Stochastic Block Models of Mixed Membership

Edoardo M. Airoldi [1] EDO@CMU.EDU
David M. Blei [2] BLEI@CS.PRINCETON.EDU
Stephen E. Fienberg [1,3] FIENBERG@STAT.CMU.EDU
Eric P. Xing [1] EPXING@CS.CMU.EDU
[1] School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA
[2] Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540 USA
[3] Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

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 many applications, clustering the objects of study or situating them in a low dimensional space (e.g., a simplex) is only one of the goals of the analysis. Begin able to estimate relational structures among the clusters themselves is often times as important. 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 & Fienberg, 2005) and block models for relational data (Holland et al., 1983) in a hierarchical Bayesian framework. We develop a novel "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 synthetic data and biological case study.

1. Introduction

In many applications, clustering the objects of study or situating them in a low dimensional space (e.g., a simplex) is only one of the goals of the analysis. Begin able to estimate relational structures among the clusters themselves is often times as important. 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 & Fienberg, 2005) and block models for relational data (Holland et al., 1983; Anderson et al., 1992; Nowicki & Snijders, 2001) in a hierarchical Bayesian framework.

In this paper we make the following contributions: (a) we introduce a general formulation of stochastic block models of mixed membership, which is amenable to theoretical analysis; and (b) we develop a novel "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

2.1. The Data

The data we are interested in modeling is a collection of directed, unipartite graphs. A unipartite graph, is

Preliminary work under review. Please, do not distribute without the authors' permission. Contact: edo@cmu.edu.

a graph whose constituent nodes are of a single type, e.g., proteins; as opposed to bipartite and multipartite graphs, whose constituent nodes are of a two and multiple types, respectively.

Four examples follow. **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. This information is representable as a single graph where the random variables associated with the edges are binary. **2.** Consider the output of a battery of yeast-two-hybrid experiments, on the same set of proteins, $\mathcal{N}$, and different, R, experimental conditions (**?**). Without entering in the biological details[1], a typical output consists of probabilities of interactions between pairs of proteins, on the logarithmic scale with base two. This information is representable as a collection of R graphs, one for each run, where the random variables associated with the edges are real-valued. **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). This information is representable as a collection of graphs with non-negative, integer edges, i.e., the number of emails. Equivalently, this information is representable as the collection of adjacency matrices corresponding to such graphs. **4.** Consider a collection of sociometric relations among a group of monks (Sampson, 1968). We observe responses to multiple, J, questions about social relations between pairs of monks, e.g., "Do you like X?" or "Do you trust X?", and multiple replicates for each response, R_j, where $j = 1, \ldots, J$. These relations are typically asymmetric. This information is representable as a collection of J collections of $R_{1:J}$ graphs, where the edges encode, say, binary responses. Alternatively, this information is representable as a collection of $\sum_j R_j$ graphs where the random variables associated with the edges have support $\{0, 1\}^J$.

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.

2.2. 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 f 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 (**?**). 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.

Typically, we perform unsupervised learning experiments, or semi-supervised learning experiments with minimal information available in terms of membership of proteins to known protein complexes. Working in the hierarchical Bayes framework, we choose to estimate the constants underlying the distribution of random quantities at the top level of the hierarchy (i.e., the hyper-parameters) via empirical Bayes. It is possible for researchers to fix the probabilities according to which clusters interact (elements of the stochastic block model) to test hypotheses in various ways. Alternative strategies are possible, e.g., supervised learning, but we do not discuss them here.

2.3. Notation

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), $n, m \in \mathcal{N}$. In particular, when we work with quantities over pairs of nodes it is sometimes easier to outline an algorithm by substituting the pair of subscripts $(n \rightarrow m, n \leftarrow m)$ for the original ones (nm, mn). 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, $g, h = 1, \ldots, K$

[1] Such details are very important as we may want to perform the analysis on data with different degrees of preprocessing.

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

3. General Model Formulation

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

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

$$Pr\ (y_{jnm}|\theta_n,\theta_m,\eta) = \sum_{g,h=1}^{K} \theta_{ng} f(y_{jnm}|\eta_{gh}) \theta_{mh}. \tag{2}$$

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

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

$$\begin{aligned} &Pr\ (y_{jnm}|\alpha,\eta) = \\ &= \int \left(\sum_{g,h=1}^{K} \theta_{ng} f(y_{jnm}|\eta_{gh}) \theta_{mh} \right) D_\alpha(d\theta). \end{aligned} \tag{3}$$

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 Equation 1. 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 1, left panel.

3.1. Example: Admixture of Latent Blocks

A vanilla data mining model that is encompassed by our general formulation is the "Admixture of Latent Blocks" introduced by Airoldi et al. (2006) for the analysis of a collection of protein-protein interactions.

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 to simplify the inference, an explicit pair of indicator variables $(z_{\overrightarrow{nm}}, z_{\overleftarrow{nm}})$ 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 $f(y_{nm}|\eta_{gh}) = Pr\ (y_{nm} = 1|z_{\overrightarrow{nm}} = g, z_{\overleftarrow{nm}} = h)$ $= Bernoulli\ (\eta_{gh})$, 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 2 integrating out the protein complex indicator variables $(z_{\overrightarrow{nm}}, z_{\overleftarrow{nm}})$ at the interactions level—the latent indicators $z_{\overrightarrow{nm}}$ are distributed according to a $Multinomial\ (1,\theta_n)$, whereas the latent indicators $z_{\overleftarrow{nm}}$ are distributed according to a $Multinomial\ (1,\theta_m)$.

A graphical representation of this specific model is given in Figure 1, right panel.

4. Nested Variational Inference

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

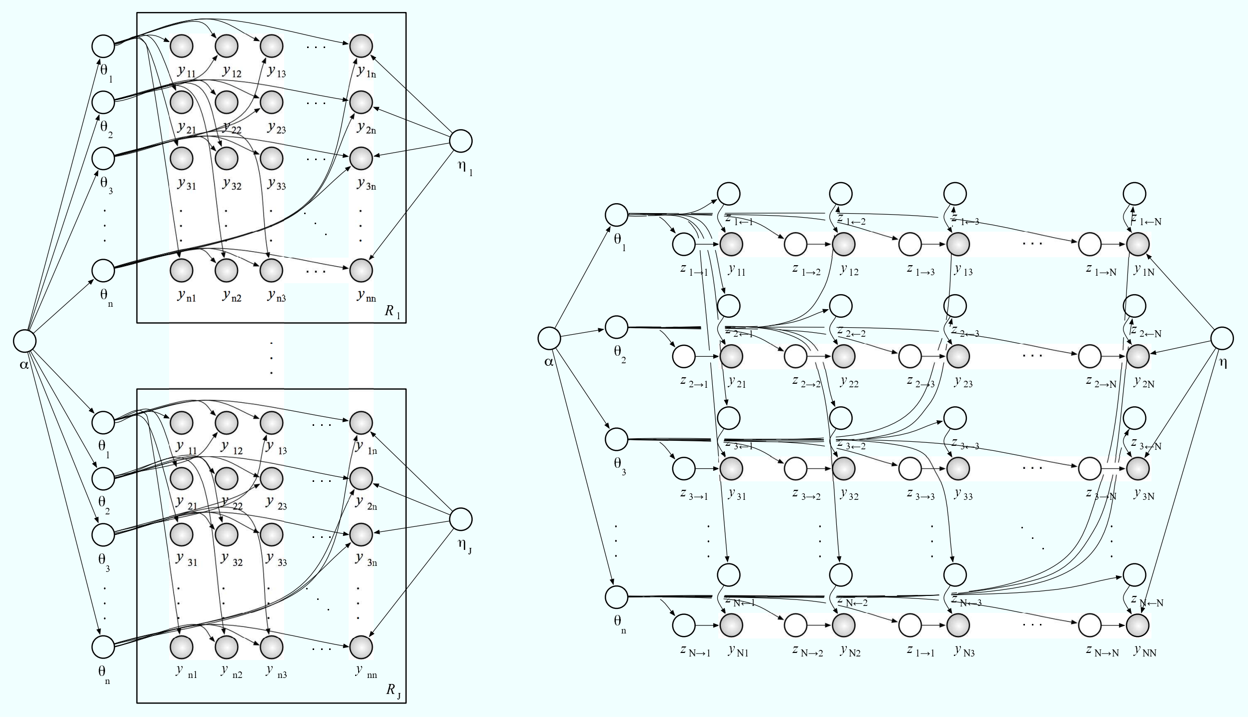

Figure 1. The graphical representation of stochastic block models of mixed membership using plates (left panel), and the graphical representation of the admixture of latent blocks introduced by Airoldi et al. (2006) (right panel). Notes: (i) the mixed-membership vectors, $\theta_{1:N}$, are sampled once for all response graphs, we plotted two sets of them, for clarity; (ii) we did not draw all the arrows out of η, for clarity, since all the interactions y_{nm} depend on it.

non-tractable integral in Equation 1. 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.

The variational method prescribes the use of a mean-field approximation to the posterior distribution of the latent variables given data and hyper-parameters, which we described below. Such an approximation leads to a lower bound for the likelihood of a document, which depends upon an set of free parameters, $\{\gamma_n, \phi_{\overrightarrow{jrnm}}, \phi_{\overleftarrow{jrnm}}\}$. These free parameters are introduced in the mean-field approximation, and are set to minimize the Kullback-Leibler (KL henceforth) divergence between true and approximate posteriors.

The "variational EM" algorithm we develop for performing posterior inference is then an approximate EM algorithm. 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 & Louis, 2005). 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, $\{\gamma_n, \phi_{\overrightarrow{jrnm}}, \phi_{\overleftarrow{jrnm}}\}$, given the most recent estimates for the hyper-parameters.

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, $\{\gamma_n, \phi_{\overrightarrow{jrnm}}, \phi_{\overleftarrow{jrnm}}\}$. 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:

$$\eta^*_{jgh} = \frac{1}{R_j} \sum_{r=1}^{R_j} \frac{\sum_{n,m=1}^{N} \phi_{\overrightarrow{jrnm}g}\, \phi_{\overleftarrow{jrnm}h}\, y_{jrnm}}{\sum_{n,m=1}^{N} \phi_{\overrightarrow{jrnm}g}\, \phi_{\overleftarrow{jrnm}h}}. \tag{9}$$

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, 2000; Blei et al., 2003). We can produce a method that

$$
\begin{aligned}
\frac{\partial L}{\partial \alpha_{[k]}} &= N \left(\Psi \left(\sum_{k=1}^{K} \alpha_k \right) - \Psi(\alpha_k) \right) + \sum_{n=1}^{N} \left(\Psi(\gamma_{nk}) - \Psi \left(\sum_{k=1}^{K} \gamma_{nk} \right) \right), & (4) \\
\frac{\partial L}{\partial \alpha_{k_1} \alpha_{k_2}} &= N \left(\delta_{k_1 = k_2} \cdot \Psi'(\alpha_{k_1}) - \Psi' \left(\sum_{k_2=1}^{K} \alpha_{k_2} \right) \right), & (5) \\
\phi_{\overrightarrow{nmg}} &\propto \exp \left\{ \psi(\gamma_{ng}) - \psi(\sum_{g=1}^{K} \gamma_{ng}) \right\} \cdot \prod_{h=1}^{K} f \left(y_{nm} \mid \eta_{gh} \right)^{\phi_{\overleftarrow{nmh}}} & (6) \\
\phi_{\overleftarrow{nmh}} &\propto \exp \left\{ \psi(\gamma_{mh}) - \psi(\sum_{h=1}^{K} \gamma_{mh}) \right\} \cdot \prod_{g=1}^{K} f \left(y_{nm} \mid \eta_{gh} \right)^{\phi_{\overrightarrow{nmg}}} & (7) \\
\gamma_{ng} &= \alpha_g + \sum_{j=1}^{J} \sum_{r=1}^{R_j} \left(\sum_{m=1}^{N} \phi_{\overrightarrow{jrnmg}} + \sum_{m=1}^{N} \phi_{\overleftarrow{jrnmg}} \right). & (8)
\end{aligned}
$$

is linear in time by using Newton-Raphson, with the gradient and Hessian for the log-likelihood in Equations 4 and 5.

In the approximate E step we update the free parameters for the mean-field approximation of the posterior distribution of the latent variables, $\{\gamma_n, \phi_{\overrightarrow{jrnm}}, \phi_{\overleftarrow{jrnm}}\}$, given the most recent estimates of the hyper-parameters of the model, $(\alpha, \eta_{1:J})$, according to Equations 6, 7 and 8. This minimizes the posterior KL divergence between true and approximate posteriors, at the document 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 fairly general in that they do not depend on a specific observation model. However, our derivations do assume a fully factorized variational distribution.

In order to develop the mean-field approximation for the posterior distribution over the latent variables we used in the E step above, we posit a fully-factorized joint distributions over the latent variables,

$$
\begin{aligned}
& q \left(\{\theta_n, z_{\overrightarrow{jrnm}}, z_{\overleftarrow{jrnm}}\} \mid \{\gamma_n, \phi_{\overrightarrow{jrnm}}, \phi_{\overleftarrow{jrnm}}\} \right) = \\
& = \prod_{n \in \mathcal{N}} q \left(\theta_n | \gamma_n \right) \prod_{j=1}^{J} \prod_{r=1}^{r_j} \prod_{n,m \in \mathcal{N}} \\
& \left(q \left(z_{\overrightarrow{jrnm}} \mid \phi_{\overrightarrow{jrnm}} \right) q \left(z_{\overleftarrow{jrnm}} \mid \phi_{\overleftarrow{jrnm}} \right) \right),
\end{aligned}
$$

which depends on the set of previously mentioned free parameters, $\{\gamma_n, \phi_{\overrightarrow{jrnm}}, \phi_{\overleftarrow{jrnm}}\}$. The mean-field approximation consists in finding an approximate posterior distribution,

$$
\tilde{p} \left(\{\theta_n, z_{\overrightarrow{jrnm}}, z_{\overleftarrow{jrnm}}\} \mid \{\gamma_n, \phi_{\overrightarrow{jrnm}}, \phi_{\overleftarrow{jrnm}}\}, \alpha, \eta_{1:J} \right)
$$

where the conditioning on the data is now obtained indirectly, trough the free parameters,

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

The factorized distribution leads to a lower bound for the likelihood; in fact it is possible to find a closed form solution to the integral in Equation 1 by applying Jensen's inequality, and then integrating the latent variables out with respect to the factorized distribution. An approximate posterior, $\tilde{p}$, is obtained by substituting the lower bound for the likelihood in the calculations, as appropriate. The mean-field approximation in then obtained by minimizing the Kullback-Leibler divergence between the true and the approximate posteriors, over the free parameters.

4.1. The Algorithms: Naïve vs. Nested

The variational inference algorithm presented in Figure 2 is a novel "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 $\phi_{\overrightarrow{nm}}$ and $\phi_{\overleftarrow{nm}}$ 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 experiments the naïve variational algorithm often failed to converge, 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

```
1. initialize $\gamma_{ng}^0 = \frac{2N}{K}$ 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 $\gamma_n^{t+1}$, $\gamma_m^{t+1}$ and $\eta^{t+1}$
7. until convergence
```

```
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 2. **Left:** 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 the right panel. **Right:** Details Step 5. in the left panel; inference 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.

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[2]. 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, $\{\phi_{nm}^{\rightarrow}, \phi_{nm}^{\leftarrow}\}$, 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 MCMC variations (i.e., blocked and collapsed Gibbs samplers) in terms of memory requirements and/or convergence rates.

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.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2006). Mixed membership stochastic block models for relational data with application to protein-protein interactions. *Proceedings of the International Biometrics Socity (ENAR) Annual Meetings*. Forthcoming.

Anderson, C. J., Wasserman, S., & Faust, K. (1992). Building stochastic blockmodels. *Social Networks*, *14*, 137–161.

Blei, D. M., Ng, A., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Carlin, B. P., & Louis, T. A. (2005). *Bayes and Empirical Bayes methods for data analysis*. Chapman & Hall.

Erosheva, E. A., & Fienberg, S. E. (2005). Bayesian mixed membership models for soft clustering and classification. In C. Weihs and W. Gaul (Eds.), *Classification—the ubiquitous challenge*, 11–26. Springer-Verlag.

Holland, P., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: Some first steps. *Social Networks*, *5*, 109–137.

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

Minka, T. (2000). *Estimating a Dirichlet distribution* (Technical Report). M.I.T.

Nowicki, K., & 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., & Park, Y. (2005). Scan statistics on enron graphs. *Computational and Mathematical Organization Theory*, *11*, 229–247.

Sampson, F. (1968). *A novitiate in a period of change: An experimental and case study of social relationships*. Doctoral dissertation, Cornell University.

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