Mixed Membership Stochastic Blockmodels

Edoardo M. Airoldi [1,2], **David M. Blei** [1], **Stephen E. Fienberg** [3,4] **& Eric P. Xing** [4*]
[1] Department of Computer Science, [2] Lewis-Sigler Institute, Princeton University
[3] Department of Statistics, [4] School of Computer Science, Carnegie Mellon University
eairoldi@Princeton.EDU

Abstract

In many settings, such as protein interactions and gene regulatory networks, collections of author-recipient email, and social networks, the data consist of pairwise measurements, e.g., presence or absence of links between pairs of objects. Analyzing such data with probabilistic models requires non-standard assumptions, since the usual independence or exchangeability assumptions no longer hold. In this paper, we introduce a class of latent variable models for pairwise measurements: mixed membership stochastic blockmodels. Models in this class combine a global model of dense patches of connectivity (blockmodel) with a local model to 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 blockmodel with applications to social networks and 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 citation, the Web connects pages by links, and protein-protein interaction data connect proteins by physical interaction records. In these settings, we often wish to infer hidden attributes of the objects from the pairwise observations. For example, we might want to compute a clustering of the web-pages, predict the functions of a protein, or assess the degree of relevance of a scientific abstract to a scholar's query. Unlike traditional attribute data measured over individual objects, *relational data* violate the classical independence or exchangeability assumptions made in machine learning and statistics. The objects are dependent by their very nature, and this interdependence suggests that a different set of assumptions is more appropriate.

Recently proposed models aim at resolving relational information into a collection of connectivity motifs. Such models are based on assumptions that often ignore useful technical necessities, or important empirical regularities. For instance, exponential random graph models [11] summarize the variability in a collection of paired measurements with a set of relational motifs, but do not provide a representation useful for making unit-specific predictions. Latent space models [4] project individual units of analysis into a low-dimensional latent space, but do not provide a group structure into such space useful for clustering. Stochastic blockmodels [8, 6] resolve paired measurements into groups and connectivity between pairs of groups, but constrain each unit to instantiate the connectivity patterns of a single group as observed in most applications. Mixed membership models, such as latent Dirichlet allocation [1], have emerged in recent years as a flexible modeling tool for data where the single group assumption is violated by the heterogeneity within a unit of analysis—e.g., a document, or a node in a graph. They have been successfully applied in many domains, such as document analysis [1], image processing [7], and population genetics [9]. Mixed membership models associate each unit of analysis with multiple groups rather than a single groups, via a membership

[*] A longer version of this work is available online, at http://jmlr.csail.mit.edu/papers/v9/airoldi08a.html

probability-like vector. The concurrent membership of a data in different groups can capture its different aspects, such as different underlying topics for words constituting each document. The mixed membership formalism is a particularly natural idea for relational data, where the objects can bear multiple latent roles or cluster-memberships that influence their relationships to others. Existing mixed membership models, however, are not appropriate for relational data because they assume that the data are conditionally independent given their latent membership vectors. Conditional independence assumptions that technically instantiate mixed membership in recent work, however, are inappropriate for the relational data settings. In such settings, an objects is described by its relationships to others. Thus assuming that the ensemble of mixed membership vectors help govern the relationships of each object would be more appropriate.

Here we develop mixed membership models for relational data and we describe a fast variational inference algorithm for inference and estimation. 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. We apply our model to protein interaction and social networks.

2 The Basic Mixed Membership Blockmodel

Observations consist of pairwise measurements, represented as a graph $G = (\mathcal{N}, Y)$, where $Y(p, q)$ denotes the measurement taken on the pair of nodes (p, q). In this section we consider observations consisting of a single binary matrix, where $Y(p, q) \in \{0, 1\}$, i.e., the data can be represented with a directed graph. The model generalizes to two important settings, however, as we discuss below—a collection of matrices and/or other types of measurements. We summarize a collection of pairwise measurements with a mapping from nodes to sets of nodes, called *blocks*, and pairwise relations among the blocks themselves. Intuitively, the inference process aims at identifying nodes that are similar to one another in terms of their connectivity to blocks of nodes. Similar nodes are mapped to the same block. Individual nodes are allowed to instantiate connectivity patterns of multiple blocks. Thus, the goal of the analysis with a Mixed Membership Blockmodel (MMB) is to identify (i) the *mixed membership* mapping of nodes, i.e., the units of analysis, to a fixed number of blocks, K, and (ii) the pairwise relations among the blocks. Pairwise measurements among N nodes are then generated according to latent distributions of block-membership for each node and a matrix of block-to-block interaction strength. Latent per-node distributions are specified by simplicial vectors. Each node is associated with a randomly drawn vector, say $\vec{\pi}_i$ for node i, where $\pi_{i,g}$ denotes the probability of node i belonging to group g. In this fractional sense, each node can belong to multiple groups with different degrees of membership. 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 connection from a node in group g to a node in group h. The indicator vector $\vec{z}_{p\to q}$ denotes the specific block membership of node p when it connects to node q, while $\vec{z}_{p\leftarrow q}$ denotes the specific block membership of node q when it is connected from node p. The complete generative process for a graph $G = (\mathcal{N}, Y)$ is as follows:

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

Note that the group membership of each node is *context dependent*, i.e., each node may assume different membership when interacting with 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 $\{\vec{z}_{p\leftarrow q} : p, q \in \mathcal{N}\} =: Z_{\leftarrow}$. Further, the pairs of group memberships that underlie interactions, e.g., $(\vec{z}_{p\to q}, \vec{z}_{p\leftarrow q})$ for $Y(p, q)$, need not be equal; this fact is useful for characterizing asymmetric interaction networks. Equality may be enforced when modeling symmetric interactions. The joint probability of the data Y and the latent variables $\{\vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow}\}$ sampled according to the MMB is:

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

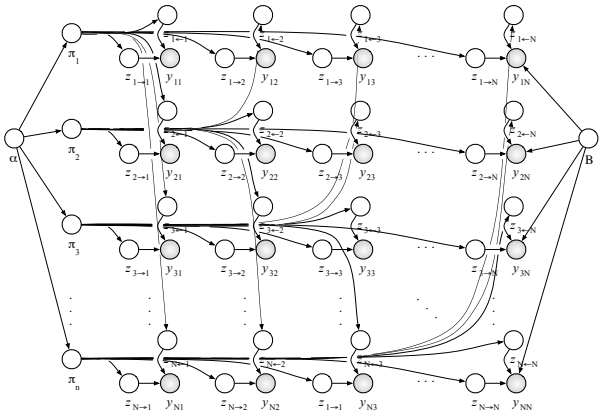

Figure 1: The graphical model of the mixed membership blockmodel (MMB). We did not draw all the arrows out of the block model B for clarity. All the pairwise measurements, $Y(p, q)$, depend on it.

Introducing Sparsity. Adjacency matrices encoding binary pairwise measurements often contain a large amount of zeros, or non-interactions; they are *sparse*. 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. In this setting, it is reasonable to expect that a large portion of the non-interactions is due to limited opportunities of contact between people in a large population, or by design of the questionnaire, 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 B's elements.

We introduce a sparsity parameter $\rho \in [0, 1]$ in the model above 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 $\{\vec{\alpha}, B, \vec{\pi}_{1:N}\}$.

Recall that $\{\vec{\alpha}, B\}$ are constant quantities to be estimated, while $\{\vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow}\}$ are unknown variable quantities whose posterior distribution needs to be determined. Below, we detail the variational expectation-maximization (EM) procedure to carry out approximate estimation and inference.

2.1 Variational E-Step

During the E-step, we update the posterior distribution over the unknown variable quantities $\{\vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow}\}$. The normalizing constant of the posterior is the marginal probability of the data, which requires an intractable integral over the simplicial vectors $\vec{\pi}_p$,

$$p(Y \mid \vec{\alpha}, B) = \int_{\vec{\pi}_{1:N}} \sum_{z_{p\leftarrow q}, z_{p\to q}} p(Y, \vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow} | \vec{\alpha}, B). \tag{1}$$

We appeal to mean-field variational methods [5] to approximate the posterior of interest. The main idea behind variational methods is to posit a simple distribution of the latent variables with free parameters, which are fit to make the approximation close in Kullback-Leibler divergence to the true posterior of interest. The log of the marginal probability in Equation 1 can be bound as follows,

$$\log p(Y \mid \alpha, B) \geq \mathbb{E}_q \left[\log p(Y, \vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow} | \alpha, B) \right] - \mathbb{E}_q \left[\log q(\vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow}) \right], \tag{2}$$

by introducing 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, $q(\vec{\pi}_{1:N}, Z_{\to}, Z_{\leftarrow} \mid \vec{\gamma}_{1:N}, \Phi_{\to}, \Phi_{\leftarrow})$, where $\{\vec{\gamma}_{1:N}, \Phi_{\to}, \Phi_{\leftarrow}\}$ is the set of free *variational parameters* that must be set to tighten the bound. We

tighten the bound with respect to the variational parameters, to minimize the KL divergence between q and the true posterior. The update for the variational multinomial parameters is

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

$$\hat{\phi}_{p\leftarrow q,h} \propto e^{\mathbb{E}_q\left[\log \pi_{q,h}\right]} \cdot \prod_g \left(B(g,h)^{Y(p,q)} \cdot \left(1 - B(g,h)\right)^{1-Y(p,q)} \right)^{\phi_{p\to q,g}}, \quad (4)$$

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

$$\hat{\gamma}_{p,k} = \alpha_k + \sum_q \phi_{p\to q,k} + \sum_q \phi_{p\leftarrow q,k}, \quad (5)$$

for all nodes $p = 1, \ldots, N$ and $k = 1, \ldots, K$.

Nested Variational Inference. To improve convergence, we developed a *nested variational inference* scheme based on an alternative schedule of updates to the traditional ordering [5]. In a naïve iteration scheme for variational inference, one initializes the variational Dirichlet parameters $\vec{\gamma}_{1:N}$ and the variational multinomial parameters $(\vec{\phi}_{p\to q}, \vec{\phi}_{p\leftarrow q})$ to non-informative values, and then iterates until convergence the following two steps: (i) update $\vec{\phi}_{p\to q}$ and $\phi_{p\leftarrow q}$ for all edges (p, q), and (ii) update $\vec{\gamma}_p$ for all nodes $p \in \mathcal{N}$. At each variational inference cycle one needs to allocate $NK + 2N^2K$ scalars. In our experiments, the naïve variational algorithm often failed to converge, or converged only after many iterations. We attribute this behavior to dependence between $\vec{\gamma}_{1:N}$ and B in the model, which is not accounted for by the naïve algorithm. The nested variational inference algorithm retains portion of this dependence across iterations by following a particular path to convergence. We keep the block of free parameters $(\vec{\phi}_{p\to q}, \vec{\phi}_{p\leftarrow q})$ at their optimal values conditionally on the other variational parameters. These parameters are involved in the updates of parameters in $\vec{\gamma}_{1:N}$ and in B, thus effectively providing a channel to maintain some dependence among them. From a computational perspective, the nested algorithm trades time for space thus allowing us to deal with large graphs. At each variational cycle we allocate $NK + 2K$ scalars only. The algorithm can be parallelized, and, empirically, leads to a better likelihood bound per unit of running time.

2.2 M-Step

During the M-step, we maximize the lower bound in Equation 2, used as a surrogate for the likelihood, with respect to the unknown constants $\{\vec{\alpha}, B\}$. In other words, we compute the empirical Bayes estimates of the hyper-parameters. The 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. We used linear-time Newton-Raphson, with gradient and Hessian

$$\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_k} = N\left(\psi\left(\sum_k \alpha_k \right) - \psi(\alpha_k) \right) + \sum_p \left(\psi(\gamma_{p,k}) - \psi\left(\sum_k \gamma_{p,k} \right) \right), \text{ and}$$

$$\frac{\partial \mathcal{L}_{\vec{\alpha}}}{\partial \alpha_{k_1} \alpha_{k_2}} = N\left(\mathbb{I}_{(k_1=k_2)} \cdot \psi'(\alpha_{k_1}) - \psi'\left(\sum_k \alpha_k \right) \right),$$

to find optimal values for $\vec{\alpha}$, numerically. The approximate MLE of B is

$$\hat{B}(g,h) = \frac{\sum_{p,q} Y(p,q) \cdot \phi_{p\to qg}\, \phi_{p\leftarrow qh}}{\sum_{p,q} \phi_{p\to qg}\, \phi_{p\leftarrow qh}}, \quad (6)$$

for every pair $(g, h) \in [1, K]^2$. Finally, the approximate MLE of the sparsity parameter ρ is

$$\hat{\rho} = \frac{\sum_{p,q} \left(1 - Y(p,q)\right) \cdot \left(\sum_{g,h} \phi_{p\to qg}\, \phi_{p\leftarrow qh} \right)}{\sum_{p,q} \sum_{g,h} \phi_{p\to qg}\, \phi_{p\leftarrow qh}}. \quad (7)$$

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, i.e., to latent sources other than the block model B or the mixed membership vectors $\vec{\pi}_{1:N}$. It can be used, however, as a quick recipe to reduce the computational burden during exploratory analyses.

Several model selection strategies exist for hierarchical models. In our setting, model selection translates into the choice of the number of blocks, K. Below, we chose K with held-out likelihood in a cross-validation experiment, on large networks, and with approximate BIC, on small networks.

2.3 Summarizing and De-Noising Pairwise Measurements

It is useful to consider two data analysis perspectives the MMB can offer: (i) it summarizes the data, Y, in terms of the global blockmodel, B, and the node-specific mixed memberships, Πs, (ii) it de-noises 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 reasons for including the set of latent variables Zs differ. 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 $\vec{\pi}_p{}' B\ \vec{\pi}_q$, and $\vec{\phi}_{p\to q}{}' B\ \vec{\phi}_{p\leftarrow q}$, in the two cases.

3 Empirical Results

We evaluated the MMB on simulated data and on three collections of pairwise measurements. Results on simulated data sampled accordingly to the model show that variational EM accurately recovers the mixed membership map, $\vec{\pi}_{1:N}$, and the blockmodel, B. Cross-validation suggests an accurate estimate for K. Nested variational scheduling of parameter updates makes inference parallelizable and a typically reaches a better solution than the naïve scheduling.

First we consider, whom-do-like relations among 18 novices in a New England monastery. The unsupervised analysis demonstrates the type of patterns that MMB recovers from data, and allows us to contrast the summaries of the original measurements achieved through prediction and de-noising.

The data was collected by Sampson during his stay at the monastery, while novices were preparing to join the monastic order [10]. Sampson's original analysis is rooted in direct anthropological observations. He made a strong case for 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). Scholars in the social sciences typically regard the faction labels assigned by Sampson to the novices (and his conclusions, more in general) as ground truth to the extent of assessing the quality of results of quantitative analyses; we shall do the same here. Using the nested variational EM algorithm above, we fit an array of mixed membership blockmodels with different values of K, and collected model estimates $\{\hat{\alpha}, \hat{B}\}$ and posterior mixed membership vectors $\vec{\pi}_{1:18}$ for the novices. We used an approximation of BIC to choose the value of K supported by the data. This criterion selects $\hat{K} = 3$, the same number of proper groups that Sampson identified based on anthropological observations—the waverers are interstitial members, rather than a group. Figure 2 shows the patterns that the mixed membership blockmodel with $\hat{K} = 3$ recovers from data. In particular, the top-left panel shows a graphical representation of the blockmodel $\hat{B}$. The block that we can identify a-posteriori with the loyal opposition is portrayed as central to the monastery, while the block identified with the outcasts shows the lowest internal coherence, in accordance with Sampson's observations. The top-right panel illustrates the posterior means of the mixed membership scores, $\mathbb{E}[\vec{\pi}|Y]$, for the 18 monks in the monastery. The model (softly) partitions the monks according to Sampson's classification, with Young Turks, Loyal Opposition, and Outcasts dominating each corner respectively. Notably, we can quantify 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 bottom panels contrast the different resolution of the original adjacency matrix of whom-do-like sociometric relations (left panel) obtained with the two analyses MMB enables.

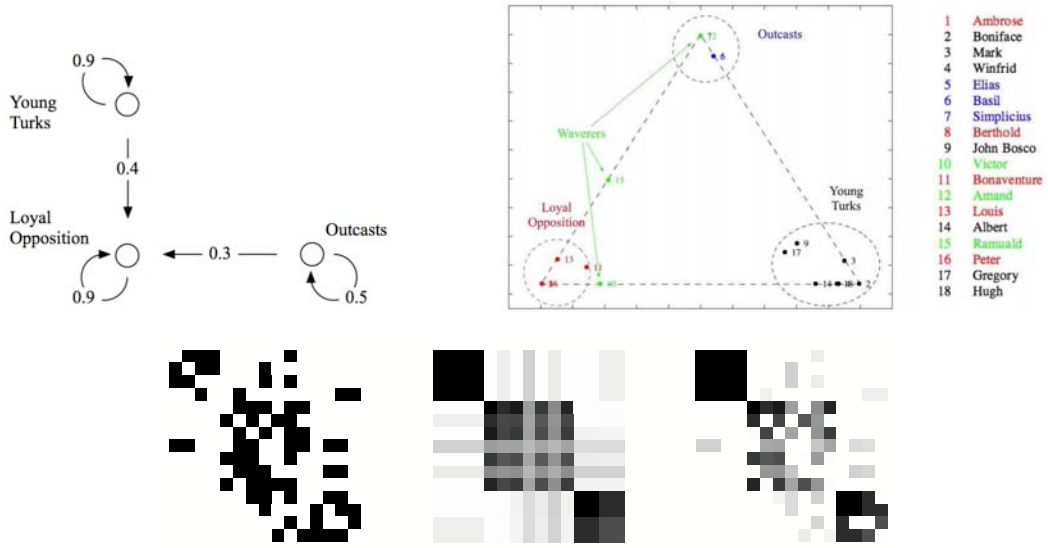

Figure 2: Top-Left: Estimated blockmodel, $\hat{B}$. Top-Right: Posterior mixed membership vectors, $\vec{\pi}_{1:18}$, projected in the simplex. The estimates correspond to a model with $\hat{B}$ top-left, and $\hat{\alpha} = 0.058$. Numbered points can be mapped to monks' names using the legend on the right. The colors identify the four factions defined by Sampson's anthropological observations. Bottom: 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).

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 $\hat{\alpha}, \hat{B}$ 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). Substantively, the unsupervised analysis of the sociometric relations with MMB offers quantitative support to several of Sampson's observations.

Second, we consider a friendship network among a group of 69 students in grades 7–12. The analysis here directly compares clustering results obtained with MMB to published results obtained with competing models, in a setting where a fair amount of social segregation is expected [2, 3].

The data is a collection of friendship relations among 69 students in a school surveyed in the National Study of Adolescent Health. The original population in the school of interest consisted of 71 students. Two students expressed no friendship preferences and were excluded from the analysis. We used variational EM algorithm to fit an array of mixed membership blockmodels with different values of K, collected model estimates, and used an approximation to BIC to select K. This procedure identified $\hat{K} = 6$ as the model-size that best explains the data; note that six is the number of grade-groups in the student population. The blocks are clearly interpretable a-posteriori in terms of grades, thus providing 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 with 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 [2], and the latent space cluster model (LSCM) in [3]. 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 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 extra-flexibility MMB offers over MB and LSCM reduces bias in the prediction of the membership of students to blocks, in this problem. 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.

	MMB Clusters						MB 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. MMSB is the proposed mixed membership stochastic blockmodel, MSB is the mixture blockmodel in [2], and LSCM is the latent space cluster model in [3].

Third, we consider 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—consider two models that suggest different sets of interactions as reliable; we prefer the model that reveals *functionally relevant* interactions.

The pairwise measurements consist of a hand-curated collection of physical protein interactions made available by the Munich Institute of Protein Sequencing (MIPS). The yeast genome database provides independent functional annotations for each protein, which we use for evaluating the functional content of the protein networks estimated with the MMB from the MIPS data, as detailed below. We explored a large model space, $K = 2 \dots 225$, and used five-fold cross-validation to identify a blockmodel B that reduces the dimensionality of the physical interactions among proteins in the training set, while revealing robust aspects of connectivity that can be leveraged to predict physical interactions among proteins in the test set. We determined that a fairly parsimonious model, $K = 50$, provides a good description of the observed physical interaction network. 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. 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 protein complex. We then evaluated the functional content of the posterior induced by MMB. The goal is to assess to what extent MMB reveals substantive information about the functionality of proteins that can be used to inform subsequent analyses. To do this, first, we fit a model on the whole data set to estimate the blockmodel, $B_{(50\times 50)}$, and the mixed membership vectors between proteins and blocks, $\vec{\pi}_{1:871}$, and second, we either impute physical interactions by thresholding the posterior expectations computed using blockmodel and node-specific memberships (summarization task), or we de-noise the observed interactions using blockmodel and pair-specific memberships (de-noising task). Posterior expectations of each interaction are in $[0, 1]$. Thresholding such expectations at q, for instance, leads to a collection of binary physical interactions that are at reliable with probability $p \geq q$. We used an independent set of functional annotations from the yeast database (SGD at www.yeastgenome.org) to decide which interactions are functionally meaningful; namely those between pairs of proteins that share at least one functional annotation. In this sense, between two models that suggest different sets of interactions as reliable, our evaluation assigns a higher score

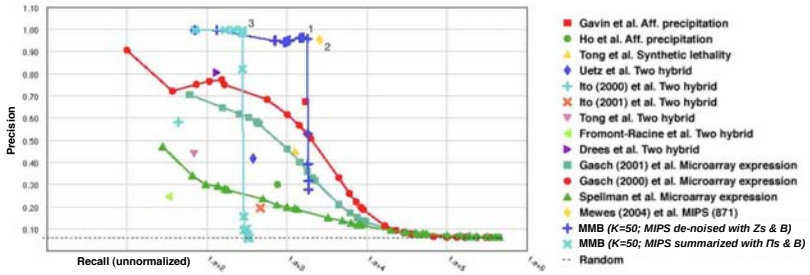

Figure 3: Functional content of the MIPS collection of protein interactions (yellow diamond) on a precision-recall plot, compared against other published collections of interactions and microarray data, and to the posterior estimates of the MMB models—computed as described in the text.

to the model that reveals *functionally relevant* interactions according to SGD. Figure 3 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 protein 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). Importantly, the estimated networks corresponding to lower levels of recall for both model variants (i.e., $\times$ and $+$) feature a more precise functional content than the original network. This means that the proposed latent block structure is helpful in effectively denoising 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 protein binding data that instantiate cellular activity of specific biological contexts, such as biopolymer catabolism and homeostasis. In conclusion, our results suggest that MMB successfully reduces the dimensionality of the data, while discovering information about the multiple functionality of proteins that can be used to inform follow-up analyses.

Remarks. **A.** 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. **B.** MMB includes two sources of variability, $B, \Pi s$, 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. **C.** 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 instance, for analyzing multivariate sociometric relations simultaneously. Second, in the MMSB the data generating distribution is a Bernoulli, but B can be a matrix of parameterizes for any kind of distribution. For instance, technologies for measuring interactions between pairs of proteins, such as mass spectrometry and tandem affinity purification, which return a probabilistic assessment about the presence of interactions, thus setting the range of $Y \in [0, 1]$.

References

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

[2] 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.

[3] 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.

[4] 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.

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

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

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

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

[9] 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.

[10] F. S. Sampson. *A Novitiate in a period of change: An experimental and case study of social relationships*. PhD thesis, Cornell University, 1968.

[11] S. Wasserman, G. Robins, and D. Steinley. A brief review of some recent research. In: *Statistical Network Analysis: Models, Issues and New Directions*, Lecture Notes in Computer Science. Springer-Verlag, 2007.