\}$, given the most recent estimates for the hyper-parameters. During the M step, we maximize the lower bound for the likelihood over the hyper-parameters of the model, $(\alpha, \eta_{1:J})$, to obtain to (approximate) maximum likelihood estimates (Carlin and Louis 2005).

[4]Note that the set of latent variables $\{z^{\rightarrow}_{nm}, z^{\leftarrow}_{nm} : n, m \in \mathcal{N}\}$ are introduced to simplify the variational inference, e.g., without such parameters the γ_{ng} updates in Equation 8, needed to carry out the approximate E step of a variational EM algorithm, would not be in closed form.

In the approximate E step we update the free parameters for the mean-field approximation of the posterior distribution of the latent variables, $\{\gamma_n, \phi^{\rightarrow}_{jrnm}, \phi^{\leftarrow}_{jrnm}\}$, given the most recent estimates of the hyper-parameters of the model, $(\alpha, \eta_{1:J})$, according to Equations 6, 7 and 8,

$$\phi^{\rightarrow}_{nmg} \propto \exp\left\{ \psi(\gamma_{ng}) - \psi(\sum_{g=1}^{K} \gamma_{ng}) \right\} \cdot \prod_{h=1}^{K} f\left(y_{nm} \mid \eta_{gh} \right)^{\phi^{\leftarrow}_{nmh}} \tag{6}$$

$$\phi^{\leftarrow}_{nmh} \propto \exp\left\{ \psi(\gamma_{mh}) - \psi(\sum_{h=1}^{K} \gamma_{mh}) \right\} \cdot \prod_{g=1}^{K} f\left(y_{nm} \mid \eta_{gh} \right)^{\phi^{\rightarrow}_{nmg}} \tag{7}$$

$$\gamma_{ng} = \alpha_g + \sum_{j=1}^{J} \sum_{r=1}^{R_j} \left(\sum_{m=1}^{N} \phi^{\rightarrow}_{jrnmg} + \sum_{m=1}^{N} \phi^{\leftarrow}_{jrnmg} \right). \tag{8}$$

This minimizes the posterior KL divergence between true and approximate posteriors, at the graph level, and leads to a new lower bound for the likelihood of the collection of graphs. Note that the free parameter updates above use observations in a single graph only, hence we suppressed the indices j, r. Further, they are general in that they do not depend on a specific observation model. However, our derivations do assume a fully factorized variational distribution, i.e., the mean field approximation.

In the M step, we update the hyper-parameters of the model, $(\alpha, \eta_{1:J})$, by maximizing the tight lower bound for the likelihood over such hyper-parameters, given the most recent updates of the free parameters the bound depends on, $\{\gamma_n, \phi^{\rightarrow}_{jrnm}, \phi^{\leftarrow}_{jrnm}\}$. In the case of $f(y_{jrnm}|\eta_{gh}) = Bernoulli\ (\eta_{gh})$, for example, this argument leads to the following (approximate) maximum likelihood estimates for the parameters:

$$\eta^{*}_{jgh} = \frac{1}{R_j} \sum_{r=1}^{R_j} \frac{\sum_{n,m=1}^{N} \phi^{\rightarrow}_{jrnmg}\, \phi^{\leftarrow}_{jrnmh}\, y_{jrnm}}{\sum_{n,m=1}^{N} \phi^{\rightarrow}_{jrnmg}\, \phi^{\leftarrow}_{jrnmh}}. \tag{9}$$

It is not possible to derive closed form expression for the approximate maximum likelihood estimates of the parameters underlying $f(y_{jrnm}|\eta_{gh})$, in general, although closed form expressions exist in many cases, Bernoulli, Poisson and Gaussian among them. Further, a closed form solution for the approximate maximum likelihood estimates of α does not exist (Minka and Lafferty 2002; Blei et al. 2003). We can produce a method that is linear in time by using Newton-Raphson, with the gradient and Hessian for the log-likelihood in Equations 10 and 11,

$$\frac{\partial L}{\partial \alpha_{[k]}} = N \left(\Psi\left(\sum_{k=1}^{K} \alpha_k \right) - \Psi(\alpha_k) \right) + \sum_{n=1}^{N} \left(\Psi(\gamma_{nk}) - \Psi\left(\sum_{k=1}^{K} \gamma_{nk} \right) \right), \tag{10}$$

$$\frac{\partial L}{\partial \alpha_{k_1} \alpha_{k_2}} = N \left(\delta_{k_1 = k_2} \cdot \Psi'(\alpha_{k_1}) - \Psi'\left(\sum_{k_2=1}^{K} \alpha_{k_2} \right) \right). \tag{11}$$

4.2 Nested versus Naïve Variational Algorithms

The variational inference algorithm presented in Figures 4 and 4 is what we term a *nested* variational inference algorithm. The difference from the naïve version of the algorithm at a glance is that to carry out the latter, we initialize the variational Dirichlet parameters γ_n and the variational Multinomial parameters $\boldsymbol{\phi}_{ij}$ to non-informative values, then we iterate until convergence the following two steps: (i) update $\phi^{\rightarrow}_{nm}$ and $\phi^{\leftarrow}_{nm}$ for all edges (n, m), and (ii) update γ_n for all nodes n. In such algorithm, at each variational inference cycle we need to allocate $NK + 2N^2K$ scalars. In our simulation experiments, where the true block model is known, the naïve variational algorithm often converged to a bad solution, or converged after a large number of iterations. We attribute this behavior to a dependence that our two main assumptions (block model and mixed membership) induce between $\{\gamma_{ng}\}$ and $\{\eta_{jgh}\}$, which is not satisfied by the naïve algorithm. Some intuition about why this may happen follows. From a purely algorithmic perspective, the naïve variational EM algorithm instantiates a large coordinate ascent algorithm, where the parameters can be semantically divided into coherent blocks. Blocks are processed in a specific order, and the parameters within each block get all updated each time[5]. At every iteration the naïve algorithm sets all the elements of $\{\gamma_{ng}\}$ equal to the same constant. This dampens the likelihood by suddenly breaking the dependence between the estimates of parameters in $\{\gamma_{ng}\}$ and in $\{\eta_{jgh}\}$ that was being inferred from the data.

Instead, the nested variational inference algorithm maintains some of this dependence that is being inferred from the data across the various iterations. This is achieved mainly through a different scheduling of the parameter updates in the various blocks. To a minor extent, the dependence is maintained by always keeping the block of free parameters, $\{\phi^{\rightarrow}_{nm}, \phi^{\leftarrow}_{nm}\}$, optimized given the other variational parameters. Note that these parameters are involved in the updates of parameters in $\{\gamma_{ng}\}$ and in $\{\eta_{jgh}\}$, thus providing us with a channel to maintain some of the dependence among them, i.e., by keeping them at their optimal value given the data. Further, the nested algorithm has the advantage that it trades time for space thus allowing us to deal with large graphs; at each variational cycle we need to allocate $NK + 2K$ scalars. The increased running time is partially offset by the fact that the algorithm can be parallelized and leads to empirically observed faster convergence rates. This algorithm is also better than blocked and collapsed Gibbs sampler in terms of memory requirements.

5 Experiments and Examples

In this section we explore the behavior of the admixture of latent blocks model described in Section 3.2, which is a simple stochastic block model of mixed membership.

We present experimental evidence to show that: (i) our model recovers both the mixed membership of nodes to clusters, and the latent block structure among clusters;

[5]Within a block, the order according to which (scalar) parameters get updated is not expected to affect convergence.

Outer loop

1. initialize $\boldsymbol{\gamma}_{ng}^0 = \frac{2N}{K}$ for all n, g
2. **repeat**
3. **for** $n \in \mathcal{N}$
4. **for** $m \in \mathcal{N}$
5. get **variational** $\phi_{nm}^{\rightarrow\ t+1}$ and $\phi_{nm}^{\leftarrow\ t+1} = g(y_{nm}, \gamma_n^t, \gamma_m^t, \eta^t)$
6. partially update γ_n^{t+1}, γ_m^{t+1} and η^{t+1}
7. **until** convergence

Figure 4: The nested (two-layered) variational inference algorithm for γ and $(\phi^{\rightarrow}, \phi^{\leftarrow})$. The inner layer consists of Step 5. The function g is described in details in Figure 5.

Inner loop

1. initialize $\phi_{nmg}^{\rightarrow\ 0} = \phi_{nmh}^{\leftarrow\ 0} = \frac{1}{K}$ for all g, h
2. **repeat**
3. **for** $g = 1$ to K
4. update $\phi_{nmg}^{\rightarrow\ s+1} \propto g_1(\phi_{nm}^{\leftarrow\ s}, \gamma, \eta)$
5. normalize $\phi_{nm}^{\rightarrow\ s+1}$ to sum to 1
6. **for** $h = 1$ to K
7. update $\phi_{nmh}^{\leftarrow\ s+1} \propto g_2(\phi_{nm}^{\rightarrow\ s}, \gamma, \eta)$
8. normalize $\phi_{nm}^{\leftarrow\ s+1}$ to sum to 1
9. **until** convergence

Figure 5: Details Step 5. in Figure 4; the inference algorithm for the variational parameters $(\phi_{nm}^{\rightarrow}, \phi_{nm}^{\leftarrow})$ corresponding to the basic observation y_{nm}. The functions g_1 and g_2 are updates for $\phi_{nmg}^{\rightarrow}$ and $\phi_{nmh}^{\leftarrow}$ described in the text of Section 4.

(ii) the nested variational algorithm drives the log-likelihood to converge faster to its peak than the naïve algorithm; (iii) a cross-validation experiment is sufficient to perform model selection for the parametric formulation of the admixture of latent blocks model, where the number of clusters K is fixed prior to the analysis. We then present some of the analysis and results in Airoldi et al. (2006a) about a set of protein-protein interactions in Yeast, to motivate the discussion of some technical issues related to the class of stochastic block model of mixed membership.

5.1 Inference and Parameter Estimation

Using the admixture of latent blocks model of Section 3.2, we simulated example graphs of $100, 300$, and 600 nodes from block models with $4, 10$, and 20 clusters, respectively. We used values of $\alpha \in \{0.05, 0.1, 0.25\}$ to simulate a range of settings in terms of membership of nodes to clusters—from unique to mixed.

The variational EM algorithm described in Section 4 successfully recovers both the latent block model, η, and the latent mixed membership vectors $(\theta_{1:N})$. In Figure 6 we show the adjacency matrices of binary interactions where rows (corresponding

to nodes) are reordered according to their most likely membership. The estimated reordering reveals the block model that was originally used to simulate the interactions. As α increases (left to right) each node is allowed to belong to more clusters; as a consequence it may express interaction patterns of clusters other than the one it is most likely to belong to, as we simulate the single interactions with other nodes. This phenomenon reflects in the reordered interaction matrices as the block structure is less evident.

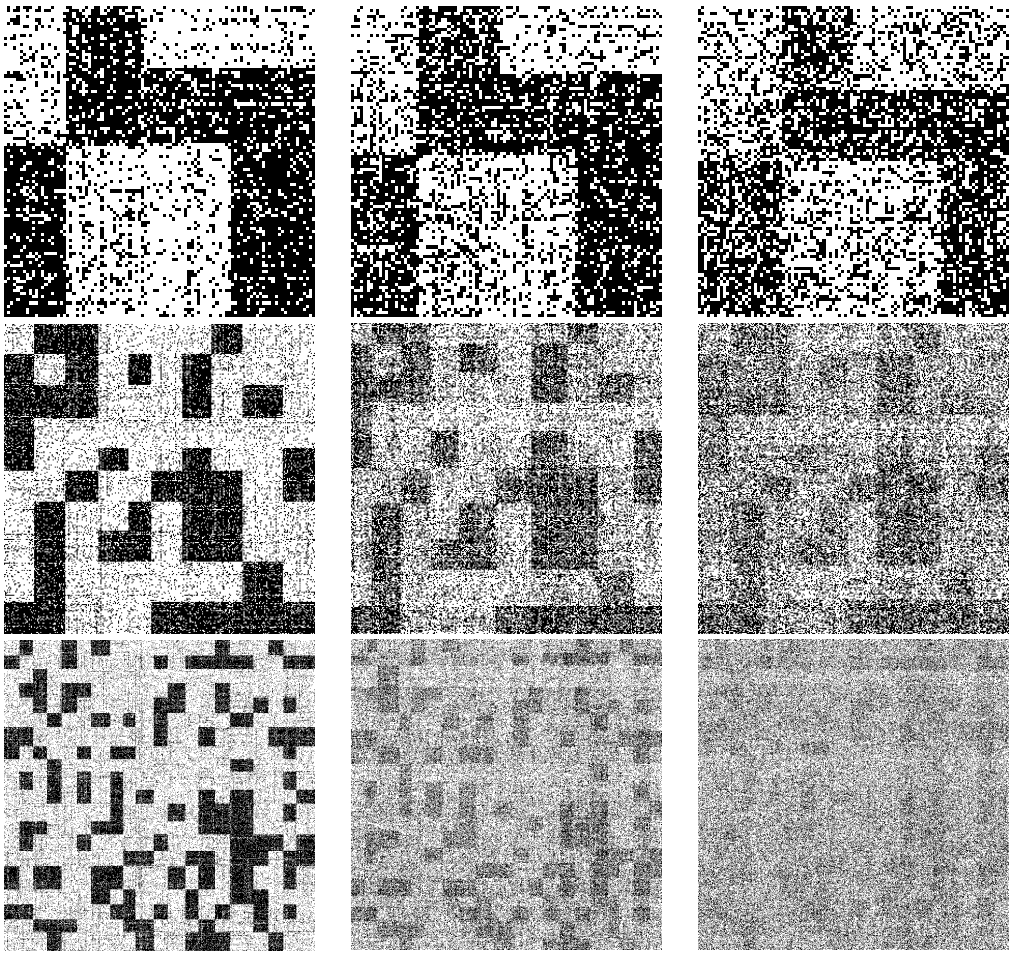

Figure 6: Adjacency matrices of corresponding to simulated interaction graphs with 100 nodes and 4 clusters, 300 nodes and 10 clusters, 600 nodes and 20 clusters (top to bottom) and α equal to $0.05, 0.1$ and 0.25 (left to right). Rows, which corresponds to nodes, are reordered according to their most likely membership. The estimated reordering reveals the block model that was originally used to simulate the interactions.

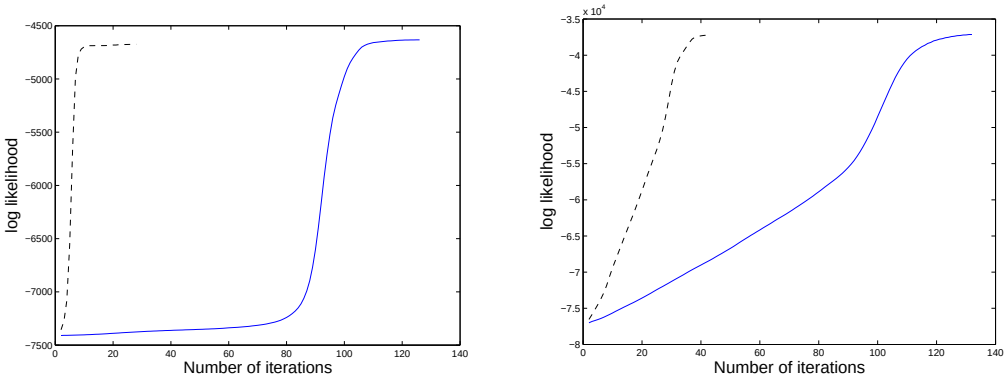

Figure 7: The running time of the naïve variational inference (solid, blue line) against the running time of our enhanced (nested) variational inference algorithm (dashed, black line), in two experimental settings: 100 nodes with 4 clusters, and 300 nodes with 10 clusters. We measure the number of iterations on the X axis and the log-likelihood on the Y axis. The two curves (iterations/log-likelihood) in each panel correspond to the same initial values for the parameters. Both algorithms reach the same plateau in terms of log-likelihood, and converge to the same parameter estimates.

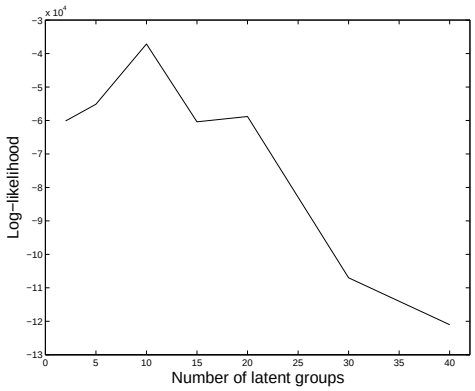

Figure 8: The held-out log-likelihood is indicative of the true number of latent clusters, on simulated data. We measure the number of latent clusters on the X axis and the log-likelihood on a test set on the Y axis. In the example shown the peak corresponds to the correct number of clusters, i.e., $K^* = 10$.

In Figure 7 we compare the running times of the nested variational-EM algorithm versus the naïve implementation. The nested algorithm, which is more efficient in terms of space, converged faster in our simulations. Further, note that the nested variational algorithm can be parallelized given that the updates for each interaction (i, j) are independent of one another.

Figure 8 shows an example where cross-validation is sufficient to perform model selection for the parametric formulation of the admixture of latent blocks model. In the parametric formulation of the model, the number of clusters K is fixed prior to the analysis. Cross-validation suggests the value of K that maximizes the likelihood on a test set. In Figure 8 cross-validation suggests a latent number of clusters K^* equals to 10. For a through exploration of model selection issues in hierarchical Bayesian models of mixed membership we refer to Airoldi et al. (2006b).

5.2 Protein-Protein Interactions

Here we present elements of the analysis of a set of protein-protein interactions in Yeast (Airoldi et al. 2006a) with the goal of discussing technical issues related to the class of stochastic block model of mixed membership.

The data consists of a collection of interactions among a subset of proteins in Yeast, such as those obtained with a yeast-two-hybrid experiment, which were experimentally verified by researchers at the Munich institute for protein sequencing (MIPS), along with a set of hand-curated functional annotations for the same subset of proteins (Mewes et al. 2004). The scientific investigation that provides the background theme for the analysis aims at finding out whether a collection of protein-protein interactions alone contains information about the functionality of the proteins involved.

Functional annotations of proteins in Yeast are organized in a tree. A possible strategy to find out whether functional annotations correlate to latent aspects underlying protein interactions is as follows. We cut the annotations tree at an arbitrary level, e.g., the first level return the 15 functional categories in Table 1. This leads to

#	Category	Count	#	Category	Count
1	Metabolism	125	9	Interaction w/ cell. environment	18
2	Energy	56	10	Cellular regulation	37
3	Cell cycle & DNA processing	162	11	Cellular other	78
4	Transcription (tRNA)	258	12	Control of cell organization	36
5	Protein synthesis	220	13	Sub-cellular activities	789
6	Protein fate	170	14	Protein regulators	1
7	Cellular transportation	122	15	Transport facilitation	41
8	Cell rescue, defence & virulence	6			

Table 1: In the table we report the 15 high-level functional categories we obtain by cutting the annotation tree at the first level. We also report how many proteins, among the 871 we considered in the MIPS collection, participate in each of such categories. Most proteins participate in more than one function, with an average of ≈ 2.4 functional annotations for each protein.

a 15-dimensional representation of each protein, using MIPS hand-curated functional annotations—see Figure 9. The admixture of latent blocks model described in Section 3.2 on the other hand, as we set $K = 15$, provides us with estimates of 15-dimensional mixed membership vectors. If the interactions contain information about the annota-

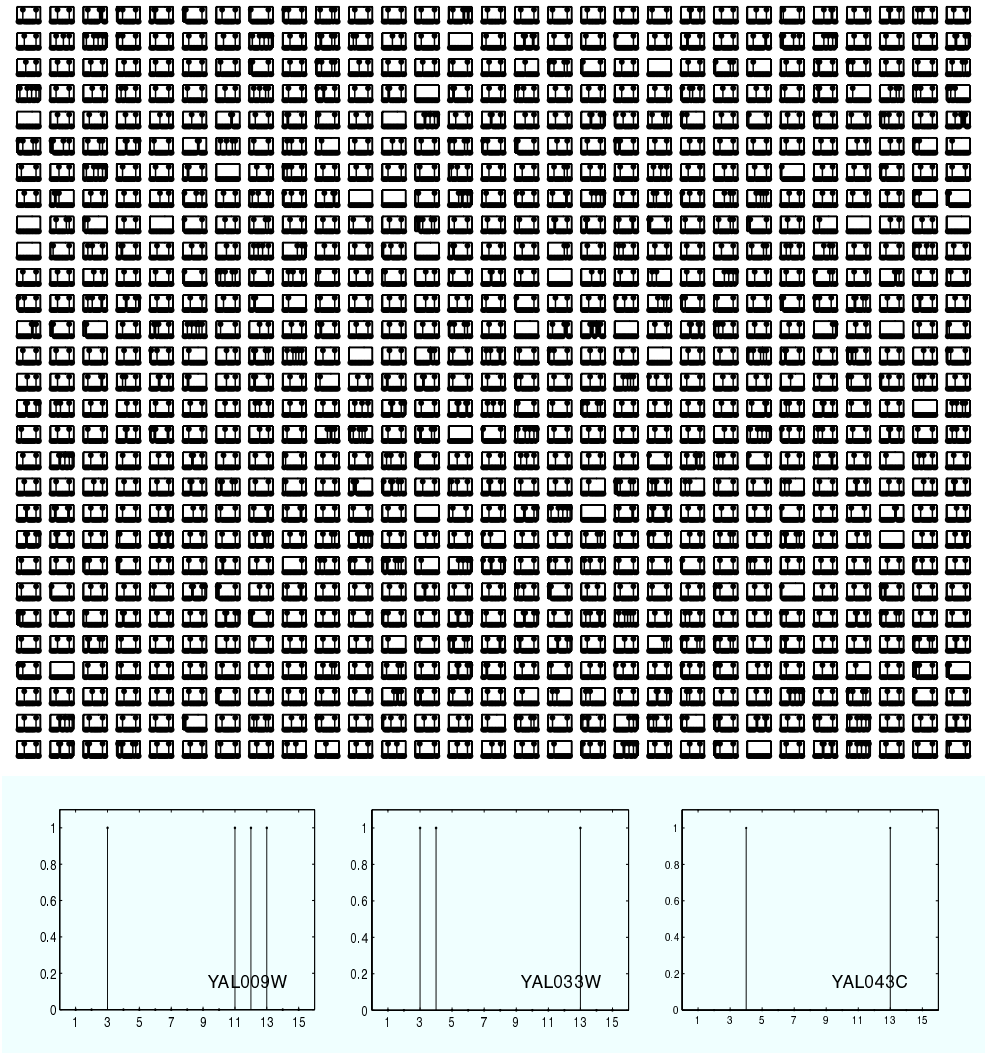

Figure 9: By cutting the annotations tree at the first level we find the 15 functional categories in Table 1. Here we plot the 15-dimensional representations of each protein, using the MIPS hand-curated functional annotations. Each panel corresponds to a protein; the 15 functional categories are displayed on the X axis, whereas the presence or absence of the corresponding functional annotation is displayed on the Y axis. The plots at the bottom zoom into the panels corresponding to three example proteins.

tions, it is conceivable that we can interpret the latent aspects the model estimates in terms of functions, and that the 15-dimensional representation of proteins in terms of MIPS functions will correlate with a thresholded version of the 15-dimensional mem-

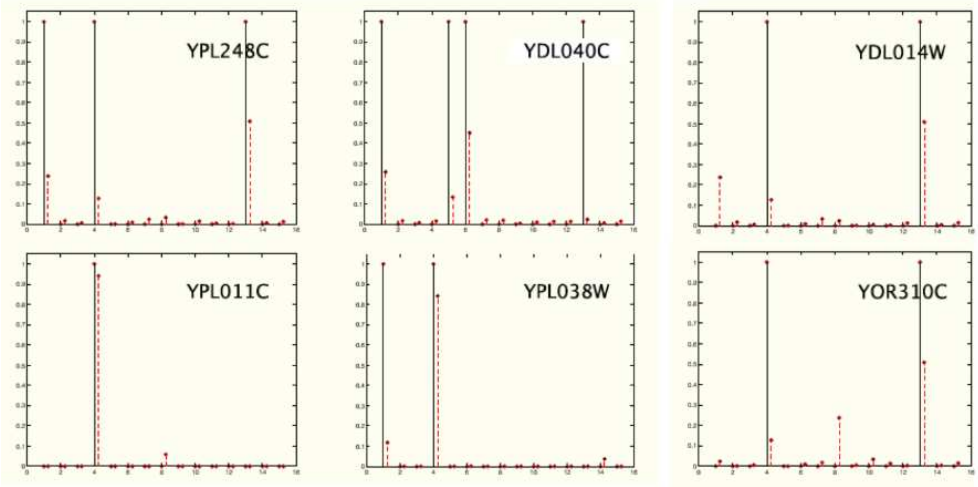

Figure 10: True annotations (solid, black bars) versus estimated mixed membership vectors (dashed, red bars) for six example proteins. Note that these panels implicitly assume a mapping between latent clusters and distinct functional categories. Section 5.3 discusses strategies to estimate one.

bership vectors. Indeed, the protein-protein interactions contain information about abundant annotations to a large degree (Airoldi et al. 2006a, see Figure 10).

A thorough and more extensive evaluation is ongoing, which integrates interaction data from different studies (Gavin et al. 2002; Ho et al. 2002; Mewes et al. 2004; Krogan et al. 2006).

5.3 Discussion

From the simulations and the case study two main points for discussion emerge, which apply to stochastic block models of mixed membership in general.

The first point concerns the lack of identifiability of latent clusters. Fitting the model with no information about node-to-cluster memberships leads to non-identifiable clusters. In the biological case study, for example, the estimates of the mixed-membership vectors, $\hat{\theta}_{1:N}$, do not carry information about which of the 15 dimensions corresponds to which functional category. However, it is possible to resolve the mapping of latent clusters to distinct functions by means of known functional annotations for a small set of proteins. Assuming that the known functional annotations are given for a random set of proteins, $\mathcal{P}$, we may look for the mapping that minimizes some measure of distance between the marginal membership distribution of proteins in $\mathcal{P}$ and the estimated marginal membership distributions of all proteins, derived from $\hat{\theta}_{1:N}$. Such two distributions are shown in Figure 11, for the biological case study. Alternatively, it is possible to make use of known functional annotations of proteins in $\mathcal{P}$ by calibrating informa-

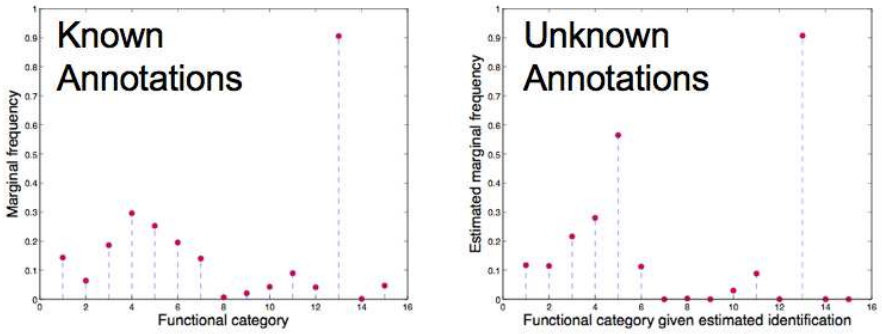

Figure 11: The observed marginal distribution of membership counts, of proteins to functional categories, in a small random sample of about 90 proteins (left panel). Note that the marginal distribution of membership counts in the sample roughly follows the true marginal distribution implied by Table 1. The estimated marginal distribution of membership counts of all proteins, derived from $\hat{\theta}_{1:N}$ (right panel).

tive priors for the corresponding mixed-membership vectors, $\{\theta_n : n \in \mathcal{P}\}$ and for the relevant elements of the block model, η.[6]

The second point concerns the contribution of the mixed membership assumption to the analysis. Figure 9 shows that mixed membership of nodes to clusters is a feature of Yeast's proteins in the MIPS collection. It is then reasonable to ask "how much" this assumption is really contributing to the model fit, at maximum likelihood. Figure 12 offers some evidence that suggests mixed membership as a reasonable assumption, which contributes to model fit. The figure shows a histogram of the estimated mixed membership vectors, $\hat{\theta}_{1:N}$. Apart from the overwhelming amount of negligible estimates,[7] among the positive estimates we find that 9.2% are smaller than 0.1, 45.5% are bigger than 0.8 and 45.3% are in between. The large percentage of estimated memberships in between 0.1 and 0.8 (i.e., the gaps in the histogram in Figure 9) portrays mixed membership as a reasonable assumption, which contributes to model fit. We recognize, however, that there may be alternative explanations for this phenomenon. For example, relaxing the constraint of single membership to allow for mixed membership of nodes to clusters introduces an extra set of random elements in the model, whose estimates can be distorted by the lack of fit due, e.g., to the fact that the model is bad in some sense. In such a case, estimated memberships "in between" would be capturing systematic error and noise. Hence a more detailed analysis is needed to support a conclusive argument regarding the contributions of the mixed membership assumption to the analysis.

[6]Further, it is possible to use the model for testing hypotheses about single elements, or specific structures, of the stochastic stochastic block model, but we do not discuss this type of analysis here.

[7]We find that $\hat{\theta}_{nk} < 0.214$ for 90% of the estimated memberships, suggesting absence of the *k-th* functional annotation for the *n-th* protein.

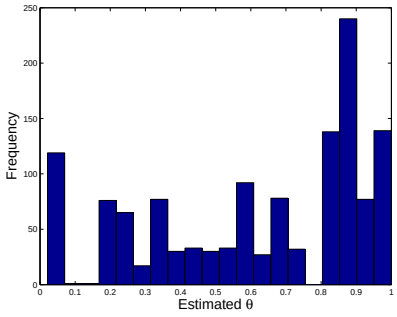

Figure 12: A histogram of the top 10% estimated memberships, $\{\hat{\theta}_{nk} > 0.214\}$.

6 Conclusions

In this paper we introduced stochastic block models of mixed membership; a novel class of latent variable models for relational data. These models provide support for scientific analyses of interest in applications where the observations can be represented as a collection of unipartite graphs; we discuss this in Section 2. Importantly, it is the data and the goals of the analyses that motivate our technical choices, e.g., latent variables in our models are introduced to represent domain-specific elements of interest. For example, the biological case study discussed in Section 5.2, which reports on the analysis of a collection of protein-protein interactions by Airoldi et al. (2006a), suggests that latent aspects, which were introduced to represent stable protein complexes, correlate with functional processes in the cell.

The applications we consider share considerable similarities in the way domain-specific semantic concepts (e.g., protein to stable protein complexes, and individuals to social groups) relate. This allows us to state a general formulation of the problem, which is no longer specific to an application domain, in Section 2.2. We subsume full model specifications aimed at analyzing diverse kinds of data, under both parametric and nonparametric assumptions on the number of non-observable clusters, into a general formulation amenable to theoretical analysis, in Section 3. Working within specifications, we develop variational methods for carrying out approximate posterior inference in these models, in Section 4.

To conclude, we note that the nested variational inference algorithm we developed is parallelizable and allows for fast approximate inference on large graphs. However, there is considerable opportunity to improve both upon computation, and efficiency of approximation.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2006a). "Mixed membership stochastic block models for relational data with application to protein-protein interactions." Manuscript under review.

Airoldi, E. M., Blei, D. M., Xing, E. P., and Fienberg, S. E. (2005). "A latent mixed-membership model for relational data." In *Workshop on Link Discovery: Issues, Approaches and Applications*. In conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Airoldi, E. M., Fienberg, S. E., Joutard, C., and Love, T. M. (2006b). "Discovery of Latent Patterns with Hierarchical Bayesian Mixed-Membership Models and the Issue of Model Choice." Technical Report CMU-MLD-06-101, School of Computer Science, Carnegie Mellon University.

Anderson, C. J., Wasserman, S., and Faust, K. (1992). "Building Stochastic Blockmodels." *Social Networks*, 14: 137–161.

Bhardwaj, N. and Lu, H. (2005). "Correlation between gene expression profiles and protein–protein interactions within and across genomes." *Bioinformatics*, 21(11): 2730–2738.

Blei, D. M., Jordan, M. I., and Ng, A. Y. (2003). "Hierarchical Bayesian Models for Applications in Information Retrieval." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics*, volume 7, 25–44. Oxford University Press.

Blei, D. M. and Lafferty, J. D. (2006). "Dynamic Topic Models." In *International Conference on Machine Learning*, volume 23, 113–120.

Carlin, B. P. and Louis, T. A. (2005). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall.

Erosheva, E. A. (2002). "Grade of membership and latent structure models with application to disability survey data." Ph.D. thesis, Carnegie Mellon University, Department of Statistics.

— (2003). "Bayesian Estimation of the Grade of Membership Model." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics*, volume 7, 501–510. Oxford University Press.

Erosheva, E. A. and Fienberg, S. E. (2005). "Bayesian Mixed Membership Models for Soft Clustering and Classification." In Weihs, C. and Gaul, W. (eds.), *Classification—The Ubiquitous Challenge*, 11–26. Springer-Verlag.

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. (2004). "Mixed-membership models of scientific publications." *Proceedings of the National Academy of Sciences*, 97(22): 11885–11892.

Fienberg, S. E., Meyer, M. M., and Wasserman, S. (1985). "Statistical Analysis of Multiple Sociometric Relations." *Journal of the American Statistical Association*, 80: 51–67.

Frank, O. and Strauss, D. (1986). "Markov Graphs." *Journal of the American Statistical Association*, 81: 832–842.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., and et. al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." *Nature*, 415: 141–147.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., and et. al, K. B. (2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." *Nature*, 415: 180–183.

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: 1090–1098.

Holland, P., Laskey, K. B., and Leinhardt, S. (1983). "Stochastic Blockmodels: Some First Steps." *Social Networks*, 5: 109–137.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). "Introduction to Variational Methods for Graphical Models." *Machine Learning*, 37: 183–233.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., PeregrÃ-n Alvarez, J., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." *Nature*, 440(7084): 637–643.

Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., and et. al (2004). "MIPS: analysis and annotation of proteins from whole genomes." *Nucleic Acids Research*, 32: D41–44.

Minka, T. and Lafferty, J. (2002). "Expectation-propagation for the generative aspect model." In *Uncertainty in Artificial Intelligence*.

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

Priebe, C. E., Conroy, J. M., Marchette, D. J., and Park, Y. (2005). "Scan Statistics on Enron Graphs." *Computational and Mathematical Organization Theory*, 11(3): 229–247.

Pritchard, J., Stephens, M., and Donnelly, P. (2000). "Inference of population structure using multilocus genotype data." *Genetics*, 155: 945–959.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). "Genetic Structure of Human Populations." *Science*, 298: 2381–2385.

Sampson, F. (1968). "A Novitiate in a period of change: An experimental and case study of social relationships." Ph.D. thesis, Cornell University.

Wasserman, S. (1980). "Analyzing social networks as stochastic processes." *Journal of the American Statistical Association*, 75: 280—294.

Wasserman, S. and Anderson, C. J. (1987). "Stochastic a Posteriori Blockmodels: Construction and Assessment." *Social Networks*, 9: 1–36.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wasserman, S. and Pattison, P. (1996). "Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p^*." *Psychometrika*, 61: 401–425.

Xing, E. P., Jordan, M. I., Karp, R. M., and Russell, S. (2003a). "A Hierarchical Bayesian Markovian Model for Motifs in Biopolymer Sequences." In *Advances in Neural Information Processing Systems*, volume 16.

Xing, E. P., Jordan, M. I., and Russell, S. (2003b). "A Generalized Mean Field Algorithm For Variational Inference In Exponential Families." In *Uncertainty in Artificial Intelligence*, volume 19.

Xing, E. P., Sohn, K., Jordan, M. I., and Teh, Y. W. (2006). "Bayesian Multi-Population Haplotype Inference via a Hierarchical Dirichlet Process Mixture." In *International Conference on Machine Learning*, volume 23, 1049–1056.

Xing, E. P., Wu, W., Jordan, M. I., and Karp, R. M. (2004). "LOGOS: A modular Bayesian model for de novo motif detection." *Journal of Bioinformatics and Computational biology*, 2(1): 127–154.

Acknowledgments

This work was partially supported by National Institutes of Health under Grant No. R01 AG023141-01, by the Office of Naval Research under Contract No. N00014-02-1-0973, by the National Science Foundation under Grants No. DMS-0240019 and DBI-0546594, and by the Department of Defense under Grant No. IIS0218466, all to Carnegie Mellon University.