Journal of Machine Learning Research 9 (2008) 1981-2014 Submitted 5/07; Revised 6/08; Published 9/08

Mixed Membership Stochastic Blockmodels

Edoardo M. Airoldi[*] EAIROLDI@PRINCETON.EDU
David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
Princeton, NJ 08544, USA

Stephen E. Fienberg[†] FIENBERG@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Eric P. Xing EPXING@CS.CMU.EDU
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Tommi Jaakkola

Abstract

Consider data consisting of pairwise measurements, such as presence or absence of links between pairs of objects. These data arise, for instance, in the analysis of protein interactions and gene regulatory networks, collections of author-recipient email, and social networks. Analyzing pairwise measurements with probabilistic models requires special assumptions, since the usual independence or exchangeability assumptions no longer hold. Here we introduce a class of variance allocation models for pairwise measurements: mixed membership stochastic blockmodels. These models combine global parameters that instantiate dense patches of connectivity (blockmodel) with local parameters that instantiate node-specific variability in the connections (mixed membership). We develop a general variational inference algorithm for fast approximate posterior inference. We demonstrate the advantages of mixed membership stochastic blockmodels with applications to social networks and protein interaction networks.

Keywords: hierarchical Bayes, latent variables, mean-field approximation, statistical network analysis, social networks, protein interaction networks

1. Introduction

The problem of modeling relational information among objects, such as pairwise relations represented as graphs, arises in a number of settings in machine learning. For example, scientific literature connects papers by citations, the Web connects pages by links, and protein-protein interaction data connects proteins by physical binding records. In these settings, we often wish to infer hidden attributes of the objects from the observed measurements on pairwise properties. For example, we might want to compute a clustering of the web-pages, predict the functions of a protein, or assess

∗. Also in the Lewis-Sigler Institute for Integrative Genomics. Address correspondence to 228 Carl Icahn Laboratory, Princeton University.
†. Also in the School of Computer Science.

the degree of relevance of a scientific abstract to a scholar's query. Unlike traditional data collected from individual objects, *relational data* violate the classical independence or exchangeability assumptions made in machine learning and statistics. The observations are dependent because of the way they are connected. This interdependence suggests that a different set of assumptions is more appropriate.

There is a history of research devoted to analyzing relational data. One well-studied problem is *clustering*, grouping the objects to uncover a structure based on the observed patterns of interactions. Standard model-based clustering methods, for example, mixture models, are not immediately applicable to relational data because they assume that the objects are conditionally independent given their cluster assignments. Rather, the latent stochastic blockmodel (Wang and Wong, 1987; Snijders and Nowicki, 1997) is an adaptation of mixture modeling to relational data. In that model, each object belongs to a cluster and the relationships between objects are governed by the corresponding pair of clusters. With posterior inference, one identifies a set of latent roles which govern the objects relationships with each other. A recent extension of this model relaxed the finite-cardinality assumption on the latent clusters with a nonparametric hierarchical Bayesian model based on the Dirichlet process prior (Kemp et al., 2004, 2006; Xu et al., 2006).

The latent stochastic blockmodel suffers from a limitation that each object can only belong to one cluster, or in other words, play a single latent role. However, many relational data sets are multi-facet. For example, when a protein or a social actor interacts with different partners, different functional or social contexts may apply and thus the protein or the actor may be acting according to different latent roles they can possible play. In this paper, we relax the assumption of single-latent-role for actors, and develop a *mixed membership model* for relational data. Mixed membership models, such as latent Dirichlet allocation (Blei et al., 2003), have re-emerged in recent years as a flexible modeling tool for data where the single cluster assumption is violated by the heterogeneity within of a data point. For almost two decades, these models have been successfully applied in many domains, such as surveys (Berkman et al., 1989; Erosheva, 2002), population genetics (Pritchard et al., 2000), document analysis (Minka and Lafferty, 2002; Blei et al., 2003; Buntine and Jakulin, 2006), image processing (Li and Perona, 2005), and transcriptional regulation (Airoldi et al., 2007).

The mixed membership model associates each unit of observation with multiple clusters rather than a single cluster, via a membership probability-like vector. The concurrent membership of a data in different clusters can capture its different aspects, such as different underlying topics for words constituting each document. This is also a natural idea for relational data, where the objects can bear multiple latent roles or cluster-memberships that influence their relationships to others. As we will demonstrate, a mixed membership approach to relational data lets us describe the interaction between objects playing multiple roles. For example, some of a protein's interactions may be governed by one function; other interactions may be governed by another function.

Existing mixed membership models are not appropriate for relational data because they assume that the data are conditionally independent given their latent membership vectors. In relational data, where each object is described by its relationships to others, we would like to assume that the ensemble of mixed membership vectors help govern the relationships of each object. The conditional independence assumptions of modern mixed membership models do not apply.

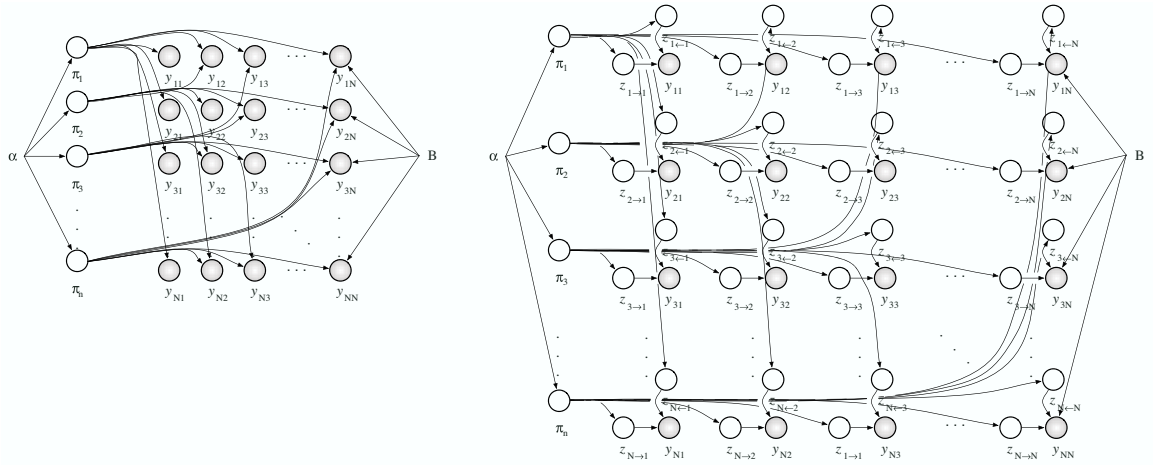

Figure 1: Two graphical model representations of the mixed membership stochastic blockmodel (MMB). Intuitively, the MMB summarized the variability of a graph with the blockmodel B and node-specific mixed membership vectors (left). In detail, a mixed membership, $\pi_n(k)$, quantifies the expected proportion of times node n instantiates the connectivity pattern of group k, according to the blockmodel. In any given interaction, $Y(n,m)$, however, node n instantiates the connectivity pattern of a single group, $z_{n\to m}(k)$. (right). We did not draw all the arrows out of the block model B for clarity; all interactions depend on it.

In this paper, we develop mixed membership models for relational data.[1] Models in this family include parameters to reduce bias due to sparsity, and can be used to analyze multiple collections of paired measurements, and collections of non-binary and multivariate paired measurements. We develop a fast nested variational inference algorithm that performs well in the relational setting and is parallelizable. We demonstrate the application of our technique to large-scale protein interaction networks and social networks. Our model captures the multiple roles that objects exhibit in interaction with others, and the relationships between those roles in determining the observed interaction matrix.

Mixed membership and the latent block structure can be recovered from relational data (Section 4.1). The application to a friendship network among students tests the model on a real data set where a well-defined latent block structure exists (Section 4.2). The application to a protein interaction network tests to what extent our model can reduce the dimensionality of the data, while revealing substantive information about the functionality of proteins that can be used to inform subsequent analyses (Section 4.3).

2. The Mixed Membership Stochastic Blockmodel

In this section, we describe the modeling assumptions if the mixed membership model of relational data. We represent observed relational data as a graph $G = (\mathcal{N}, Y)$, where $Y(p,q)$ maps pairs of nodes to values, that is, edge weights. We consider binary matrices, where $Y(p,q) \in \{0,1\}$. The data can be thought of as a directed graph.

As a running example, we consider the monk data of Sampson (1968). Sampson measured a collection of sociometric relations among a group of monks by repeatedly asking questions such as "whom do you like?" and "whom do you dislike?" to determine asymmetric social relationships within the group. The questionnaire was repeated at four subsequent epochs. Information about these repeated, asymmetric relations was collapsed into a square binary table that encodes the directed connections between monks by Breiger et al. (1975). In analyzing this data, the goal is to determine the social structure within the monastery.

In the context of the monastery example, we assume K factions, that is, latent groups, exist in the monastery, and the observed network is generated according to distributions of group-membership for each monk and a matrix of group-group interaction strength. The per-monk distributions are specified by latent simplicial vectors. Each monk is associated with a randomly drawn vector $\vec{\pi}_i$ for monk i, where $\pi_{i,g}$ denotes the probability of monk i belonging to group g. That is, each monk can simultaneously belong to multiple groups with different degrees of affiliation strength. The probabilities of interactions between different groups are defined by a matrix of Bernoulli rates $B_{(K\times K)}$, where $B(g,h)$ represents the probability of having a link between a monk from group g and a monk from group h.

For each monk, the indicator vector $\vec{z}_{p\to q}$ denotes the group membership of monk p when he responds to survey questions about monk q and $\vec{z}_{p\leftarrow q}$ denotes the group membership of monk q when he responds to survey questions about node p.[2] N denotes the number of monks in the monastery, and recall that K denotes the number of distinct groups a monk can belong to.

More in general, monks can be represented by nodes in a graph, where directed (binary) edges represent positive responses to survey questions about a specific sociometric relation. In this abstract setting, the mixed membership stochastic blockmodel (MMB) posits that a graph $G = (\mathcal{N}, Y)$ is drawn from the following procedure.

- For each node $p \in \mathcal{N}$:
  - Draw a K dimensional mixed membership vector $\vec{\pi}_p \sim$ Dirichlet ($\vec{\alpha}$).
- For each pair of nodes $(p,q) \in \mathcal{N} \times \mathcal{N}$:
  - Draw membership indicator for the initiator, $\vec{z}_{p\to q} \sim$ Multinomial ($\vec{\pi}_p$).
  - Draw membership indicator for the receiver, $\vec{z}_{q\to p} \sim$ Multinomial ($\vec{\pi}_q$).
  - Sample the value of their interaction, $Y(p,q) \sim$ Bernoulli ($\vec{z}_{p\to q}^{\top} B\, \vec{z}_{p\leftarrow q}$).

1. In previous work we combined mixed membership and blockmodels to perform analyses of a single collection of binary, paired measurements; namely, hypothesis testing, predicting and de-noising interactions within an unsupervised learning setting (Airoldi et al., 2005).
2. An indicator vector is used to denote membership in one of the K groups. Such a membership-indicator vector is specified as a K-dimensional vector of which only one element equals to one, whose index corresponds to the group to be indicated, and all other elements equal to zero.

This process is illustrated as a graphical model in Figure 1. Note that the group membership of each node is *context dependent*. That is, each node may assume different membership when interacting to or being interacted by different peers. Statistically, each node is an admixture of group-specific interactions. The two sets of latent group indicators are denoted by $\{\vec{z}_{p\to q} : p,q \in \mathcal{N}\} =: Z_{\to}$ and $\{\vec{z}_{p\leftarrow q} : p,q \in \mathcal{N}\} =: Z_{\leftarrow}$. Also note that the pairs of group memberships that underlie interactions need not be equal; this fact is useful for characterizing asymmetric interaction networks. Equality may be enforced when modeling symmetric interactions.

Under the MMB, the joint probability of the data Y and the latent variables $\{\vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow}\}$ can be written in the following factored form,

$$p(Y,\vec{\pi}_{1:N},Z_{\to},Z_{\leftarrow}|\vec{\alpha},B) = \prod_{p,q} P(Y(p,q)|\vec{z}_{p\to q},\vec{z}_{p\leftarrow q},B)P(\vec{z}_{p\to q}|\vec{\pi}_p)P(\vec{z}_{p\leftarrow q}|\vec{\pi}_q)\prod_p P(\vec{\pi}_p|\vec{\alpha}). \quad (1)$$

This model generalizes to two important cases. First, multiple networks among the same actors can be generated by the same latent vectors. This may be useful, for instance, to analyze multivariate sociometric relations. Second, in the MMB the data generating distribution is a Bernoulli, but B can be a matrix that parameterizes any kind of distribution. This may be useful, for instance, to analyze collections of paired measurements, Y, that take values in an arbitrary metric space. We elaborate on this in Section 5.

2.1 Modeling Sparsity

Adjacency matrices encoding binary pairwise measurements are often sparse, that is, they contain many zeros or non-interactions. It is useful to distinguish two sources of non-interaction: they may be the result of the rarity of interactions in general, or they may be an indication that the pair of relevant blocks rarely interact. In applications to social sciences, for instance, nodes may represent people and blocks may represent social communities. It is reasonable to expect that a large portion of the non-interactions is due to limited opportunities of contact between people rather than due to deliberate choices, the structure of which the blockmodel is trying to estimate. It is useful to account for these two sources of sparsity at the model level. A good estimate of the portion of zeros that should not be explained by the blockmodel B reduces the bias of the estimates of its elements.

Thus, we introduce a sparsity parameter $\rho \in [0,1]$ in the MMB to characterize the source of non-interaction. Instead of sampling a relation $Y(p.q)$ directly the Bernoulli with parameter specified as above, we down-weight the probability of successful interaction to $(1-\rho)\cdot \vec{z}_{p\to q}^{\top} B\, \vec{z}_{p\leftarrow q}$. This is the result of assuming that the probability of a non-interaction comes from a mixture, $1-\sigma_{pq} = (1-\rho)\cdot \vec{z}_{p\to q}^{\top}(1-B)\, \vec{z}_{p\leftarrow q} + \rho$, where the weight ρ capture the portion zeros that should not be explained by the blockmodel B. A large value of ρ will cause the interactions in the matrix to be weighted more than non-interactions, in determining plausible values for $\{\vec{\alpha}, B, \vec{\pi}_{1:N}\}$.

The sparsity parameter ρ can be estimated. Its maximum likelihood estimate provides the best data-driven guess about the proportion of zeros that the blockmodel can explain. Introducing ρ provides a strategy to rescale B, by separating zeros in the adjacency matrix into those that are likely to be due to the blockmodel and those that are not.

2.2 Summarizing and De-Noising Pairwise Measurements

It is useful to distinguish two types of data analysis that can be performed with the mixed-membership blockmodel. First, MMB can be used to summarize the data, Y, in terms of the global blockmodel, B, and the node-specific mixed memberships, Πs. Second, MMB can be used to de-noise the data, Y, in terms of the global blockmodel, B, and interaction-specific single memberships, Zs. In both cases the model depends on a small set of unknown constants to be estimated: α, and B. The likelihood is the same in both cases, although, the rationale for including the set of latent variables Zs differs. When summarizing data, we could integrate out the Zs analytically; this leads to numerical optimization of a smaller set of variational parameters, Γs. We choose to keep the Zs to simplify inference. When de-noising, the Zs are instrumental in estimating posterior expectations of each interactions individually—a network analog to the Kalman Filter. The posterior expectations of an interaction is computed as follows, in the two cases,

$$\mathbb{E}\,[\,Y(p,q)=1\,] \approx \widehat{\vec{\pi}}_p{}' \widehat{B}\, \widehat{\vec{\pi}}_q \qquad \text{and} \qquad \mathbb{E}\,[\,Y(p,q)=1\,] \approx \widehat{\vec{\phi}}_{p\to q}{}' \widehat{B}\, \widehat{\vec{\phi}}_{p\leftarrow q}.$$

2.3 An Illustration: Crisis in a Cloister

To illustrate the MMB, we return to an analysis of the monk data described above. Sampson (1968) surveyed 18 novice monks in a monastery and asked them to rank the other novices in terms of four sociometric relations: like/dislike, esteem, personal influence, and alignment with the monastic credo. We consider Breiger's collation of Sampson's data (Breiger et al., 1975). The original graph of monk-monk interaction is illustrated in Figure 2 (left).

Sampson spent several months in a monastery in New England, where novice monks were preparing to join a monastic order. Sampson's original analysis was rooted in direct anthropological observations. He suggested the existence of tight factions among the novices: the loyal opposition (whose members joined the monastery first), the young turks (who joined later on), the outcasts (who were not accepted in the two main factions), and the waverers (who did not take sides). The events that took place during Sampson's stay at the monastery supported his observations—members of the young turks resigned or were expelled over religious differences (John and Gregory). We shall

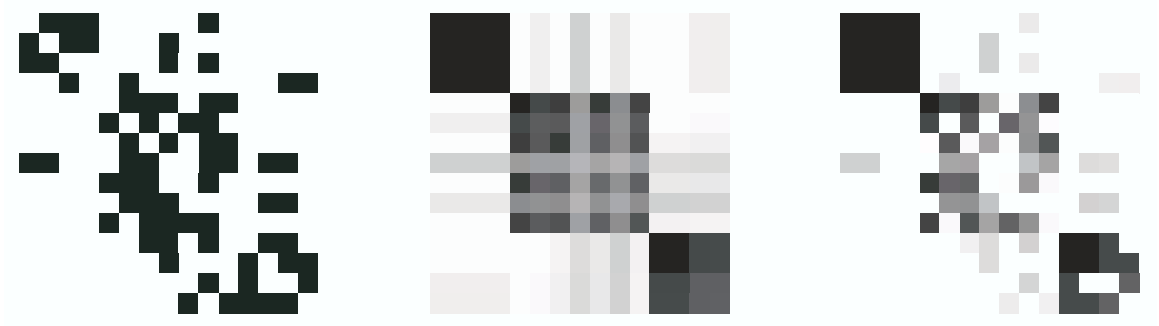

Figure 2: Original adjacency matrix of whom-do-like sociometric relations (left), relations predicted using approximate MLEs for $\vec{\pi}_{1:N}$ and B (center), and relations de-noised using the model including Zs indicators (right).

refer to the labels assigned by Sampson to the novices in the analysis below. For more analyses, we refer to Fienberg et al. (1985), Davis and Carley (2006) and Handcock et al. (2007).

Using the algorithms presented in Section 3, we fit the monks to MMB models for different numbers of groups, providing model estimates $\{\hat{\alpha}, \hat{B}\}$ and posterior mixed membership vectors $\vec{\pi}_n$ for each monk. Here, we use the following approximation to BIC to choose the number of groups in the MMB:

$$BIC = 2 \cdot \log p(Y) \approx 2 \cdot \log p(Y|\widehat{\vec{\pi}}, \widehat{Z}, \widehat{\vec{\alpha}}, \widehat{B}) - |\vec{\alpha}, B| \cdot \log |Y|,$$

which selects three groups, where $|\vec{\alpha}, B|$ is the number of hyper-parameters in the model, and $|Y|$ is the number of positive relations observed (Volinsky and Raftery, 2000; Handcock et al., 2007). Note that this is the same number of groups that Sampson identified. We illustrate the fit of model fit via the predicted network in Figure 2 (Right). The three panels contrast the different resolution of the original adjacency matrix of whom-do-like sociometric relations (left panel) obtained in different uses of MMB. If the goal of the analysis if to find a parsimonious summary of the data, the amount of relational information that is captured by in $\hat{\alpha}, \hat{B}$, and $\mathbb{E}[\vec{\pi}|Y]$ leads to a coarse reconstruction of the original sociomatrix (central panel). If the goal of the analysis if to de-noising a collection of pairwise measurements, the amount of relational information that is revealed by $\hat{\alpha}, \hat{B}, \mathbb{E}[\vec{\pi}|Y]$ and $\mathbb{E}[Z_{\rightarrow}, Z_{\leftarrow}|Y]$ leads to a finer reconstruction of the original sociomatrix, Y—relations in Y are re-weighted according to how much they *make sense* to the model (right panel).

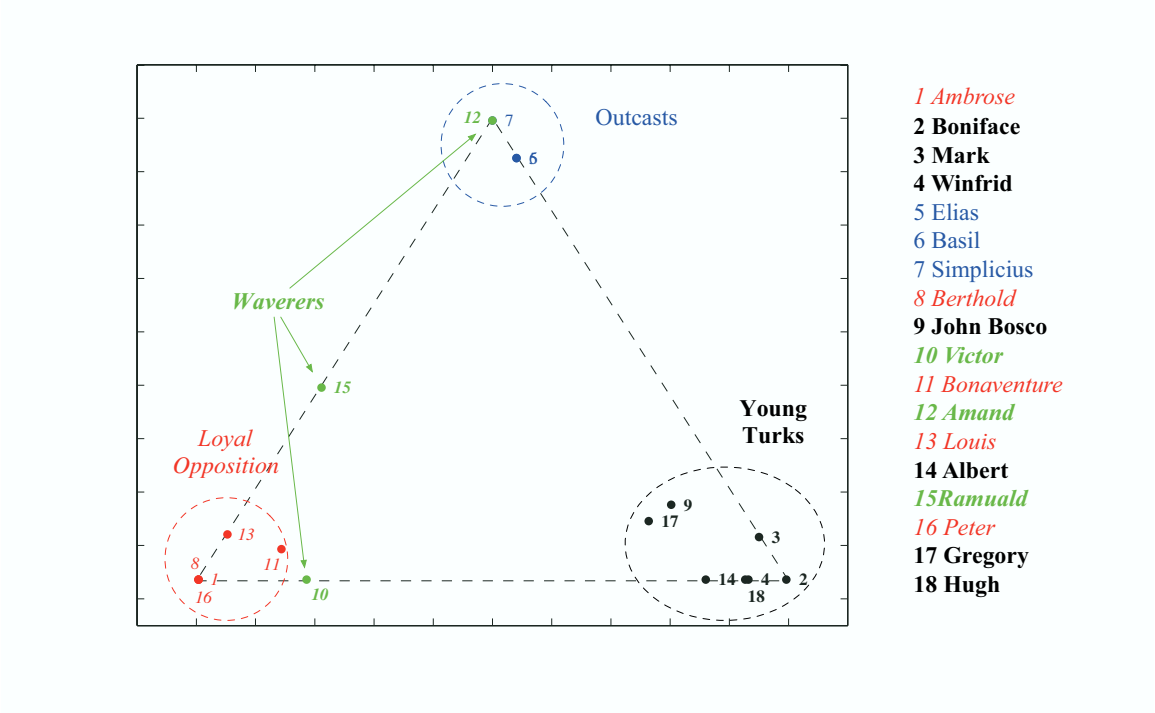

Figure 3: Posterior mixed membership vectors, $\vec{\pi}_{1:18}$, projected in the simplex. Numbered points can be mapped to monks' names using the legend on the right. The colors identify the four factions defined by Sampson's anthropological observations.

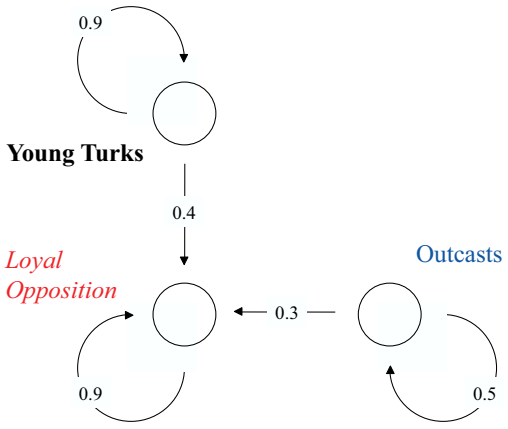

Figure 4: Estimated blockmodel in the monk data, $\hat{B}$.

The MMB provides interesting descriptive statistics about the actors in the observed graph. In Figure 3 we illustrate the the posterior means of the mixed membership scores, $\mathbb{E}[\vec{\pi}|Y]$, for the 18 monks in the monastery. Note that the monks cluster according to Sampson's classification, with Young Turks, Loyal Opposition, and Outcasts dominating each corner respectively. We can see the central role played by John Bosco and Gregory, who exhibit relations in all three groups, as well as the uncertain affiliations of Ramuald and Victor. (Amand's uncertain affiliation, however, is not captured.) The estimated blockmodel is shown in Figure 4.

3. Parameter Estimation and Posterior Inference

Two computational problems are central to the MMB: posterior inference of the per-node mixed membership vectors and per-pair roles, and parameter estimation of the Dirichlet parameters and Bernoulli rate matrix. We derive empirical Bayes estimates of the parameters $(\vec{\alpha}, B)$, and employ a mean-field approximation scheme for posterior inference.

3.1 Posterior Inference

The posterior inference problem is to compute the posterior distribution of the latent variables given a collection of observations. The normalizing constant of the posterior distribution is the marginal probability of the data, which requires an integral over the simplicial vectors $\vec{\pi}_p$,

$$p(Y|\vec{\alpha}, B) = \int_{\Pi} \sum_{Zs} \left(\prod_{p,q} P(Y(p,q)|\vec{z}_{p\to q}, \vec{z}_{p\leftarrow q}, B) P(\vec{z}_{p\to q}|\vec{\pi}_p) P(\vec{z}_{p\leftarrow q}|\vec{\pi}_q) \prod_{p} P(\vec{\pi}_p|\vec{\alpha}) \right) d\vec{\pi},$$

which is not solvable in closed form (Blei et al., 2003). A number of approximate inference algorithms for mixed membership models have appeared in recent years, including mean-field variational methods (Blei et al., 2003; Teh et al., 2007), expectation propagation (Minka and Lafferty, 2002), and Monte Carlo Markov chain sampling (MCMC) (Erosheva and Fienberg, 2005; Griffiths and Steyvers, 2004).

We appeal to variational methods (Jordan et al., 1999; Wainwright and Jordan, 2003). The main idea behind variational methods is to first posit a distribution of the latent variables with free parameters, and then fit those parameters such that the distribution is close in Kullback-Leibler divergence to the true posterior. The variational distribution is simpler than the true posterior so that the optimization problem can be approximately solved. Good reviews of variational methods can be found in Wainwright and Jordan (2003), Xing et al. (2003), Bishop et al. (2003) and Airoldi (2007).

In the MMB, we begin by bounding the log of the marginal probability of the data with Jensen's inequality,

$$\log p(Y \mid \alpha, B) \geq \mathbb{E}_q \left[\log p(Y, \vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \alpha, B) \right] - \mathbb{E}_q \left[\log q(\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}) \right].$$

We have introduced a distribution of the latent variables q that depends on a set of free parameters. We specify q as the mean-field fully-factorized family,

$$q(\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow} | \vec{\gamma}_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow}) = \prod_p q_1(\vec{\pi}_p | \vec{\gamma}_p) \prod_{p,q} \left(q_2(\vec{z}_{p \rightarrow q} | \vec{\phi}_{p \rightarrow q}) \, q_2(\vec{z}_{p \leftarrow q} | \vec{\phi}_{p \leftarrow q}) \right),$$

where q_1 is a Dirichlet, q_2 is a multinomial, and $\{\vec{\gamma}_{1:N}, \Phi_{\rightarrow}, \Phi_{\leftarrow}\}$ are the set of free *variational parameters* that are optimized to tighten the bound.

Tightening the bound with respect to the variational parameters is equivalent to minimizing the KL divergence between q and the true posterior. When all the nodes in the graphical model are conjugate pairs or mixtures of conjugate pairs, we can directly write down a coordinate ascent algorithm for this optimization to reach a local maximum of the bound. The updates for the variational multinomial parameters are

$$\hat{\phi}_{p \rightarrow q, g} \propto e^{\mathbb{E}_q[\log \pi_{p,g}]} \cdot \prod_h \left(B(g,h)^{Y(p,q)} \cdot (1 - B(g,h))^{1 - Y(p,q)} \right)^{\phi_{p \leftarrow q, h}} \tag{2}$$

$$\hat{\phi}_{p \leftarrow q, h} \propto e^{\mathbb{E}_q[\log \pi_{q,h}]} \cdot \prod_g \left(B(g,h)^{Y(p,q)} \cdot (1 - B(g,h))^{1 - Y(p,q)} \right)^{\phi_{p \rightarrow q, g}}, \tag{3}$$

for $g, h = 1, \ldots, K$. The update for the variational Dirichlet parameters $\gamma_{p,k}$ is

$$\hat{\gamma}_{p,k} = \alpha_k + \sum_q \phi_{p \rightarrow q, k} + \sum_q \phi_{p \leftarrow q, k}, \tag{4}$$

for all nodes $p = 1, \ldots, N$ and $k = 1, \ldots, K$. The complete coordinate ascent algorithm is described in Figure 5.

To improve convergence, we employed a nested variational inference scheme based on an alternative schedule of updates to the traditional ordering. In a typical schedule for coordinate ascent (which we call "naïve variational inference"), one initializes the variational Dirichlet parameters $\vec{\gamma}_{1:N}$ and the variational multinomial parameters $(\vec{\phi}_{p \rightarrow q}, \vec{\phi}_{p \leftarrow q})$ to non-informative values, and then iterates the following two steps until convergence: (i) update $\vec{\phi}_{p \rightarrow q}$ and $\phi_{p \leftarrow q}$ for all edges (p,q), and (ii) update $\vec{\gamma}_p$ for all nodes $p \in \mathcal{N}$. In such algorithm, at each variational inference cycle we need to allocate $NK + 2N^2K$ scalars.

In our experiments, the naïve variational algorithm often converged only after many iterations. We attribute this behavior to the dependence between $\vec{\gamma}_{1:N}$ and B, which is not satisfied by the naïve algorithm. Some intuition about why this may happen follows. From a purely algorithmic

1. initialize $\vec{\gamma}^0_{pk} = \frac{2N}{K}$ for all p,k
2. **repeat**
3. **for** $p = 1$ to N
4. **for** $q = 1$ to N
5. get **variational** $\vec{\phi}^{t+1}_{p\to q}$ and $\vec{\phi}^{t+1}_{p\leftarrow q} = f\ (\ Y(p,q), \vec{\gamma}^t_p, \vec{\gamma}^t_q, B^t\ )$
6. partially update $\vec{\gamma}^{t+1}_p$, $\vec{\gamma}^{t+1}_q$ and B^{t+1}
7. **until** convergence

5.1. initialize $\phi^0_{p\to q,g} = \phi^0_{p\leftarrow q,h} = \frac{1}{K}$ for all g,h
5.2. **repeat**
5.3. **for** $g = 1$ to K
5.4. update $\phi^{s+1}_{p\to q} \propto f_1\ (\ \vec{\phi}^s_{p\leftarrow q}, \vec{\gamma}_p, B\ )$
5.5. normalize $\vec{\phi}^{s+1}_{p\to q}$ to sum to 1
5.6. **for** $h = 1$ to K
5.7. update $\phi^{s+1}_{p\leftarrow q} \propto f_2\ (\ \vec{\phi}^s_{p\to q}, \vec{\gamma}_q, B\ )$
5.8. normalize $\vec{\phi}^{s+1}_{p\leftarrow q}$ to sum to 1
5.9. **until** convergence

Figure 5: **Top:** The two-layered variational inference for $(\vec{\gamma}, \phi_{p\to q,g}, \phi_{p\leftarrow q,h})$ and $M = 1$. The inner algorithm consists of Step 5. The function f is described in details in the bottom panel. The partial updates in Step 6 for $\vec{\gamma}$ and B refer to Equation 4 of Section B.4 and Equation 5 of Section B.5, respectively. **Bottom:** Inference for the variational parameters $(\vec{\phi}_{p\to q}, \vec{\phi}_{p\leftarrow q})$ corresponding to the basic observation $Y(p,q)$. This nested algorithm details Step 5 in the top panel. The functions f_1 and f_2 are the updates for $\phi_{p\to q,g}$ and $\phi_{p\leftarrow q,h}$ described in Equations 2 and 3 of Section B.4.

perspective, the naïve variational EM algorithm instantiates a large coordinate ascent algorithm, where the parameters can be divided into blocks. Blocks are processed in a specific order, and the parameters within each block get all updated each time.[3] At every new iteration the naïve algorithm sets all the elements of $\vec{\gamma}^{t+1}_{1:N}$ equal to the same constant. This dampens the likelihood by suddenly breaking the dependence between the estimates of parameters in $\hat{\vec{\gamma}}^t_{1:N}$ and in $\hat{B}^t$ that was being inferred from the data during the previous iteration.

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

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

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, $(\vec{\phi}_{p\to q}, \vec{\phi}_{p\leftarrow q})$, optimized given the other variational parameters. Note that these parameters are involved in the updates of parameters in $\vec{\gamma}_{1:N}$ and in B, thus providing us with a channel to maintain some of the dependence among them, that is, by keeping them at their optimal value given the data.

Furthermore, 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 only. The increased running time is partially offset by the fact that the algorithm can be parallelized and leads to empirically observed faster convergence rates.

An alternative strategy to perform inference is given by Monte Carlo Markov chain (e.g., see Griffiths and Steyvers, 2004; Kemp et al., 2004). While powerful in some settings, MCMC is impractical here. There are too many variables to sample. The proposed nested variational EM algorithm outperforms MCMC variations (i.e., blocked and collapsed Gibbs samplers) in terms of memory requirements and convergence rates.

3.2 Parameter Estimation

We compute the empirical Bayes estimates of the model hyper-parameters $\{\vec{\alpha}, B\}$ with a variational expectation-maximization (EM) algorithm. Alternatives to empirical Bayes have been proposed to fix the hyper-parameters and reduce the computation. The results, however, are not always satisfactory and often times cause of concern, since the inference is sensitive to the choice of the hyper-parameters (Joutard et al., 2007). Empirical Bayes, on the other hand, guides the posterior inference towards a region of the hyper-parameter space that is supported by the data.

Variational EM uses the lower bound in Equation 5 as a surrogate for the likelihood. To find a local optimum of the bound, we iterate between fitting the variational distribution q to approximate the posterior and maximizing the corresponding bound with respect to the parameters. The latter M-step is equivalent to finding the MLE using expected sufficient statistics under the variational distribution. We consider the maximization step for each parameter in turn.

A closed form solution for the approximate maximum likelihood estimate of $\vec{\alpha}$ does not exist (Minka, 2003). We use a linear-time Newton-Raphson method, where the gradient and Hessian are

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_k} &= N\left(\psi\left(\sum_k \alpha_k\right) - \psi(\alpha_k)\right) + \sum_p \left(\psi(\gamma_{p,k}) - \psi\left(\sum_k \gamma_{p,k}\right)\right), \\
\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_{k_1}\alpha_{k_2}} &= N\left(\mathbb{I}_{(k_1=k_2)} \cdot \psi'(\alpha_{k_1}) - \psi'\left(\sum_k \alpha_k\right)\right).
\end{aligned}$$

The approximate MLE of B is

$$\hat{B}(g,h) = \frac{\sum_{p,q} Y(p,q) \cdot \phi_{p\to qg}\, \phi_{p\leftarrow qh}}{(1-\rho) \cdot \sum_{p,q} \phi_{p\to qg}\, \phi_{p\leftarrow qh}},$$

for every index pair $(g,h) \in [1,K] \times [1,K]$. Finally, the approximate MLE of the sparsity parameter ρ is

$$\hat{\rho} = \frac{\sum_{p,q} \left(1 - Y(p,q)\right) \cdot \left(\sum_{g,h} \phi_{p\to qg}\, \phi_{p\leftarrow qh}\right)}{\sum_{p,q} \sum_{g,h} \phi_{p\to qg}\, \phi_{p\leftarrow qh}}.$$

Alternatively, we can fix ρ prior to the analysis; the density of the interaction matrix is estimated with $\hat{d} = \sum_{p,q} Y(p,q)/N^2$, and the sparsity parameter is set to $\tilde{\rho} = (1 - \hat{d})$. This latter estimator attributes all the information in the non-interactions to the point mass, that is, to latent sources other than the block model B or the mixed membership vectors $\vec{\pi}_{1:N}$. It does however provide a quick recipe to reduce the computational burden during exploratory analyses.[4]

Several model selection strategies are available for complex hierarchical models (Joutard et al., 2007). In our setting, model selection translates into the determination of a plausible value of the number of groups K. In the various analyses presented, we selected the optimal value of K according to two strategies. On large networks, we selected K corresponding to the highest averaged held-out likelihood in a cross-validation experiment. On small networks—where cross-validation cannot be expected to work well, as we discuss in Section 5—we selected K using an approximation to BIC.

4. Experiments and Results

We present a study of simulated data and applications to social and protein interaction networks.

Simulations are performed in Section 4.1 to show that both mixed membership, $\vec{\pi}_{1:N}$, and the latent block structure, B, can be recovered from data, when they exist, and that the nested variational inference algorithm is faster than the naïve implementation while reaching the same peak in the likelihood—all other things being equal.

The application to a friendship network among students in Section 4.2 tests the model on a real data set where we expect a well-defined latent block structure to inform the observed connectivity patterns in the network. In this application, the blocks are interpretable in terms of grades. We compare our results with those that were recently obtained with a simple mixture of blocks (Doreian et al., 2007) and with a latent space model (Handcock et al., 2007) on the same data.

The application to a protein interaction network in Section 4.3 tests the model on a real data set where we expect a noisy, vague latent block structure to inform the observed connectivity patterns in the network to some degree. In this application, the blocks are interpretable in terms functional biological contexts. This application tests to what extent our model can reduce the dimensionality of the data, while revealing substantive information about the functionality of proteins that can be used to inform subsequent analyses.

4.1 Exploring Expected Model Behavior with Simulations

In developing the MMB and the corresponding computation, our hope is the the model can recover both the mixed membership of nodes to clusters and the latent block structure among clusters in situations where a block structure exists and the relations are measured with some error. To substantiate this claim, we sampled graphs of $100, 300,$ and 600 nodes from blockmodels with $4, 10,$ and 20 clusters, respectively, using the MMB. We used different values of α to simulate a range of settings in terms of membership of nodes to clusters—from unique $(\alpha = 0.05)$ to mixed $(\alpha = 0.25)$.

Recovering the truth. The variational EM algorithm successfully recovers both the latent block model B and the latent mixed membership vectors $\vec{\pi}_{1:N}$. In Figure 6 we show the adjacency matrices of binary interactions where rows, that is, nodes, are reordered according to their most likely membership. The estimated reordering reveals the block model that was originally used to simulate

4. Note that $\tilde{\rho} = \hat{\rho}$ in the case of single membership. In fact, that implies $\phi^m_{p \to qg} = \phi^m_{p \leftarrow qh} = 1$ for some (g,h) pair, for any (p,q) pair.

the interactions. As α increases, each node is likely to belong to more clusters. As a consequence, they express interaction patterns of clusters. This phenomenon reflects in the reordered interaction matrices as the block structure is less evident.

Nested variational inference. The nested variational algorithm drives the log-likelihood to converge faster to its peak than the naïve algorithm. In Figure 7 (left panel) 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. Furthermore, the nested variational algorithm can be parallelized given that the updates for each interaction (i, j) are independent of one another.

Choosing the number of blocks. The right panel of Figure 7 shows an example where cross-validation is sufficient to perform model selection for the MMB. The example shown corresponds to a network among 300 nodes with $K = 10$ clusters. We measure the number of latent clusters

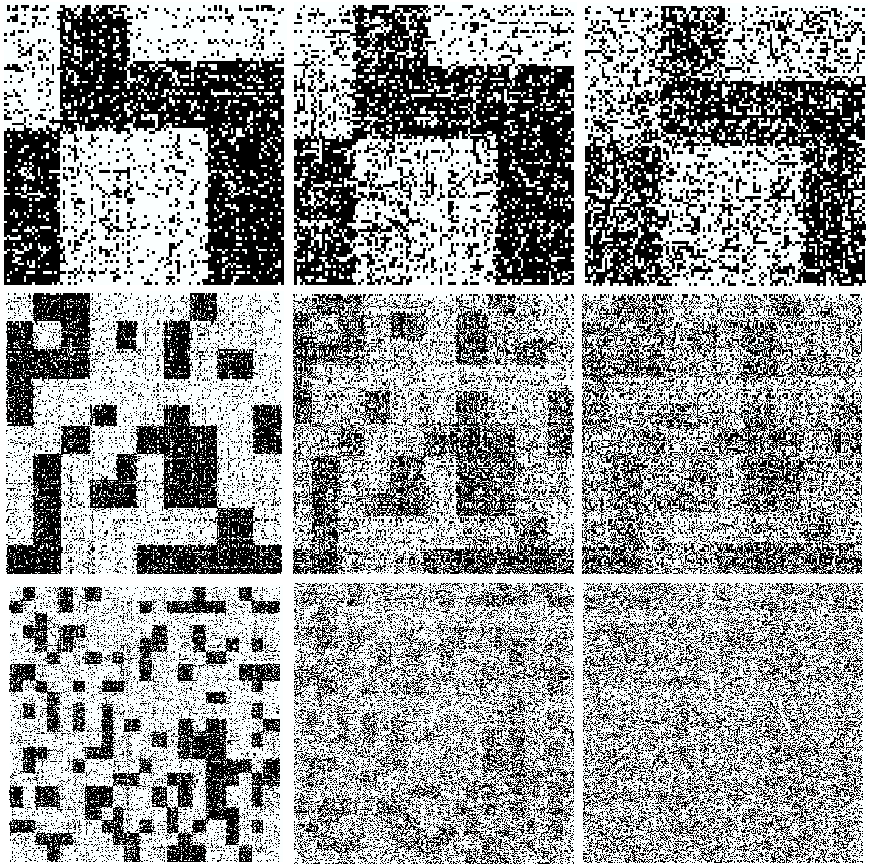

Figure 6: Adjacency matrices of corresponding to simulated interaction graphs with 100 nodes and 4 clusters, 300 nodes and 10 clusters, 600 nodes and 20 clusters (top to bottom) and α equal to $0.05, 0.1$ and 0.25 (left to right). Rows, which corresponds to nodes, are reordered according to their most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

on the X axis and the average held-out log-likelihood, corresponding to five-fold cross-validation experiments, on the Y axis. The nested variational EM algorithm was xrun until convergence, for each value of K we tested, with a tolerance of $\varepsilon = 10^{-5}$. Our estimate for K occurs at the peak in the average held-out log-likelihood, and equals the correct number of clusters, $K^* = 10$

4.2 Application to Social Network Analysis

We considered a friendship network among a group of 69 students in grades 7–12. The analysis here directly compares clustering results obtained by MMB to published clustering results obtained by competing models, in a setting where a fair amount of social segregation is expected (Doreian et al., 2007; Handcock et al., 2007).

The National Longitudinal Study of Adolescent Health is nationally representative study that explores the how social contexts such as families, friends, peers, schools, neighborhoods, and communities influence health and risk behaviors of adolescents, and their outcomes in young adulthood (Harris et al., 2003; Udry, 2003). As part of the survey, a questionnaire was administered to a sample of students in each school, who were allowed to nominate up to 10 friends. We analyzed a friendship network among the students, at the same school that was considered by Handcock et al. (2007) and discussants. Friendship nominations were collected among 71 students in grades 7 to 12; two students did not nominate any friends. The network of binary, asymmetric friendship relations among the remaining 69 students that constitutes our data is shown in Figure 9 (left).

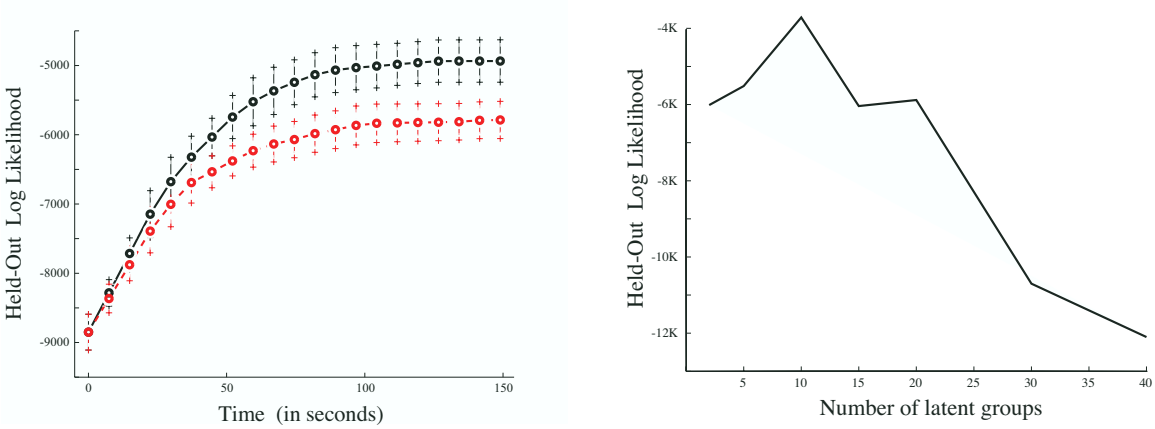

Figure 7: **Left:** The running time of the naïve variational inference (dashed, red line) against the running time of our enhanced (nested) variational inference algorithm (solid, black line), on a graph with 100 nodes and 4 clusters. We measure the number of seconds on the X axis and the log-likelihood on the Y axis. The two curves are averages over 26 experiments, and the error bars are at three standard deviations. Each of the 26 pairs of experiments was initialized with the same values for the parameters. **Right:** 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 identifies the correct number of clusters, $K^* = 10$

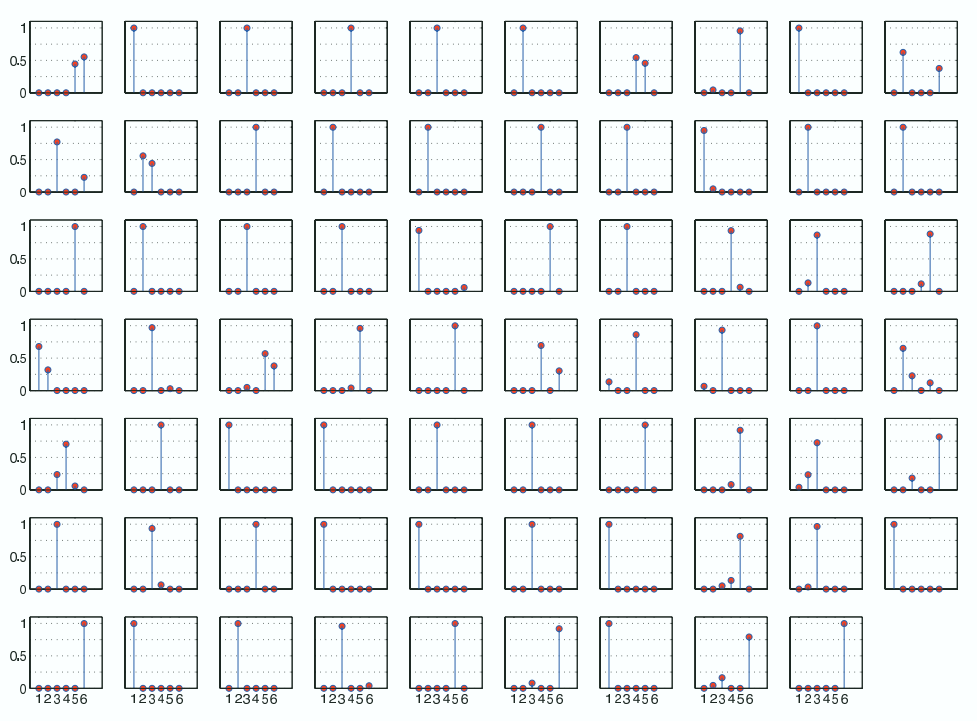

Figure 8: The posterior mixed membership scores, $\vec{\pi}$, for the 69 students. Each panel correspond to a student; we order the clusters 1 to 6 on the X axis, and we measure the student's grade of membership to these clusters on the Y axis.

Given the size of the network we used BIC to perform model selection, as in the monks example of Section 2.3. The results suggest a model with $K^* = 6$ groups. (We fix $K^* = 6$ in the analyses that follow.) The hyper-parameters estimated with the nested variational EM. They are $\hat{\alpha} = 0.0487$, $\hat{\rho} = 0.936$, and a fairly diagonal blockmodel,

$$\hat{B} = \begin{bmatrix} 0.3235 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.3614 & 0.0002 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.2607 & 0.0 & 0.0 & 0.0002 \\ 0.0 & 0.0 & 0.0 & 0.3751 & 0.0009 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0002 & 0.3795 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3719 \end{bmatrix}.$$

Figure 8 shows the expected posterior mixed membership scores for the 69 students in the sample; few students display mixed membership. The rarity of mixed membership in this context is expected, while mixed membership may signal unexpected social situations for further investigation. For instance, it may signal a family bond such as brotherhood, or a student that is repeating a grade and is thus part of a broader social clique. In Figure 9, we contrast the friendship relation data (left) to the estimates obtained by thresholding the estimated probabilities of a relation, using the blockmodel and the node-specific latent variables (center) and the interactions-specific latent variables (right). The model provides a good summary of the social structure in the school; students

tend to befriend other students in the same grade, with a few exceptions. The low degree of mixed membership explains the absence of obvious differences between the model-based reconstructions of the friendship relations with the two model variants (center and right).

Figure 9: Original matrix of friensdhip relations among 69 students in grades 7 to 12 (left), and friendship estimated relations obtained by thresholding the posterior expectations $\vec{\pi}_p{}' B \vec{\pi}_q | Y$ (center), and $\vec{\phi}_p{}' B \vec{\phi}_q | Y$ (right).

Next, we attempted a quantitative evaluation of the goodness of fit. In this data, the blocks are clearly interpretable a-posteriori in terms of grades. The mixed membership vectors provide a mapping between grades and blocks. Conditionally on such a mapping, we assign students to the grade they are most associated with, according to their posterior-mean mixed membership vectors, $\mathbb{E}[\vec{\pi}_n | Y]$. To be fair in the comparison with competing models, we assign students to a unique grade—despite MMB allows for mixed membership. Table 1 computes the correspondence of grades to blocks by quoting the number of students in each grade-block pair, for MMB versus the mixture blockmodel (MB) in Doreian et al. (2007), and the latent space cluster model (LSCM) in Handcock et al. (2007). The higher the sum of counts on diagonal elements is the better is the correspondence, while the higher the sum of counts off diagonal elements is the worse is the correspondence. MMB performs best by allocating 63 students to their grades, versus 57 of MB, and 37 of LSCM. Correspondence only partially captures goodness of fit, however, it is a good metric in the setting we consider, where a fair amount of clustering is present. The results suggest that the extra-flexibility MMB offers over MB and LSCM reduces bias in the prediction of the membership of students to blocks. In other words, mixed membership does not absorb noise in this example; rather it accommodates variability in the friendship relation that is instrumental in producing better predictions.

Concluding this example, we note how the model decouples the observed friendship patterns into two complementary sources of variability. On the one hand, the connectivity matrix B is a global, unconstrained set of hyper-parameters. On the other hand, the mixed membership vectors $\vec{\pi}_{1:N}$ provide a collection of node-specific latent vectors, which inform the directed connections in the graph in a symmetric fashion.

4.3 Application to Protein Interactions in *Saccharomyces Cerevisiae*

We considered physical interactions among 871 proteins in yeast. The analysis allows us to evaluate the utility of MMB in summarizing and de-noising complex connectivity patterns quantitatively, using an independent set of functional annotations. For instance, between two models that sug-

	MMB Clusters						MSB Clusters						LSCM Clusters					
Grade	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
7	13	1	0	0	0	0	13	1	0	0	0	0	13	1	0	0	0	0
8	0	9	2	0	0	1	0	10	2	0	0	0	0	11	1	0	0	0
9	0	0	16	0	0	0	0	0	10	0	0	6	0	0	7	6	3	0
10	0	0	0	10	0	0	0	0	0	10	0	0	0	0	0	0	3	7
11	0	0	1	0	11	1	0	0	1	0	11	1	0	0	0	0	3	10
12	0	0	0	0	0	4	0	0	0	0	0	4	0	0	0	0	0	4

Table 1: Grade levels versus (highest) expected posterior membership for the 69 students, according to three alternative models. MMB is the proposed mixed membership stochastic blockmodel, MSB is a simpler stochastic block mixture model (Doreian et al., 2007), and LSCM is the latent space cluster model (Handcock et al., 2007).

gest different sets of interactions as reliable, we prefer the model that reveals *functionally relevant* interactions—as measured using the annotations.

Protein interactions (PPI) form the physical basis for the formation of stable protein complexes (i.e., protein clusters) and signaling pathways (i.e., cascades of protein interaction events) that carry out all major biological processes in the cell. A number of high-throughput experimental technologies have been devised to determine the set of interacting proteins on a global scale in yeast. These include two-hybrid (Y2H) screens and mass spectrometry methods (Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006). High-throughput technologies, however, often miss to identify interactions that are not present under the given conditions. Specific wet-lab methods employed by a certain technology, such as tagging, may disturb the formation of a stable protein complex, and weakly associated components may dissociate and escape detection. Statistical models that encode information about functional processes with high precision are an essential tool for carrying out probabilistic de-noising of biological signals from high-throughput experiments.

The goal of the analysis of protein interactions with MMB is to reveal the proteins' diverse functional roles by analyzing their local and global patterns of interaction. The biochemical composition of individual proteins make them suitable for carrying out a specific set of cellular operations, or *functions*. The main intuition behind our methodology is that pairs of protein interact because they participate in the same cellular process, as part of the same stable protein complex, that is, co-location, or because they are part of interacting protein complexes, as they carry out compatible cellular operations (Alberts et al., 2002). Below, we describe the MIPS protein interactions data and the possible interpretations of the blocks in MMB in terms of biological functions, and we report results of two experiments.

4.3.1 PROTEIN INTERACTION DATA AND FUNCTIONAL ANNOTATION DATA

The Munich Institute for Protein Sequencing (MIPS) database was created in 1998 based on evidence derived from a variety of experimental techniques (Mewes et al., 2004). It includes a hand-curated collection of protein interactions that does not include interactions obtained with high-throughput technologies. The collection covers about 8000 protein complex associations in yeast.

We analyzed a subset of this collection containing 871 proteins, the interactions amongst which were hand-curated.

The MIPS institute also provides a set of functional annotations for each protein. These annotations are organized in a tree, with 15 nodes (i.e., high-level functions) at the first level, 72 nodes (i.e., the mid-level functions) at the second level, and 255 nodes (i.e., the low-level functions) at the the leaf level. We mapped the 871 proteins in our collections to the high-level functions of the MIPS annotation tree. Table 2 quotes the number of proteins annotated to each of these 15 functions. Most proteins participate in more than one functional category, with an average of ≈ 2.4 functional annotations for each protein.. The relative importance of functional categories in our collection, in terms of the number of proteins involved, is similar to the relative importance of functional categories over the entire MIPS collection. We can also represent each protein in terms of its MIPS functional annotations. This leads to a 15-dimensional, binary representation for each protein, $\vec{b}_p$, where a component $\vec{b}_p(k) = 1$ indicates that protein p is annotated with function k in Table 2. Figure 10 shows the binary representations, $\vec{b}_{1:871}$, of the proteins in our collections; each panel corresponds to a protein; the 15 functional categories are ordered as in Table 2 on the X axis, whereas the presence or absence of the corresponding functional annotation is displayed on the Y axis. In Section 4.3.2, we fit a mixed membership blockmodel with $K = 15$, and we explore the direct correspondence between protein-specific mixed memberships to blocks, $\vec{\pi}_{1:871}$, and MIPS-derived functional annotations, $\vec{b}_{1:871}$.

An alternative source of functional annotations is the gene ontology (GO), distributed as part of the Saccharomyces genome database (Ashburner et al., 2000). GO provides vocabularies for describing the molecular function, biological process, and cellular component of gene products—such as proteins. Terms are organized in a directed acyclic graph. Terms at the top represent broader, more general concepts, terms lower down represent more specific concepts. There are two different relationship types between (parent-child) terms: "is a" and "part of". Proteins are annotated to terms, and, most importantly, a protein is typically annotated to multiple terms, in different portions of the GO annotation graph. We restrict our evaluations to a collection of GO terms that is specific enough for a co-annotation (i.e., two proteins annotated to the same term) to be functionally relevant to molecular biologists (Myers et al., 2006). In Section 4.3.3, we select the mixed membership blockmodel best for predicting out-of-sample interactions, corresponding 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 2: The 15 high-level functional categories obtained by cutting the MIPS annotation tree at the first level and how many proteins (out of 871) participate in each.

$K^* = 50$, and we explore its goodness-of-fit indirectly—rather than attempting a direct interpretation of the model's parameters—, in terms of the number of predicted interactions that are functionally relevant according to GO functional annotations.

4.3.2 DIRECT EVALUATION: THE MODEL CAPTURES SUBSTANTIVE BIOLOGY

In the first experiment, we fit a model with $K = 15$ blocks, and we attempt a direct interpretation of the blocks in terms of the 15 high-level functional categories in the MIPS annotation tree—separate from the MIPS protein interaction data, and independently conceived. We discuss results

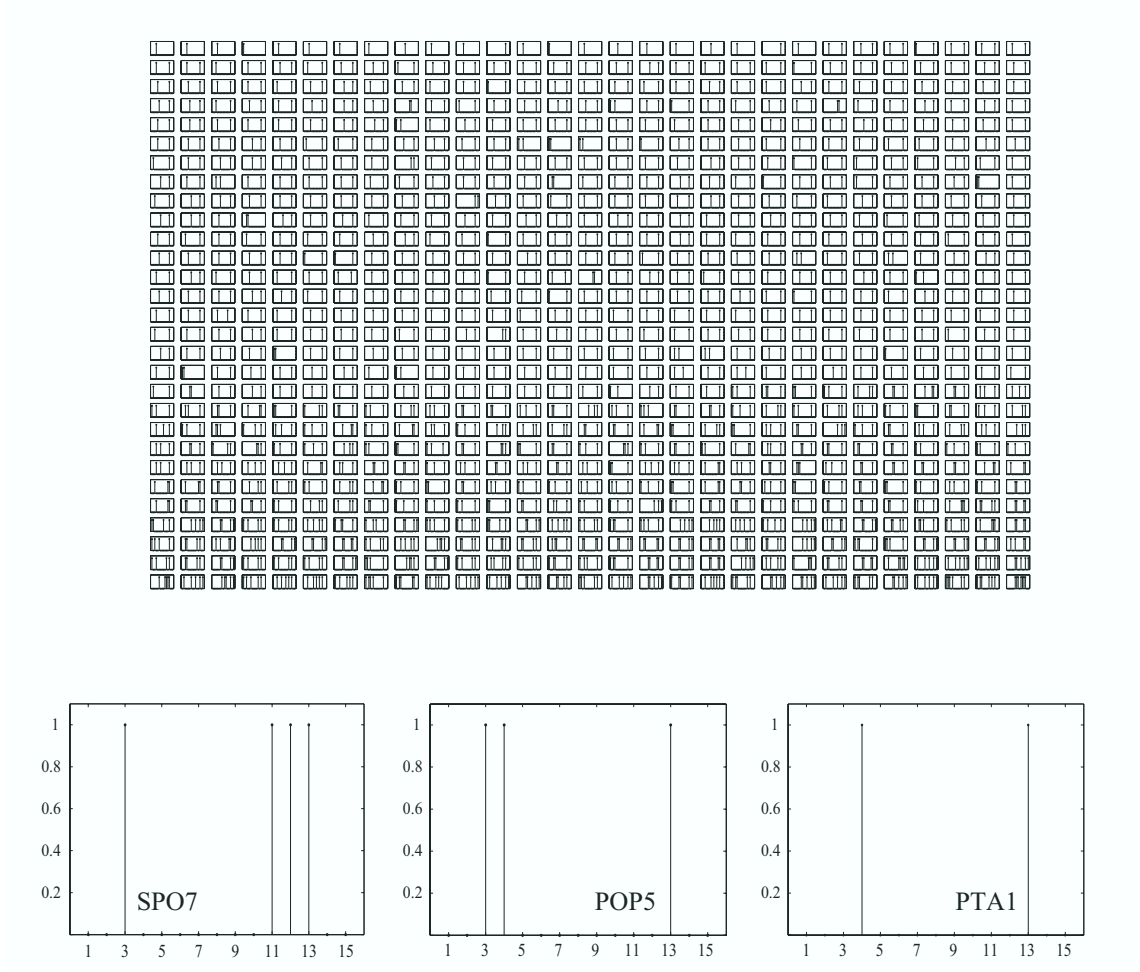

Figure 10: By mapping individual proteins to the 15 general functions in Table 2, we obtain a 15-dimensional representation for each protein. Here, each panel corresponds to a protein; the 15 functional categories are displayed on the X axis, whereas the presence or absence of the corresponding functional annotation is displayed on the Y axis. The plots at the bottom zoom into three example panels (proteins).

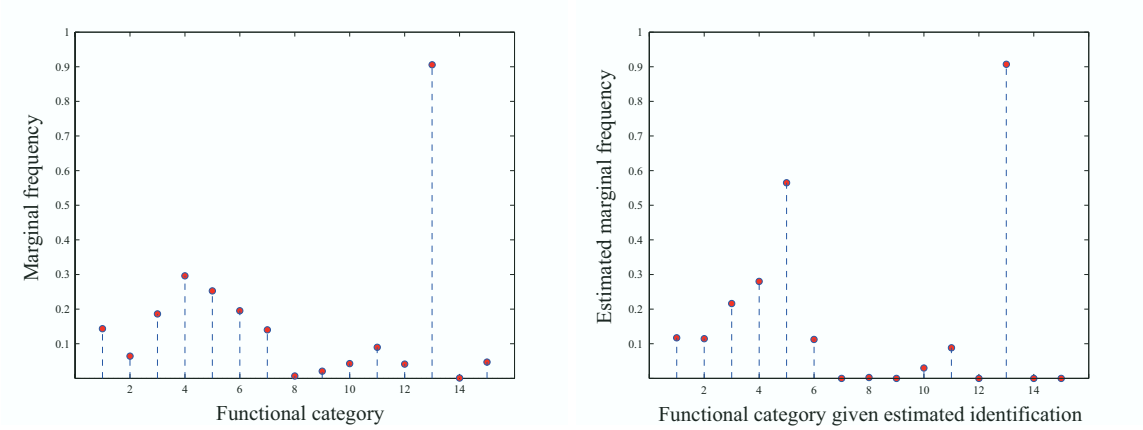

Figure 11: The mapping of blocks to functions is estimated by maximizing the accuracy of the predicted annotations of 87 proteins. We plot marginal frequencies of proteins' membership to true functions (left) and to predicted functions (right).

that portray the relevance of mixed membership, the resolution of the identification of blocks with functional categories, and selected predictions.

We want to compute the correspondence between protein-specific mixed memberships to blocks, $\vec{\pi}_{1:871}$, and MIPS-derived functional annotations, $\vec{b}_{1:871}$. The $K = 15$ blocks in the blockmodel B are not directly identifiable in terms of functional categories. In other words, we need to estimate a permutation of the components of $\vec{\pi}_n$ in order to be able to interpret $E[\pi_n(k)|Y]$ as the expected degree of membership of protein n in function k of Table 2—rather than simply the expected degree of membership of protein n in block k, out of 15. To estimate the permutation that best identifies blocks to functions, we proceeded as follows. We sampled 87 proteins and their corresponding MIPS annotations, $\vec{b}_{1:87}$. We predicted membership of the 87 proteins by thresholding their mixed membership representations,

$$\hat{b}_n(k) = \begin{cases} 1 & \text{if } \pi_n(k) > \tau \\ 0 & \text{otherwise,} \end{cases}$$

where τ is the 95th percentile of the ensemble of elements of $\vec{\pi}_{1:87}$, corresponding to the 87 proteins in the training set. We then greedily identified the mapping that maximizing the accuracy of the predicted annotations of 87 proteins. We used this mapping to compare predicted versus known functional annotations for all proteins; in Figure 11 we plot marginal frequencies of proteins' membership to true functions (left panel) and to predicted functions (right panel). The accuracy on the 90% testing set is about 87%. An algorithm that randomly guesses annotations, knowing the right proportions of annotations in each category, leads to a baseline accuracy of about 70%. Figure 12 shows predicted mixed memberships (dashed, red lines) versus the true annotations (solid, black lines), given the estimated mapping of blocks to functions, for six example proteins.

4.3.3 INDIRECT EVALUATION: FUNCTIONAL CONTENT OF PREDICTED INTERACTIONS

In the second experiment, we selected the mixed membership blockmodel best for predicting out-of-sample interactions, and we explored its goodness-of-fit indirectly, in terms of the number of

predicted interactions that are functionally relevant according to GO present in estimated protein interaction networks obtained with the two types of analyses that MMB supports; summarization and de-noising.

We fit models with K ranging between 2 and 255. We selected the best model ($K = 50$) using cross-validated held-out log likelihood, as in Figure 7. This finding supports the hypothesis that proteins derived from the MIPS data are interpretable in terms functional biological contexts. Alternatively, the blocks might encode signal at a finer resolution, such as that of protein complexes.

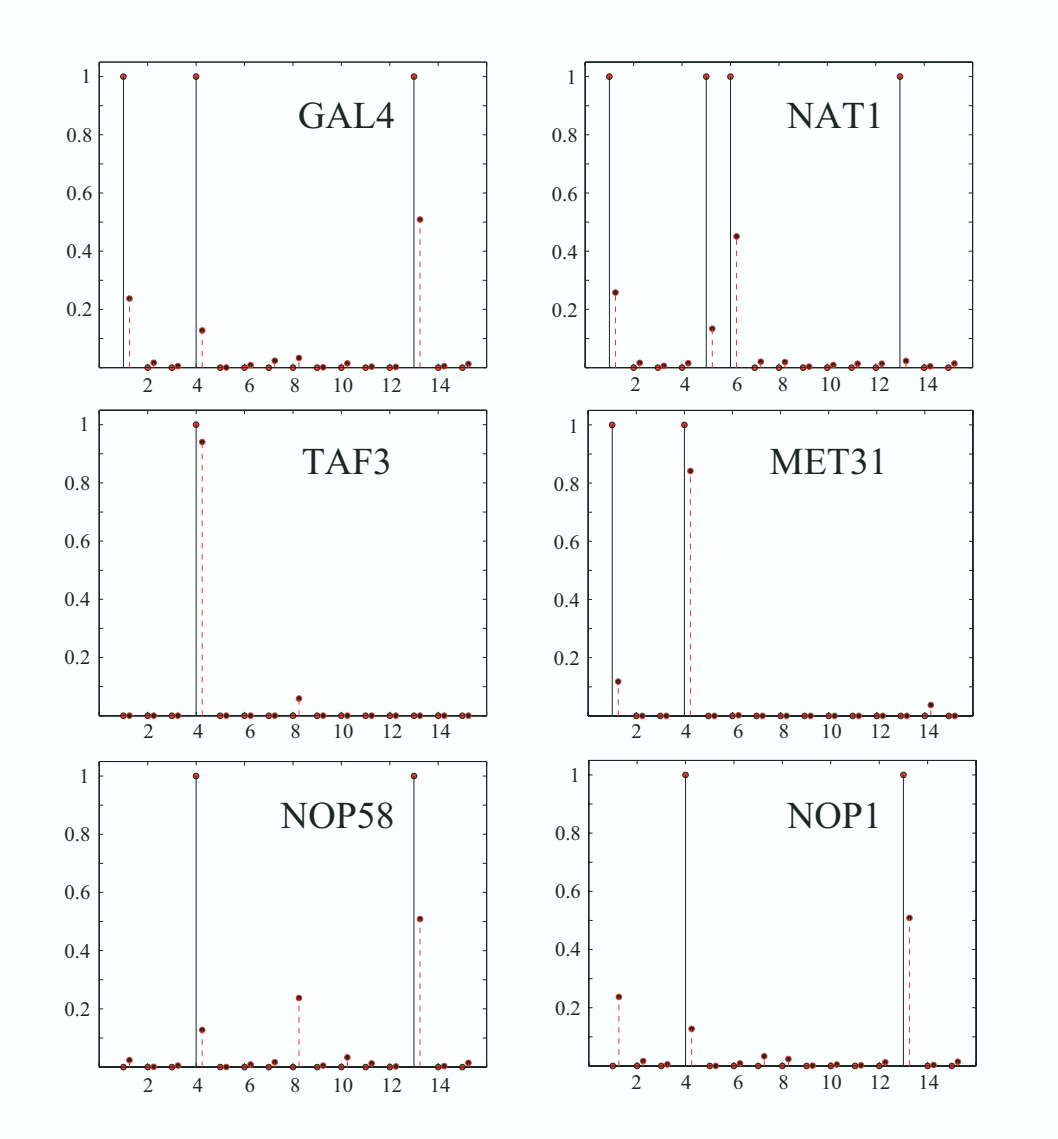

Figure 12: Predicted mixed-memberships (dashed, red lines) versus binary manually curated functional annotations (solid, black lines) for six example proteins, given the estimated mapping of blocks to functions in Figure 11.

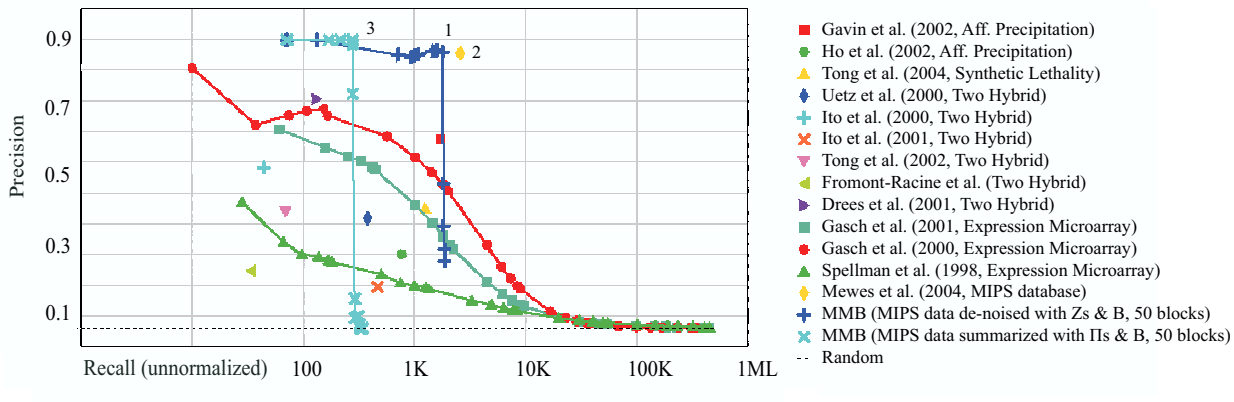

Figure 13: In the top panel we measure the functional content of the the MIPS collection of protein interactions (yellow diamond), and compare it against other published collections of interactions and microarray data, and to the posterior estimates of the MMB models—computed as described in Section 4.3.3. A breakdown of three estimated interaction networks (the points annotated 1, 2, and 3) into most represented gene ontology categories is detailed in Table 3.

If that was the case, however, we would expect the optimal number of blocks to be significantly higher; $871/5 \approx 175$, given an average size of five proteins in a complex (Krogan et al., 2006).

Using this model, we computed posterior model-based expectations of each interaction as follows,

$$\mathbb{E}\,[\,Y(p,q)\,] \approx \widehat{\vec{\pi}}_p{}' \widehat{B}\, \widehat{\vec{\pi}}_q \qquad \text{and} \qquad \mathbb{E}\,[\,Y(p,q)\,] \approx \widehat{\vec{\phi}}_{p\to q}{}' \widehat{B}\, \widehat{\vec{\phi}}_{p\leftarrow q}.$$

These computations lead to two estimated protein interaction networks with expected probabilities of interactions taking values in $[0,1]$. We obtained binary protein interaction networks by thresholding these expected probabilities at ten different values. In terms of the two analyses described in Section 2.2, this amount to either (i)predicting physical interactions by thresholding the posterior expectations computed using blockmodel B and mixed membership map $\vec{\pi}s$, essentially a prediction task, or (ii) we de-noise the observed interactions Y using the blockmodel B and interaction-specific membership indicators Zs, essentially a de-noising task. We use the independent set of functional annotations from the gene ontology to decide which interactions are functionally meaningful; namely those between pairs of proteins that share at least one functional annotation (Myers et al., 2006). In this sense, between two models that suggest different sets of interactions as reliable, our evaluation assigns a higher score to the model that reveals *functionally relevant* interactions. Figure 13 shows the functional content of the original MIPS collection of physical interactions (point no.2), and of the collections of interactions computed using $(B,\Pi s)$, the light blue $(-\times)$ line, and using (B,Zs), the dark blue $(-+)$ line, thresholded at ten different levels—precision-recall curves. The posterior means of Πs provide a parsimonious representation for the MIPS collection, and lead to precise interaction estimates, in moderate amount ($-\times$ line). The posterior means of Zs provide a richer representation for the data, and describe most of the functional content of the MIPS collection with high precision ($-+$ line). Figure 13 also shows the functional content of the original MIPS collection (the yellow diamond). Most importantly, notice the estimated protein interaction

networks, that is, ex-es and crosses, corresponding to lower levels of recall feature a more precise functional content than the original. This means that the proposed latent block structure is helpful in summarizing the collection of interactions—by ranking them properly. On closer inspection, dense blocks of predicted interactions contain known functional predictions that were not in the MIPS collection, thus effectively improving the quality of the data that instantiate activity specific to few biological contexts, such as biopolymer catabolism and homeostasis. In conclusion, results suggest that MMB successfully reduces the dimensionality of the data, while revealing substantive information about the multiple functionality of proteins that can be used to inform subsequent analyses.

Table 3 provides more information about three instances of predicted interaction networks displayed in Figure 13; those corresponding the points annotated 1, 2, and 3. Specifically, the table shows a breakdown of the predicted (posterior) collections of interactions in each example network into the gene ontology categories. A count in the table corresponds to the fact that both proteins are annotated with the same GO functional category.[5]

In this application, the MMB learned information about (i) the mixed membership of objects to latent groups, and (ii) the connectivity patterns among latent groups. These estimates were useful in describing and summarizing the functional content of the MIPS collection of protein interactions. This suggests the use of MMB as a dimensionality reduction approach that may be useful for performing model-driven de-noising of new collections of interactions, such as those measured via high-throughput experiments.

5. Discussion

Modern probabilistic models for relational data analysis are rooted in the stochastic blockmodels for psychometric and sociological analysis, pioneered by Lorrain and White (1971) and by Holland and Leinhardt (1975). In statistics, this line of research has been extended in various contexts over the years (Fienberg et al., 1985; Wasserman and Pattison, 1996; Snijders, 2002; Hoff et al., 2002; Doreian et al., 2004). In machine learning, the related technique of Markov random networks (Frank and Strauss, 1986) have been used for link prediction (Taskar et al., 2003) and the traditional blockmodels have been extended to include nonparametric Bayesian priors (Kemp et al., 2004, 2006; Xu et al., 2006) and to integrate relations and text (McCallum et al., 2007).

There is a close relationship between the MMB and the latent space models (Hoff et al., 2002; Handcock et al., 2007). In the latent space models, the latent vectors are drawn from Gaussian distributions and the interaction data is drawn from a Gaussian with mean $\vec{\pi}_p{}' \mathbb{I} \vec{\pi}_q$. In the MMB, the marginal probability of an interaction takes a similar form, $\vec{\pi}_p{}' B \vec{\pi}_q$, where B is the matrix of probabilities of interactions for each pair of latent groups. Two major differences exist between these approaches. In MMB, the distribution over the latent vectors is a Dirichlet and the underlying data distribution is arbitrary—we have chosen Bernoulli. The posterior inference in latent space models (Hoff et al., 2002; Handcock et al., 2007) is carried out via MCMC sampling, while we have developed a scalable variational inference algorithm to analyze large network structures. (It would be interesting to develop a variational algorithm for the latent space models as well.) A number of well-designed numerical investigations and comparisons between variational EM and variants of MCMC have been performed in existing literature; for instance, see Buntine and Jakulin (2006),[6]

5. Note that, in GO, proteins are typically annotated to multiple functional categories.
6. See corresponding slides with additional results. (`http://www.hiit.fi/~buntine/dpca\_slides.pdf`)

#	GO Term	Description	Pred.	Tot.
1	GO:0043285	Biopolymer catabolism	561	17020
1	GO:0006366	Transcription from RNA polymerase II promoter	341	36046
1	GO:0006412	Protein biosynthesis	281	299925
1	GO:0006260	DNA replication	196	5253
1	GO:0006461	Protein complex assembly	191	11175
1	GO:0016568	Chromatin modification	172	15400
1	GO:0006473	Protein amino acid acetylation	91	666
1	GO:0006360	Transcription from RNA polymerase I promoter	78	378
1	GO:0042592	Homeostasis	78	5778
2	GO:0043285	Biopolymer catabolism	631	17020
2	GO:0006366	Transcription from RNA polymerase II promoter	414	36046
2	GO:0016568	Chromatin modification	229	15400
2	GO:0006260	DNA replication	226	5253
2	GO:0006412	Protein biosynthesis	225	299925
2	GO:0045045	Secretory pathway	151	18915
2	GO:0006793	Phosphorus metabolism	134	17391
2	GO:0048193	Golgi vesicle transport	128	9180
2	GO:0006352	Transcription initiation	121	1540
3	GO:0006412	Protein biosynthesis	277	299925
3	GO:0006461	Protein complex assembly	190	11175
3	GO:0009889	Regulation of biosynthesis	28	990
3	GO:0051246	Regulation of protein metabolism	28	903
3	GO:0007046	Ribosome biogenesis	10	21528
3	GO:0006512	Ubiquitin cycle	3	2211

Table 3: Breakdown of three example interaction networks into most represented gene ontology categories—see text for more details. The digit in the first column indicates the example network in Figure 13 that any given line refers to. The last two columns quote the number of predicted, and possible pairs for each GO term.

and Braun and McAuliffe (2007). We refer readers interested in the comparison between variational vs. MCMC to these resources.

The model decouples the observed connectivity patterns into two sources of variability, B, Πs, that are apparently in competition for explaining the data, possibly raising an identifiability issue. This is not the case, however, as the blockmodel B captures global/asymmetric relations, while the mixed membership vectors Πs capture local/symmetric relations. This difference practically eliminates the issue, unless there is no signal in the data to begin with.

A recurring question, which bears relevance to mixed membership models in general, is why we do not integrate out the single membership indicators—$(\vec{z}_{p\to q}, \vec{z}_{p\leftarrow q})$. While this may lead to computational efficiencies we would often lose interpretable quantities that are useful for making predictions, for de-noising new measurements, or for performing other tasks. In fact, the posterior distributions of such quantities typically carry substantive information about elements of the appli-

cation at hand. In the application to protein interaction networks of Section 4.3, for example, they encode the interaction-specific memberships of individual proteins to protein complexes.

In the relational setting, cross-validation is feasible if the blockmodel estimated on training data can be expected to hold on test data; for this to happen the network must be of reasonable size, so that we can expect members of each block to be in both training and test sets. In this setting, scheduling of variational updates is important; nested variational scheduling leads to efficient and parallelizable inference.

A limitation of our model can be best appreciated in a simulation setting. If we consider structural properties of the network MMB is capable of generating, we count a wide array of local and global connectivity patterns. But the model does not readily generate *hubs*, that is, nodes connected with a large number of directed or undirected connections, or networks with skewed degree distributions.

From a data analysis perspective, we speculate that the value of MMB in capturing substantive information about a problem will increase in semi-supervised setting—where, for example, information about the membership of genes to functional contexts is included in the form of prior distributions. In such a setting we may be interested in looking at the change between prior and posterior membership; a sharp change may signal biological phenomena worth investigating. We need not assume that the number of groups/blocks, K, is finite. It is possible, for example, to posit that the mixed-membership vectors are sampled form a stochastic process, in the nonparametric setting. To maintain mixed membership of nodes to groups/blocks in such setting, we need to sample them from a hierarchical Dirichlet process (Teh et al., 2006), rather than from a Dirichlet Process (Escobar and West, 1995).

MMB generalizes to two important cases. First, multiple data collections $Y_{1:M}$ on the same objects can be generated by the same latent vectors. This might be useful, for example, for simultaneously analyzing the relational measurements about esteem and disesteem, liking and disliking, positive influence and negative influence, praise and blame, for example, see Sampson (1968), or those about the collection of 17 relations measured by Bradley (1987). Second, in the MMB the data generating distribution is a Bernoulli, but B can be a matrix that parameterizes any kind of distribution. For example, technologies for measuring interactions between pairs of proteins such as mass spectrometry (Ho et al., 2002) and tandem affinity purification (Gavin et al., 2002) return a probabilistic assessment about the presence of interactions, thus setting the range of $Y(p,q)$ to $[0,1]$. This is not the case for the manually curated collection of interactions we analyze in Section 4.3.

6. Conclusions

In this paper we introduced mixed membership stochastic blockmodels, a novel class of latent variable models for relational data. These models provide exploratory tools for scientific analyses in applications where the observations can be represented as a collection of unipartite graphs. The nested variational inference algorithm is parallelizable and allows fast approximate inference on large graphs.

Acknowledgments

This work was partially supported by National Institutes of Health under Grant No. R01 AG023141-01, by the Office of Naval Research under Contracts N00014-02-1-0973 and 175-6343, by the National Science Foundation under Grants No. DMS-0240019, IIS-0218466, IIS-0745520and DBI-0546594, by the Pennsylvania Department of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739, and by the Department of Defense, all to Carnegie Mellon University. The authors would like to thank David Banks and Jim Berger at Duke University, Alan Karr at the National Institute of Statistical Sciences for insight and advice, and acknowledge generous support from the Statistical and Applied Mathematical Sciences Institute.

Appendix A. General Model Formulation

In general, mixed membership stochastic blockmodels can be specified in terms of assumptions at four levels: population, node, latent variable, and sampling scheme level.

A.1 Population Level

Assume that there are K classes or sub-populations in the population of interest. We denote by $f\ (\ Y(p,q) \mid B(g,h)\ )$ the probability distribution of the relation measured on the pair of nodes (p,q), where the *p-th* node is in the *h-th* sub-population, the *q-th* node is in the *h-th* sub-population, and $B(g,h)$ contains the relevant parameters. The indices i, j run in $1,\ldots,N$, and the indices g,h run in $1,\ldots,K$.

A.2 Node Level

The components of the membership vector $\vec{\pi}_p = [\vec{\pi}_p(1),\ldots,\vec{\pi}_p(k)]'$ encodes the mixed membership of the *n-th* node to the various sub-populations. The distribution of the observed response $Y(p,q)$ given the relevant, node-specific memberships, $(\vec{\pi}_p,\vec{\pi}_q)$, is then

$$Pr\ (\ Y(p,q) \mid \vec{\pi}_p,\vec{\pi}_q,B\ )\ \ =\ \ \sum_{g,h=1}^{K} \vec{\pi}_p(g)\ f(Y(p,q) \mid B(g,h))\ \vec{\pi}_q(h).$$

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

A.3 Latent Variable Level

Assume that the mixed membership vectors $\vec{\pi}_{1:N}$ are realizations of a latent variable with distribution $D_{\vec{\alpha}}$, with parameter vector $\vec{\alpha}$. The probability of observing $Y(p,q)$, given the parameters, is then

$$Pr\ (\ Y(p,q) \mid \vec{\alpha},B\ )\ \ =\ \ \int Pr\ (\ Y(p,q) \mid \vec{\pi}_p,\vec{\pi}_q,B\ )\ D_{\vec{\alpha}}(d\vec{\pi}).$$

A.4 Sampling Scheme Level

Assume that the M independent replications of the relations measured on the population of nodes are independent of one another. The probability of observing the whole collection of graphs, $Y_{1:M}$, given the parameters, is then given by the following equation.

$$Pr\ (\ Y_{1:M} \mid \vec{\alpha},B\ ) = \prod_{m=1}^{M} \prod_{p,q=1}^{N} Pr\ (\ Y_m(p,q) \mid \vec{\alpha},B\ )\ .$$

Full model specifications immediately adapt to the different kinds of data, for example, 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 a distribution for the πs, D_α.

Appendix B. Details of the Variational Approximation

Here we present more details about the derivation of the variational EM algorithm presented in Section 3. Furthermore, we address a setting where M replicates are available about the paired measurements, $G_{1:M} = (N, Y_{1:M})$, and relations $Y_m(p,q)$ take values into an arbitrary metric space according to $f\ (\ Y_m(p,q) \mid .. \ )$. An extension of the inference algorithm to address the case or multivariate relations, say J-dimensional, and multiple blockmodels $B_{1:J}$ each corresponding to a distinct relational response, can be derived with minor modifications of the derivations that follow.

B.1 Variational Expectation-Maximization

We begin by briefly summarizing the general strategy we intend to use. The approximate variant of EM we describe here is often referred to as *Variational EM* (Beal and Ghahramani, 2003). Recall that Y denotes the data. Rewrite $X = (\vec{\pi}_{1:N}, Z_{\rightarrow}, Z_{\leftarrow})$ for the latent variables, and $\Theta = (\vec{\alpha}, B)$ for the model's parameters. Briefly, it is possible to lower bound the likelihood, $p(Y|\Theta)$, making use of Jensen's inequality and of any distribution on the latent variables $q(X)$,

$$\begin{aligned} p(Y|\Theta) &= \log \int_{\mathcal{X}} p(Y,X|\Theta)\, dX \\ &= \log \int_{\mathcal{X}} q(X)\, \frac{p(Y,X|\Theta)}{q(X)}\, dX \qquad \text{(for any } q\text{)} \\ &\geq \int_{\mathcal{X}} q(X)\, \log \frac{p(Y,X|\Theta)}{q(X)}\, dX \qquad \text{(Jensen's)} \\ &= \mathbb{E}_q\, [\, \log p(Y,X|\Theta) - \log q(X)\,] \ =:\ \mathcal{L}(q,\Theta) \end{aligned}$$

In EM, the lower bound $\mathcal{L}(q,\Theta)$ is then iteratively maximized with respect to Θ, in the M step, and q in the E step (Dempster et al., 1977). In particular, at the *t-th* iteration of the E step we set

$$q^{(t)} = p(X|Y, \Theta^{(t-1)}), \tag{5}$$

that is, equal to the posterior distribution of the latent variables given the data and the estimates of the parameters at the previous iteration.

Unfortunately, we cannot compute the posterior in Equation 5 for the admixture of latent blocks model. Rather, we define a direct parametric approximation to it, $\tilde{q} = q_\Delta(X)$, which involves an extra set of *variational parameters*, Δ, and entails an approximate lower bound for the likelihood $\mathcal{L}_\Delta(q,\Theta)$. At the *t-th* iteration of the E step, we then minimize the Kullback-Leibler divergence between $q^{(t)}$ and $q^{(t)}_\Delta$, with respect to Δ, using the data.[7] The optimal parametric approximation is, in fact, a proper posterior as it depends on the data Y, although indirectly, $q^{(t)} \approx q^{(t)}_{\Delta^*(Y)}(X) = p(X|Y)$.

B.2 Lower Bound for the Likelihood

According to the mean-field theory (Jordan et al., 1999), one can approximate an intractable distribution such as the one defined by Equation (1) by a fully factored distribution $q(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M})$

7. This is equivalent to maximizing the approximate lower bound for the likelihood, $\mathcal{L}_\Delta(q,\Theta)$, with respect to Δ.

defined as follows:

$$\begin{aligned} & q(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M} | \vec{\gamma}_{1:N}, \Phi^{\rightarrow}_{1:M}, \Phi^{\leftarrow}_{1:M}) \\ = & \prod_p q_1(\vec{\pi}_p|\vec{\gamma}_p) \prod_m \prod_{p,q} \Big(q_2(\vec{z}^m_{p\rightarrow q}|\vec{\phi}^m_{p\rightarrow q}, 1)\; q_2(\vec{z}^m_{p\leftarrow q}|\vec{\phi}^m_{p\leftarrow q}, 1) \Big), \end{aligned}$$

where q_1 is a Dirichlet, q_2 is a multinomial, and $\Delta = (\vec{\gamma}_{1:N}, \Phi^{\rightarrow}_{1:M}, \Phi^{\leftarrow}_{1:M})$ represent the set of free *variational parameters* need to be estimated in the approximate distribution.

Minimizing the Kulback-Leibler divergence between this $q(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M}|\Delta)$ and the original $p(\vec{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M}$ defined by Equation (1) leads to the following approximate lower bound for the likelihood.

$$\begin{aligned} \mathcal{L}_\Delta(q,\Theta) &= \mathbb{E}_q\left[\log \prod_m \prod_{p,q} p_1(Y_m(p,q)|\vec{z}^m_{p\rightarrow q}, \vec{z}^m_{p\leftarrow q}, B)\right] \\ &+ \mathbb{E}_q\left[\log \prod_m \prod_{p,q} p_2(\vec{z}^m_{p\rightarrow q}|\vec{\pi}_p, 1)\right] + \mathbb{E}_q\left[\log \prod_m \prod_{p,q} p_2(\vec{z}^m_{p\leftarrow q}|\vec{\pi}_q, 1)\right] \\ &+ \mathbb{E}_q\left[\log \prod_p p_3(\vec{\pi}_p|\vec{\alpha})\right] - \mathbb{E}_q\left[\prod_p q_1(\vec{\pi}_p|\vec{\gamma}_p)\right] \\ &- \mathbb{E}_q\left[\log \prod_m \prod_{p,q} q_2(\vec{z}^m_{p\rightarrow q}|\vec{\phi}^m_{p\rightarrow q}, 1)\right] - \mathbb{E}_q\left[\log \prod_m \prod_{p,q} q_2(\vec{z}^m_{p\leftarrow q}|\vec{\phi}^m_{p\leftarrow q}, 1)\right]. \end{aligned}$$

Working on the single expectations leads to

$$\begin{aligned} \mathcal{L}_\Delta(q,\Theta) &= \sum_m \sum_{p,q} \sum_{g,h} \phi^m_{p\rightarrow q,g} \phi^m_{p\leftarrow q,h} \cdot f\left(Y_m(p,q), B(g,h)\right) \\ &+ \sum_m \sum_{p,q} \sum_g \phi^m_{p\rightarrow q,g}\left[\psi(\gamma_{p,g}) - \psi(\sum_g \gamma_{p,g})\right] \\ &+ \sum_m \sum_{p,q} \sum_h \phi^m_{p\leftarrow q,h}\left[\psi(\gamma_{p,h}) - \psi(\sum_h \gamma_{p,h})\right] \\ &+ \sum_p \log\Gamma(\sum_k \alpha_k) - \sum_{p,k} \log\Gamma(\alpha_k) + \sum_{p,k} (\alpha_k - 1)\left[\psi(\gamma_{p,k}) - \psi(\sum_k \gamma_{p,k})\right] \\ &- \sum_p \log\Gamma(\sum_k \gamma_{p,k}) + \sum_{p,k} \log\Gamma(\gamma_{p,k}) - \sum_{p,k} (\gamma_{p,k} - 1)\left[\psi(\gamma_{p,k}) - \psi(\sum_k \gamma_{p,k})\right] \\ &- \sum_m \sum_{p,q} \sum_g \phi^m_{p\rightarrow q,g} \log \phi^m_{p\rightarrow q,g} - \sum_m \sum_{p,q} \sum_h \phi^m_{p\leftarrow q,h} \log \phi^m_{p\leftarrow q,h} \end{aligned}$$

where

$$f\left(Y_m(p,q), B(g,h)\right) = Y_m(p,q) \log B(g,h) + \left(1 - Y_m(p,q)\right) \log\left(1 - B(g,h)\right);$$

m runs over $1,\ldots,M$; p,q run over $1,\ldots,N$; g,h,k run over $1,\ldots,K$; and $\psi(x)$ is the derivative of the log-gamma function, $\frac{d\log\Gamma(x)}{dx}$.

B.3 The Expected Value of the Log of a Dirichlet Random Vector

The computation of the lower bound for the likelihood requires us to evaluate $\mathbb{E}_q\left[\log \vec{\pi}_p\right]$ for $p = 1,\ldots,N$. Recall that the density of an exponential family distribution with natural parameter $\vec{\theta}$ can be written as

$$\begin{aligned} p(x|\alpha) &= h(x) \cdot c(\alpha) \cdot \exp\left\{\sum_k \theta_k(\alpha) \cdot t_k(x)\right\} \\ &= h(x) \cdot \exp\left\{\sum_k \theta_k(\alpha) \cdot t_k(x) - \log c(\alpha)\right\}. \end{aligned}$$

Omitting the node index p for convenience, we can rewrite the density of the Dirichlet distribution p_3 as an exponential family distribution,

$$p_3(\vec{\pi}|\vec{\alpha}) = \exp\left\{ \sum_k (\alpha_k - 1)\log(\pi_k) - \log\frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \right\},$$

with natural parameters $\theta_k(\vec{\alpha}) = (\alpha_k - 1)$ and natural sufficient statistics $t_k(\vec{\pi}) = \log(\pi_k)$. Let $c'(\vec{\theta}) = c(\alpha_1(\vec{\theta}), \ldots, \alpha_K(\vec{\theta}))$; using a well known property of the exponential family distributions (Schervish, 1995) we find that

$$\mathbb{E}_q\left[\log\pi_k\right] \;=\; \mathbb{E}_{\vec{\theta}}\left[\log t_k(x)\right] = \psi\left(\alpha_k\right) - \psi\left(\sum_k \alpha_k\right),$$

where $\psi(x)$ is the derivative of the log-gamma function, $\frac{d\log\Gamma(x)}{dx}$.

B.4 Variational E Step

The approximate lower bound for the likelihood $\mathcal{L}_\Delta(q,\Theta)$ can be maximized using exponential family arguments and coordinate ascent (Wainwright and Jordan, 2003).

Isolating terms containing $\phi^m_{p\to q,g}$ and $\phi^m_{p\leftarrow q,h}$ we obtain $\mathcal{L}_{\phi^m_{p\to q,g}}(q,\Theta)$ and $\mathcal{L}_{\phi^m_{p\to q,g}}(q,\Theta)$. The natural parameters $\vec{g}^m_{p\to q}$ and $\vec{g}^m_{p\leftarrow q}$ corresponding to the natural sufficient statistics $\log(\vec{z}^m_{p\to q})$ and $\log(\vec{z}^m_{p\leftarrow q})$ are functions of the other latent variables and the observations. We find that

$$\begin{aligned} g^m_{p\to q,g} &= \log\pi_{p,g} + \sum_h z^m_{p\leftarrow q,h}\cdot f\left(Y_m(p,q), B(g,h)\right), \\ g^m_{p\leftarrow q,h} &= \log\pi_{q,h} + \sum_g z^m_{p\to q,g}\cdot f\left(Y_m(p,q), B(g,h)\right), \end{aligned}$$

for all pairs of nodes (p,q) in the *m-th* network; where $g,h = 1,\ldots,K$, and

$$f\left(Y_m(p,q), B(g,h)\right) = Y_m(p,q)\log B(g,h) + \left(1 - Y_m(p,q)\right)\log\left(1 - B(g,h)\right).$$

This leads to the following updates for the variational parameters $(\vec{\phi}^m_{p\to q}, \vec{\phi}^m_{p\leftarrow q})$, for a pair of nodes (p,q) in the *m-th* network:

$$\begin{aligned} \hat{\phi}^m_{p\to q,g} &\propto e^{\,\mathbb{E}_q\left[g^m_{p\to q,g}\right]} \\ &= e^{\,\mathbb{E}_q\left[\log\pi_{p,g}\right]} \cdot e^{\,\sum_h \phi^m_{p\leftarrow q,h}\cdot\, \mathbb{E}_q\left[f\left(Y_m(p,q),B(g,h)\right)\right]} \\ &= e^{\,\mathbb{E}_q\left[\log\pi_{p,g}\right]} \cdot \prod_h \left(B(g,h)^{Y_m(p,q)} \cdot \left(1 - B(g,h)\right)^{1 - Y_m(p,q)} \right)^{\phi^m_{p\leftarrow q,h}}, \\ \hat{\phi}^m_{p\leftarrow q,h} &\propto e^{\,\mathbb{E}_q\left[g^m_{p\leftarrow q,h}\right]} \\ &= e^{\,\mathbb{E}_q\left[\log\pi_{q,h}\right]} \cdot e^{\,\sum_g \phi^m_{p\to q,g}\cdot\, \mathbb{E}_q\left[f\left(Y_m(p,q),B(g,h)\right)\right]} \\ &= e^{\,\mathbb{E}_q\left[\log\pi_{q,h}\right]} \cdot \prod_g \left(B(g,h)^{Y_m(p,q)} \cdot \left(1 - B(g,h)\right)^{1 - Y_m(p,q)} \right)^{\phi^m_{p\to q,g}}, \end{aligned}$$

for $g,h = 1,\ldots,K$. These estimates of the parameters underlying the distribution of the nodes' group indicators $\vec{\phi}^m_{p\to q}$ and $\vec{\phi}^m_{p\leftarrow q}$ need be normalized, to make sure $\sum_k \phi^m_{p\to q,k} = \sum_k \phi^m_{p\leftarrow q,k} = 1$.

Isolating terms containing $\gamma_{p,k}$ we obtain $\mathcal{L}_{\gamma_{p,k}}(q,\Theta)$. Setting $\frac{\partial \mathcal{L}_{\gamma_{p,k}}}{\partial \gamma_{p,k}}$ equal to zero and solving for $\gamma_{p,k}$ yields:

$$\hat{\gamma}_{p,k} = \alpha_k + \sum_m \sum_q \phi^m_{p\to q,k} + \sum_m \sum_q \phi^m_{p\leftarrow q,k},$$

for all nodes $p \in \mathcal{P}$ and $k = 1,\ldots,K$.

The *t-th* iteration of the variational E step is carried out for fixed values of $\Theta^{(t-1)} = (\vec{\alpha}^{(t-1)}, B^{(t-1)})$, and finds the optimal approximate lower bound for the likelihood $\mathcal{L}_{\Delta^*}(q,\Theta^{(t-1)})$.

B.5 Variational M Step

The optimal lower bound $\mathcal{L}_{\Delta^*}(q^{(t-1)},\Theta)$ provides a tractable surrogate for the likelihood at the *t-th* iteration of the variational M step. We derive empirical Bayes estimates for the hyper-parameters Θ that are based upon it.[8] That is, we maximize $\mathcal{L}_{\Delta^*}(q^{(t-1)},\Theta)$ with respect to Θ, given expected sufficient statistics computed using $\mathcal{L}_{\Delta^*}(q^{(t-1)},\Theta^{(t-1)})$.

Isolating terms containing $\vec{\alpha}$ we obtain $\mathcal{L}_{\vec{\alpha}}(q,\Theta)$. Unfortunately, a closed form solution for the approximate maximum likelihood estimate of $\vec{\alpha}$ does not exist (Blei et al., 2003). We can produce a Newton-Raphson method that is linear in time, where the gradient and Hessian for the bound $\mathcal{L}_{\vec{\alpha}}$ are

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_k} &= N\left(\psi\left(\sum_k \alpha_k\right) - \psi(\alpha_k)\right) + \sum_p \left(\psi(\gamma_{p,k}) - \psi\left(\sum_k \gamma_{p,k}\right)\right), \\
\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_{k_1}\alpha_{k_2}} &= N\left(\mathbb{I}_{(k_1=k_2)}\cdot\psi'(\alpha_{k_1}) - \psi'\left(\sum_k \alpha_k\right)\right).
\end{aligned}$$

Isolating terms containing B we obtain $\mathcal{L}_B$, whose approximate maximum is

$$\hat{B}(g,h) = \frac{1}{M}\sum_m \left(\frac{\sum_{p,q} Y_m(p,q)\cdot \phi^m_{p\to qg}\phi^m_{p\leftarrow qh}}{(1-\rho)\cdot\sum_{p,q}\phi^m_{p\to qg}\phi^m_{p\leftarrow qh}}\right),$$

for every index pair $(g,h) \in [1,K]\times[1,K]$.

In Section 2.1 we introduced an extra parameter, ρ, to control the relative importance of presence and absence of interactions in likelihood, that is, the score that informs inference and estimation. Isolating terms containing ρ we obtain $\mathcal{L}_\rho$. We may then estimate the sparsity parameter ρ by

$$\hat{\rho} = \frac{1}{M}\sum_m \left(\frac{\sum_{p,q}\left(1 - Y_m(p,q)\right)\cdot\left(\sum_{g,h}\phi^m_{p\to qg}\phi^m_{p\leftarrow qh}\right)}{\sum_{p,q}\sum_{g,h}\phi^m_{p\to qg}\phi^m_{p\leftarrow qh}}\right).$$

Alternatively, we can fix ρ prior to the analysis; the density of the interaction matrix is estimated with $\hat{d} = \sum_{m,p,q} Y_m(p,q)/(N^2M)$, and the sparsity parameter is set to $\tilde{\rho} = (1-\hat{d})$. This latter estimator attributes all the information in the non-interactions to the point mass, that is, to latent sources other than the block model B or the mixed membership vectors $\vec{\pi}_{1:N}$. It does, however, provide a quick recipe to reduce the computational burden during exploratory analyses.[9]

8. We could term these estimates *pseudo* empirical Bayes estimates, since they maximize an approximate lower bound for the likelihood, $\mathcal{L}_{\Delta^*}$.

9. Note that $\tilde{\rho} = \hat{\rho}$ in the case of single membership. In fact, that implies $\phi^m_{p\to qg} = \phi^m_{p\leftarrow qh} = 1$ for some (g,h) pair, for any (p,q) pair.

References

E. M. Airoldi. Getting started in probabilistic graphical models. *PLoS Computational Biology*, 3 (12):e252, 2007.

E. M. Airoldi, D. M. Blei, E. P. Xing, and S. E. Fienberg. A latent mixed-membership model for relational data. In *ACM SIGKDD Workshop on Link Discovery: Issues, Approaches and Applications*, 2005.

E. M. Airoldi, S. E. Fienberg, and E. P. Xing. Mixed membership analysis of expression studies—attribute data. Manuscript, 2007. URL `http://arxiv.org/abs/0711.2520/`.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland, 4th edition, 2002.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubinand, and G. Sherlock. Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nature Genetics*, 25(1):25–29, 2000.

M. J. Beal and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics*, volume 7, pages 453–464. Oxford University Press, 2003.

L. Berkman, B. H. Singer, and K. Manton. Black/white differences in health status and mortality among the elderly. *Demography*, 26(4):661–678, 1989.

C. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 777–784. MIT Press, Cambridge, MA, 2003.

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

R. T. Bradley. *Charisma and Social Structure*. Paragon House, 1987.

M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. Manuscript, 2007. URL `http://arxiv.org/abs/0712.2526/`.

R. L. Breiger, S. A. Boorman, and P. Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison to multidimensional scaling. *Journal of Mathematical Psychology*, 12:328–383, 1975.

W. L. Buntine and A. Jakulin. Discrete components analysis. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag, 2006. URL `http://arxiv.org/abs/math.ST/0604410/`.

G. B. Davis and K. M. Carley. Clearing the FOG: Fuzzy, overlapping groups for social networks. Manuscript, 2006.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized Blockmodeling*. Cambridge University Press, 2004.

P. Doreian, V. Batagelj, and A. Ferligoj. Discussion of "Model-based clustering for social networks". *Journal of the Royal Statistical Society, Series A*, 170, 2007.

E. A. Erosheva. *Grade of Membership and Latent Structure Models with Application to Disability Survey Data*. PhD thesis, Carnegie Mellon University, Department of Statistics, 2002.

E. A. Erosheva and S. E. Fienberg. Bayesian mixed membership models for soft clustering and classification. In C. Weihs and W. Gaul, editors, *Classification—The Ubiquitous Challenge*, pages 11–26. Springer-Verlag, 2005.

M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

S. E. Fienberg, M. M. Meyer, and S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80:51–67, 1985.

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

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

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A*, 170:1–22, 2007.

K. M. Harris, F. Florey, J. Tabor, P. S. Bearman, J. Jones, and R. J. Udry. The national longitudinal study of adolescent health: research design. Technical report, Caorlina Population Center, University of North Carolina, Chapel Hill, 2003.

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

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

P. W. Holland and S. Leinhardt. Local structure in social networks. In D. Heise, editor, *Sociological Methodology*, pages 1–45. Jossey-Bass, 1975.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

C. Joutard, E. M. Airoldi, S. E. Fienberg, and T. M. Love. Discovery of latent patterns with hierarchical bayesian mixed-membership models and the issue of model choice. In *Data Mining Patterns, New Methods and Applications*, 2007. Forthcoming.

C. Kemp, T. L. Griffiths, and J. B. Tenenbaum. Discovering latent classes in relational data. Technical Report AI Memo 2004-019, MIT, 2004.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

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

F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *IEEE Computer Vision and Pattern Recognition*, 2005.

F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1:49–80, 1971.

A. McCallum, X. Wang, and N. Mohanty. Joint group and topic discovery from relations and text. In *Statistical Network Analysis: Models, Issues and New Directions*, Lecture Notes in Computer Science. Springer-Verlag, 2007.

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

T. Minka. Estimating a Dirichlet distribution. Manuscript, 2003.

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

C. L. Myers, D. A. Barret, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya. Finding function: An evaluation framework for functional genomics. *BMC Genomics*, 7(187), 2006.

J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American Journal of Human Genetics*, 67:170–181, 2000.

F. S. Sampson. *A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships*. PhD thesis, Cornell University, 1968.

Mark J. Schervish. *Theory of Statistics*. Springer, 1995.

T. A. B. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 2002.

T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Neural Information Processing Systems 15*, 2003.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Y. W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

R. J. Udry. The national longitudinal study of adolescent health: (add health) waves i and ii, 1994–1996; wave iii 2001–2002. Technical report, Caorlina Population Center, University of North Carolina, Chapel Hill, 2003.

C. T. Volinsky and A. E. Raftery. Bayesian information criterion for censored survival models. *Biometrics*, 56:256–262, 2000.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.

Y. J. Wang and G. Y. Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82:8–19, 1987.

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

E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*, volume 19, 2003.

Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel. Infinite hidden relational models. In *Uncertainty in Artificial Intelligence*, 2006.