Mixed Membership Stochastic Block Models for Relational Data with Application to Protein-Protein Interactions

E. M. Airoldi,[1,*] **D. M. Blei,**[2] **S. E. Fienberg**[1,3] **and E. P. Xing**[1]

[1] School of Computer Science, Carnegie Mellon University

[2] Department of Computer Science, Princeton University

[3] Department of Statistics, Carnegie Mellon University

SUMMARY. Modeling relational data is an important problem for modern data analysis and machine learning. In this paper we propose a Bayesian model that uses a hierarchy of probabilistic assumptions about the way objects interact with one another in order to learn latent groups, their typical interaction patterns, and the degree of membership of objects to groups. Our model explains the data using a small set of parameters that can be reliably estimated with an efficient inference algorithm. In our approach, the set of probabilistic assumptions may be tailored to a specific application domain in order to incorporate intuitions and/or semantics of interest.

We demonstrate our methods on simulated data, where they outperform spectral clustering techniques, and we apply our model to a data set of protein-to-protein interactions, to reveal proteins' diverse functional roles.

KEY WORDS: Bayesian inference; Latent Variables; Hierarchical mixture model; Variational approximation; Social networks.

---

[*] *email:* eairoldi@cs.cmu.edu

1. Introduction

Modeling relational data is an important problem for modern data analysis and machine learning. Many data sets contain interrelated observations. For example, scientific literature connects papers by citation, web graphs connect pages by links, and protein-protein interaction data connect proteins by physical interaction records. Such data violate the classical exchangeability assumptions made in machine learning and statistics; moreover, the relationships between data are often of interest as observations themselves. One may try to predict citations of newly written papers, predict the likely links of a web-page, or cluster proteins based on patterns of interaction between them.

There is a history of probabilistic models for relational data analysis in Statistics. Part of this literature is rooted in the stochastic block modeling ideas from psychometrics and sociology. These ideas are due primarily to Holland and Leinhardt (1975), and later elaborated upon by others, e.g., see Fienberg et al. (1985), Wasserman and Pattison (1996), Snijders (2002), Hoff et al. (2002). In machine learning, Markov random networks have been used for link prediction (Taskar et al., 2003) and the traditional block models from Statistics have been extended with nonparametric Bayesian priors (Kemp et al., 2004).

In this paper, we develop a mixed membership model for analyzing patters of interaction between data. Mixed membership models for soft classification have emerged as a powerful and popular analytical tool for analyzing large databases involving text (Blei et al., 2003), text and references (Cohn and Hofmann, 2001; Erosheva et al., 2004), text and images (Barnard et al., 2003), multiple disability measures (Erosheva and Fienberg, 2005; Manton

et al., 1994), and genetics information (Rosenberg et al., 2002; Pritchard et al., 2000; Xing et al., 2003). These models use a simple generative model, such as bag-of-words or naive Bayes, embedded in a hierarchical Bayesian framework involving a latent variable structure; this induces dependencies and introduces statistical control over the estimation of what might otherwise be an extremely large set of parameters.

We propose a Bayesian model that uses a hierarchy of probabilistic assumptions about how objects interact with one another in order to learn latent groups, their typical interaction patterns, and the degree of membership of objects to groups. Given data, we find an approximate posterior distribution with an efficient variational inference algorithm. In our approach, the set of probabilistic assumptions may be tailored to a specific application domain in order to incorporate semantics of interest. We demonstrate our methods on simulated data, and we successfully apply the model to a data set of protein-protein interactions.

2. Mixed Membership Stochastic Block Models

In this section, we describe a probabilistic model of interaction patterns in a group of objects. Each object can exhibit several patterns that determine its relationships to the others. We will use protein-protein interaction modeling as a working example; however, the model can be used for any relational data where the primary goal of the analysis is to learn latent group interaction patterns and mixed group membership of a set of objects. Throughout the paper we refer to our model as the *mixed membership stochastic block model* or MMSB.

Suppose we have observed the physical interactions between N proteins[1]. We represent the interaction data by an $N \times N$ binary adjacency matrix $\boldsymbol{r}$ where $r_{i,j} = 1$ if the i-th protein interacts with the j-th protein. Usually, an interaction between a pair of proteins is indicative of a unique biological function they both involve; it may be possible to infer the functional classes of the study proteins from the protein interactions.

In a complex biological system, many proteins are functionally versatile and can participate in multiple functions or processes at different times or under different biological conditions. Thus, when modeling functional classes of the proteins, it is natural to adopt a flexible model which allows multiple scenarios under which a protein can interact with its partners. For example, a signal transduction protein may sometimes interact with a cellular membrane protein as part of a signal receptor; at another time, it may interact with the transcription complex as an auxiliary transcription factor. By assessing the similarity of observed protein-to-protein interaction patterns, we aim to recover the latent function groups and the degree with which the proteins take part in them.

In the generative process, we model the observed adjacency matrix as a collection of Bernoulli random variables. For each pair of objects, the presence or absence of an interaction is drawn by (1) choosing a latent class for each protein from a protein-specific distribution and (2) drawing from a Bernoulli distribution with parameter associated with the pair of latent

[1]Such information can be obtained experimentally with "yeast two-hybrid" tests and others means, and in practice the data may be noisy. For simplicity, we defer explicit treatment of observation noise, although plugging in appropriate error processes is possible.

classes. A protein represents several functional groups through its distribution of latent classes; however, each protein participates in one function when determining its relationship to another.

For a model with K groups, the parameters are K-dimensional Dirichlet parameters $\boldsymbol{\alpha}$, a $K \times K$ matrix of Bernoulli parameters $\boldsymbol{\eta}$, and $\rho \in [0, 1]$ which is described below. Each $\boldsymbol{\theta}_i$ is a Dirichlet random variable (i.e., a point on the $K-1$ simplex) and each $\boldsymbol{z}_{i \to j}$ and $\boldsymbol{z}_{i \leftarrow j}$ are indicators into the K groups. The generative process of the observations, $\boldsymbol{r}_{(N \times N)}$, is as follows:

1. For each object $i = 1, \ldots, N$:

   1.1. Sample $\boldsymbol{\theta}_i \sim Dirichlet\,(\boldsymbol{\alpha})$.

2. For each pair of objects $(i, j) \in [1, N] \times [1, N]$:

   2.1. Sample group $\boldsymbol{z}_{i \to j} \sim Multinomial\,(\boldsymbol{\theta}_i, 1)$

   2.2. Sample group $\boldsymbol{z}_{i \leftarrow j} \sim Multinomial\,(\boldsymbol{\theta}_j, 1)$

   2.3. Sample $r_{ij} \sim Bernoulli\ \left(\rho\, \boldsymbol{z}'_{i \to j} \boldsymbol{\eta} \boldsymbol{z}_{i \leftarrow j} + (1 - \rho)\, \delta_0 \right)$

The parameter ρ controls how often a zero is due to noise and how often it occurs as a function of the constituent proteins' latent class memberships in the generative process. In turn, this leads to ones in the matrix being weighted more as ρ decreases, and allows for the model to pick up sparsely interconnected clusters. For the rest, the model uses three sets of latent variables. The Multinomial parameters $\boldsymbol{\theta}_i$ are sampled once for the entire collection of observations; the latent cluster indicators $\boldsymbol{z}_{i \to j}$ and $\boldsymbol{z}_{i \leftarrow j}$ are sampled once for each protein-protein interaction variable r_{ij}.

The generative process described above leads to a joint probability distribution over the observations and the latent variables,

$$p(\boldsymbol{r}, \boldsymbol{\theta}, \boldsymbol{z}_1, \boldsymbol{z}_2 | \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{i=1}^{N} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \prod_{j=1}^{N} p(\boldsymbol{z}_{i \to j} | \boldsymbol{\theta_i}) p(\boldsymbol{z}_{i \leftarrow j} | \boldsymbol{\theta_j}) \, p(r_{ij} | \boldsymbol{z}_{i \to j}, \boldsymbol{z}_{i \leftarrow j}, \boldsymbol{\eta}).$$

The marginal probability of the observations is not tractable to compute,

$$p(\boldsymbol{r} | \boldsymbol{\alpha}, \boldsymbol{\eta}) = \int_{\Theta} \int_{Z} \prod_{i=1}^{N} p(\boldsymbol{\theta}_i | \boldsymbol{\alpha}) \prod_{j=1}^{N} p(\boldsymbol{z}_{i \to j} | \boldsymbol{\theta_i}) p(\boldsymbol{z}_{i \leftarrow j} | \boldsymbol{\theta_j}) \, p(r_{ij} | \boldsymbol{z}_{i \to j}, \boldsymbol{z}_{i \leftarrow j}, \boldsymbol{\eta}) \, d\boldsymbol{z} \, d\boldsymbol{\theta}.$$

We carry out *approximate* inference and parameter estimation to deal with this issue.

The only input to this model is the number of groups. The goal is to learn the posterior distribution of the membership proportions of each protein and the group interaction probabilities. We will focus on the interpretability of these quantities, e.g., consistent functional annotations of the proteins within groups. Note that there are several ways to select the number of groups. For example, Kemp et al. (2004) use a nonparametric Bayesian prior for a single-membership block model.

In our fully generative approach, it is possible to integrate outside information about the objects into the hierarchy of probabilistic assumptions. For example, we can include outside information about the proteins into the generative process that includes the linkage. In citation data, document words can be modeled along with how the documents cite each other.

3. Inference and Estimation

In this section we present the elements of approximate inference essential for learning the hyper-parameters of the model and inferring the posterior distribution of the degrees of membership for each object.

In order to learn the hyper-parameters we need to be able to evaluate the likelihood, which involves a non-tractable integral as we stated above—see equation. In order to infer the degrees of membership corresponding to each object, we need to compute the posterior degrees of membership given the hyper-parameters and the observations

$$p(\boldsymbol{\theta}|\boldsymbol{r}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{r}|\boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\boldsymbol{r}|\boldsymbol{\alpha}, \boldsymbol{\eta})}, \tag{1}$$

Using variational methods, we can find a lower bound of the likelihood and approximate posterior distributions for each object's membership vector.

The basic idea behind variational methods is to posit a variational distribution on the latent variables $q(\boldsymbol{\theta}, \boldsymbol{z})$, which is fit to be close to the true posterior in Kullback-Leibler (KL) divergence. This corresponds to maximizing a lower bound, $\mathbb{L}\left[\, \boldsymbol{\gamma}, \boldsymbol{\phi}\,;\, \boldsymbol{\alpha}, \boldsymbol{\eta}\,\right]$, on the log probability of the observations given by Jensen's inequality:

$$\begin{aligned}
\log p(\boldsymbol{r}|\boldsymbol{\alpha}, \boldsymbol{\eta}) \;\geq\; & \sum_{i=1}^{N} \mathbb{E}_q\left[\, \log p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i) \,\right] + \\
& + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_q\left[\, \log p(\boldsymbol{z}_{i\to j}|\boldsymbol{\theta}_i) \,\right] + \\
& + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_q\left[\, \log p(\boldsymbol{r}_{ij}|\boldsymbol{z}_{i\to j}, \boldsymbol{z}_{i\leftarrow j}, \boldsymbol{\eta}) \,\right] + \\
& + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_q\left[\, \log p(\boldsymbol{z}_{i\leftarrow j}|\boldsymbol{\theta}_j) \,\right] - \\
& - \mathbb{E}_q\left[\, \log q(\boldsymbol{\theta}, \boldsymbol{z}) \,\right].
\end{aligned}$$

where the expectations are taken with respect to $q(\boldsymbol{\theta}, \boldsymbol{z})$. We choose a fully factorized variational distribution such that this optimization is tractable.

3.1 *Variational Inference*

The fully factorized variational distribution q is as follows.

$$\begin{aligned} q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) &= \prod_{i=1}^{N} q(\boldsymbol{\theta}_i|\boldsymbol{\gamma}_i) \prod_{j=1}^{N} \Big(q(\boldsymbol{z}_{i\to j}|\boldsymbol{\phi}_{i\to j})\, q(\boldsymbol{z}_{i\leftarrow j}|\boldsymbol{\phi}_{i\leftarrow j}) \Big) \\ &= \prod_{i=1}^{N} Dirichlet\,(\boldsymbol{\theta}_i|\boldsymbol{\gamma}_i) \prod_{j=1}^{N} \Big(Mult\,(\boldsymbol{z}_{i\to j}|\boldsymbol{\phi}_{i\to j})\, Mult\,(\boldsymbol{z}_{i\leftarrow j}|\boldsymbol{\phi}_{i\leftarrow j}) \Big) \end{aligned}$$

The lower bound for the log likelihood $\mathbb{L}[\boldsymbol{\gamma}, \boldsymbol{\phi}\,; \boldsymbol{\alpha}, \boldsymbol{\eta}]$ can be maximized using exponential family arguments and coordinate ascent (Wainwright and Jordan, 2003). This leads to the following updates for the variational parameters $(\boldsymbol{\phi}_{i\to j}, \boldsymbol{\phi}_{i\leftarrow j})$, for each pair (i, j):

$$\begin{aligned} \phi^*_{i\to jg} &\propto \exp\Big\{ \psi(\gamma_{ig}) - \psi(\sum_{g=1}^{K} \gamma_{ig}) \Big\} \prod_{h=1}^{K} \eta_{gh}^{r_{ij}\,\phi_{i\leftarrow jh}} \prod_{h=1}^{K} (1-\eta_{gh})^{(1-r_{ij})\,\phi_{i\leftarrow jh}} \\ \phi^*_{i\leftarrow jh} &\propto \exp\Big\{ \psi(\gamma_{jh}) - \psi(\sum_{h=1}^{K} \gamma_{jh}) \Big\} \prod_{g=1}^{K} \eta_{gh}^{r_{ij}\,\phi_{i\to jg}} \prod_{g=1}^{K} (1-\eta_{gh})^{(1-r_{ij})\,\phi_{i\to jg}} \end{aligned}$$

for $g, h = 1, \ldots, K$, and to the following updates for the variational parameters $\boldsymbol{\gamma}_i$, for each i:

$$\gamma^*_{ig} = \alpha_g + \sum_{j=1}^{N} \phi_{i\to jg} + \sum_{j=1}^{N} \phi_{i\leftarrow jg}.$$

The vectors $\boldsymbol{\phi}_{i\to j}$ and $\boldsymbol{\phi}_{i\leftarrow j}$ are normalized to sum to one. The complete algorithm to perform variational inference in the model is described in detail in Figure 1. Variational inference is carried out for fixed values of $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$, in order to maximize the lower bound for the likelihood. Then we maximize the lower bound with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$. We iterate these two steps (variational inference and maximization) until convergence. The overall procedure is a variational expectation-maximization (EM) algorithm (Xing et al., 2003).

3.2 *Remarks*

The variational inference algorithm presented in Figure 1 is not the naïve variational inference algorithm. In the naïve version of the algorithm, we initialize the variational Dirichlet parameters $\boldsymbol{\gamma}_i$ and the variational Multinomial parameters $\boldsymbol{\phi}_{i\to j}$ and $\boldsymbol{\phi}_{i\leftarrow j}$ to non-informative values, then we iterate until convergence the following two steps: (i) update $\boldsymbol{\phi}_{i\to j}$ and $\boldsymbol{\phi}_{i\leftarrow j}$ for all pairs (i, j), and (ii) update $\boldsymbol{\gamma}_i$ for all objects i. In such algorithm, at each variational inference cycle we need to allocate $NK + 2N^2K$ numbers.

The nested variational inference algorithm trades time for space thus allowing us to deal with large graphs; at each variational cycle we need to allocate $NK + 2K$ numbers. The increased running time is partially offset by the fact that the algorithm can be parallelized and leads to empirically observed faster convergence rates, as we show in Figure 3. This algorithm is also better than MCMC variations (i.e., blocked and collapsed Gibbs samplers) in terms of memory requirements and/or convergence rates.

[Figure 1 about here.]

It is also important to note that the variational Dirichlet parameters $\boldsymbol{\gamma}$ and the Bernoulli parameters $\boldsymbol{\eta}$ are closely related in this model: it is necessary to keep the $\boldsymbol{\gamma}$s across variational-EM iterations in order to better inform the M-step estimates of $\boldsymbol{\eta}$. Thus, we smooth the $\boldsymbol{\gamma}$ parameters in between EM iterations instead of resetting them to a non-informative value, $2N/K$ in our model. Using a damping parameter ϵ we obtain: $\tilde{\gamma}_{ig} = (1 - \epsilon)\,\gamma^*_{ig} + \epsilon\,\frac{2N}{K}$.

3.3 *Parameter Estimation*

Using the optimal lower bound $\mathbb{L}[\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*\,; \boldsymbol{\alpha}, \boldsymbol{\eta}]$ as a tractable surrogate for the likelihood we here look for (pseudo) empirical Bayes estimates for the hyper-parameters (Carlin and Louis, 2005).

Such maximization amounts to maximum likelihood estimation of the Dirichlet parameters α and Bernoulli parameter matrix η using expected sufficient statistics, where the expectation is taken with respect to the variational distribution. Finding the MLE of a Dirichlet requires numerical optimization (Minka, 2000). For each Bernoulli parameter, the approximate MLE is:

$$\eta^*_{gh} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{i\to jg}\,\phi_{i\leftarrow jh}\,r_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{i\to jg}\,\phi_{i\leftarrow jh}},$$

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

We also smooth the probabilities of interactions between any member of group a and any member of group b, that is $\eta_{a,b}$, by assuming $\eta_{a,b} \sim Beta\,(\beta_1, \beta_2)$ for each pair of groups $(a, b) \in [1, K] \times [1, K]$. Variational inference is modified appropriately.

4. Examples and Simulation Experiments

We first tested our model in a controlled setting. We simulated non-contrived adjacency matrices mimicking protein-protein interactions with 100 proteins and four groups, 300 proteins and 10 groups, and 600 proteins and 20 groups. In our experiment, the signal-to-noise ratio is decreasing with the size of the problem, for a fixed Dirichlet parameter $\alpha < 1$.[2] The data are display in

[2]That is, a fixed $\alpha < 1$ leads to a number of active functions for each protein that increases linearly with the total number of latent functions, but the

Figure 5, where the S/N ratio is roughly 0.5, 0.4 and 0.3 for the both the top and bottom rows, from left to right.

In Figure 2 we compare our model to spectral clustering with local scaling (Zelnik-Manor and Perona, 2004) that is particularly suited for recovering the structure of the interactions in the case when proteins take part in a single function. Note that spectral clustering (or normalized cuts) minimizes the total transition probability due to 1-step random walk of objects between clusters. Each object is assumed to have a unique cluster membership. Our model, however, is more flexible. It allows object to have different cluster membership while interacting with different objects. The simulations with the Dirichlet parameter $\alpha = 0.05$ are meant to provide mostly unique membership; spectral clustering performs well and our model has a slightly better performance. As proteins participate to more functions, that is, α increases to 0.25 in our simulations, spectral clustering is not an adequate solution anymore. Our model, on the other hand, is able to recover the mixed membership to a large degree, and performs better than spectral clustering.

[Figure 2 about here.]

In a more general formulation of our model we accommodate a collection of observations, e.g., protein-protein interaction patterns measured by different laboratories and under possible different conditions, or daily summaries of email exchanges. We used this general model to understand how the model takes advantage of the information available. Empirical results show that it

---

number of interactions sampled among functional groups decreases with the square of the total number of latent function, and causes an overall decrease of the informative part of the observed matrix $\boldsymbol{r}$.

is better to have a larger adjacency matrix rather than having a collection of small matrices, in order to overcome a fixed signal-to-noise ratio.

In Figure 3 compare the running time of our enhanced variational-EM algorithm to the naïve implementation. Our algorithm is more efficient in terms of space and converges faster. Further, it can be parallelized given that the updates for each interaction (i, j) are independent of one another.

[Figure 3 about here.]

To perform data analysis using our model we need to select the number of clusters, K, in advance. For the analysis of simulated data, we use cross-validation to this extent. That is, we pick the value of K that maximizes the likelihood on a test set. In Figure 4 we show an example, where the latent number of clusters equals 10.

[Figure 4 about here.]

5. Application to Protein-Protein Interactions

Protein-protein interactions (PPI) form the physical basis for formation of complexes and pathways which carry out different biological processes. A number of high-throughput experimental approaches have been applied to determine the set of interacting proteins on a proteome-wide scale in yeast. These include the two-hybrid (Y2H) screens and mass spectrometry methods. For example, mass spectrometry is used to identify components of protein complexes (Gavin et al., 2002; Ho et al., 2002). High-throughput methods, though, may miss complexes that are not present under the given conditions, for example, tagging may disturb complex formation and weakly associated components may dissociate and escape detection.

The MIPS database was created in 1998 based on evidence derived from a variety of experimental techniques and does not include information from high-throughput data sets (Mewes et al., 2004). It contains about 8000 protein complex associations in yeast. We analyze a subset of this collection containing 871 proteins, the interactions amongst which were hand-curated. In Table 1 we summarize the main functions of the protein in our sub-collection. Note that, since most proteins participate in more than one function, Table 1 contains more counts (2119) than proteins (871), for an average of ≈ 2.4 functions per protein. Further, the relative importance of each functional category in our sub-collection, in terms of the number of proteins involved, is different from the relative importance of the functional categories over the entire MIPS collection, as reported in Lanckriet et al. (2004).

[Table 1 about here.]

5.1 *Recovering the Ground Truth*

Our data consists of 871 proteins participating in 255 functions. The functions are organized into a hierarchy, and the 15 functions in Table 1 are those at the top level of the hierarchy. The ground truth for our analysis is constituted by the presence or absence of functional annotations at the top level of the hierarchy. That is, each proteins is associated with a 15-dimensional vector of zeros and ones, where the ones indicate participation in high-level functional categories of sub-categories. There are about 2200 functional annotations in our data sets, that is, the density of the proteins-to-functions annotation matrix is about 16%. The Dirichlet parameter α corresponding to the true mixed-membership is ≈ 0.0667. Most of the proteins in our data participate in two to four functions. In Figure 5 we show

the normalized frequencies of participation of each protein in sub-categories of the 15 high-level functions, which ware derived using the manually curated functional annotations.

[Figure 5 about here.]

5.2 *Evaluating the Performance*

In order to evaluate the performance of the competing methods in predicting the (possibly) multiple functional annotations of proteins we devised a very simple measure of accuracy. Briefly, we added the number of functional annotations correctly predicted for each proteins, divided by the total number of functional annotations.

Note that, given their exchangeable nature, the latent functional groups are not identifiable in our model. On the other hand, in order to compute the accuracy above we need to decide which latent cluster correspond to which functional class. We resolved the ambiguity by finding the one mapping that maximized the accuracy on the training data. In those cases where no training data is available, e.g., the unsupervised experiment, we minimize the divergence between marginal true and predicted marginal frequencies of membership, instead. We then used that permutation in order to compare predicted functional annotations to the ground truth, for all proteins.

[Figure 6 about here.]

In order to compute the accuracy of spectral clustering with local scaling, we implemented softened a soft version of it; we used the cluster predictions and the relative distances between proteins and the centroids of the clusters to obtain normalized scores (probabilities) of membership of each protein to

each cluster. These mixed-membership scores enabled us to compute the accuracy measure.

Further, we devised two baselines to compare the accuracy of our methods against: the "dumb" random annotator, and the "clever" random annotator. The dumb annotator knows the probability of a functional annotation in general, and annotates at random each function of each protein with that probability. The clever annotator knows the probability of a functional annotation for each function, and assigns annotations at random accordingly. Note that the clever annotator can perfectly map latent groups to functions.

5.3 *Testing Functional Interaction Hypothesis*

The mixed-membership stochastic block model is a useful tool to explore hypothesis about the nexus between latent protein interaction patterns and the functions they are able to express.

For example, it is reasonable to assume that proteins that share a common functional annotation tend to interact with one another more often than with proteins with no functional annotations in common. In order to test this hypothesis we can fix the function interaction matrix $\boldsymbol{\eta}$ to be the identity matrix. This leads to accuracies of 76.31% for the latent mixed-membership model and of 71.4% for spectral clustering. In this case the mapping of latent clusters to functions was obtained by minimizing the divergence between marginal true and predicted marginal frequencies of membership.

5.4 *De-Noising Protein-Protein Interactions*

It is reasonable to assume that a collection of PPI may inform us on the functions protein are able to express (Deng et al., 2002). In the bigger picture, the goal is to use our model to estimate interaction patterns and functional

membership of proteins from manually curated data, in order to de-noise functional annotations and interactions made available via high-throughput experiments.

In the experiments we present in this section, we aim at assessing the prediction error associated with our model. To this extent we perform cross-validation, and in order to obtain conservative estimates of the error we split the proteins into a training set and a testing set of about the same size. We slightly modify our model in order to predict the functional mixed-membership probabilities of new proteins, i.e., those in the testing set. In particular, we use available information to learn the function interaction matrix $\boldsymbol{\eta}$, which encodes the interaction patterns between pairs of proteins as they express a corresponding pair of functions. We also consider known the functional annotations of the proteins in the training data in terms of their corresponding mixed membership probabilities $\boldsymbol{\theta}_i$. In order to estimate $\boldsymbol{\eta}$ we considered all protein pairs in the training set, and estimated the strength of the interactions between pairs of expressed functions by composing the corresponding membership probabilities of the proteins involved, under assumption of independence. In the testing phase, we fixed $\boldsymbol{\eta}$, and the $\boldsymbol{\theta}_i$ for the proteins in the training set and fit our model in order to infer the mixed-membership probability vectors of the proteins in the testing set. Alternatives are possible, where the information available is used to calibrate priors for the elements of $\boldsymbol{\eta}$, rather than fixing its values.

We perform 100 such experiments, where each replicate differs for the subset of proteins used for training. In order to threshold the estimated mixed membership scores, and pick the most likely annotations, we used

the estimated frequency of functional annotations in the training set. The accuracy of the predictions obtained with MMSB is 85% on average. Our method significantly outperforms both the dumb random annotator, 74% accurate on average, and the clever random annotator, 83% accurate on average—Figure 7 shows the distribution of the accuracy in the various cases. This suggests that our method leverages the signal in the data, and that the identification of latent groups to functions is somewhat feasible in this application.

[Figure 7 about here.]

Figure 8 displays few examples of predicted mixed membership probabilities against the true annotations, given the estimated mapping of latent groups to functions.

[Figure 8 about here.]

Figure 9 shows the predicted mixed membership probabilities for 841 proteins. Most proteins are predicted as participating in at least two functions. The predicted degree of membership is reasonably good, and the estimated Dirichlet parameter is $\hat{\alpha} = 0.417$.

[Figure 9 about here.]

6. Discussion

In the experiments above, we have presented the mixed-membership stochastic block model (MMSB) for relational data with stochastic and heterogeneous interactions among objects. In particular, the mixed-membership assumption is very desirable for modeling real data. Given a collection of

interaction patterns, our model yields posterior estimation of the multiple group membership of objects, which align closely to real world scenarios (e.g., multi-functionality of proteins). Further, our model estimates interaction probabilities between pairs of latent groups.

In simulations, our model out-performs spectral clustering both in cases when objects have single membership and in cases when objects have mixed-membership. In this latter case, the differential performance of latent mixed-membership model over spectral clustering (with local scaling) is remarkable, since spectral clustering lacks a device for capturing mixed membership. The parameter ρ of MMSB enables to recover clusters whose objects are sparsely interconnected, by assigning more weight to the observed edges, i.e., the ones in the observed adjacency matrix $\boldsymbol{r}$. On the contrary, spectral clustering methods assign equal weight to both ones and zeros in the adjacency matrix $\boldsymbol{r}$, so that the classification is driven by the zeros in cases where the number of zeros is overwhelming—this may be a not desirable effect, thus it is important to be able to modulate it, e.g., with ρ.

We then applied our model to the task of predicting the functional annotation of proteins by leveraging protein-protein interaction patterns. We showed how our model provides a valuable tool to test hypothesis about the nexus between PPI and functionality. We showed a strategy to perform cross-validation experiments in this setting, to demonstrate how to fit our model and make use of reliable information (about PPI) in order to infer the functionality of unlabeled proteins. However, that is not the only strategy. An alternative strategy we are currently exploring is that of calibrating informative priors for $\boldsymbol{\eta}$ using the training data. An informative prior would

both smooth the estimates of parameters, on the testing data, and increase the identifiability of the latent groups. Last, in the analyses we presented in this paper we fixed $K = 15$ and estimated a mapping between latent groups and functions. An alternative we are currently exploring is to make the partially available functional annotation part of the model, and select K independently of the number of functional categories.

7. Conclusions

In conclusion, our mixed membership stochastic block model provides a valuable tool for "summarizing" relational data. Specifically, the MMSB both projects the observed interactions into a lower dimensional "latent" space, the space of group-to-group connectivity patterns, and assigns mixed membership of objects of study to groups. The connectivity patterns are captured by $\boldsymbol{\eta}$ and the mixed membership scores are captured by $\boldsymbol{\theta}$.

There is a relationship between the MMSB and the latent space model of relational data (Hoff et al., 2002). In the latent space model, the latent vectors are drawn from Gaussian distributions and the interaction data is drawn from a Gaussian with mean $\boldsymbol{\theta}_i' I \boldsymbol{\theta}_j$. In the MMSB, the marginal probability of an interaction takes a similar form, $p(r_{i,j}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{\eta}) = \boldsymbol{\theta}_i' \boldsymbol{M} \boldsymbol{\theta}_j$, where $M_{ij} = p(r_{i,j}|\boldsymbol{\eta})$ is a matrix of probabilities for each pair of latent functional states in the collection. In contrast to the latent space model, the interaction data can be modeled by an arbitrary distribution, in our model. With binary relationships, i.e., a graph, we can use a collection of Bernoulli parameters; with continuous relationships, we can use a collection of Gaussian parameters. While more flexible, the MMSB does not subsume the latent space model; they make different assumptions about the data.

When compared to spectral clustering techniques, MMSB allows to modulate the relative importance of presence and absence of interactions in the cost function that drives the assignments of objects to clusters, by modulating the parameter ρ. Further, MMSB empirically outperformed spectral clustering with local scaling (Zelnik-Manor and Perona, 2004) in all cases we tested.

In our applications to protein-protein interactions, recovering the mixed membership of proteins to clusters that relate to functionality provides a promising approach to learn the generative/mechanistic aspects underlying such data. This approach can be valuable for seeking deeper insight of the data, as well as for serving as informative priors for future estimation tasks. Our results confirm previous findings that information about PPI alone does not lead to accurate functional annotation of unlabeled proteins. More information is needed, for example, gene expression levels could be integrated in MMSB to boost the prediction accuracy of functional annotation.

7.1 *Future Work*

In the future we plan to explore PPI generated with high-throughput experimental methods: the tandem-affinity purification (TAP) and high-throughput mass spectrometry (HMS) complex data, described in Ho et al. (2002) and Gavin et al. (2002).

We will use all MIPS manually curated PPI, used for the analyses in this paper, to "calibrate informative priors" for the hyper-parameters in our model, in order to de-noise both the interactions and the functional annotations for the proteins in the TAP and HMS collections. The TAP collection contains 1363 proteins, 469 of which are contained in the MIPS hand-curated

collection, whereas the HMS collection contains 1578 proteins, and shares 330 of them with the MIPS hand-curated collection.

Acknowledgements

This paper was the recipient of the "John Van Ryzin Award" of the International Biometric Society, Eastern North American Region (ENAR).

This work was partially supported by National Institutes of Health (NIH) under Grant 1 R01 AG023141-01, by the Office of Naval Research (ONR) under Dynamic Network Analysis (N00014-02-1-0973), the National Science Foundation (NSF) and the Department of Defense (DOD) under MKIDS (IIS0218466). The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the NIH, the ONR, the NSF, the DOD, or the U.S. government.

References

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D. and Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research* **3**, 1107–1135.

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

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

Cohn, D. and Hofmann, T. (2001). The missing link—A probabilistic model

of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13*.

Deng, M. H., Zhang, K., Mehta, S., Chen, T. and Sun, F. Z. (2002). Prediction of protein function using protein-protein interaction data. In *IEEE Computer Society Bioinformatics Conference*.

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

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

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

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

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

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

Holland, P. W. and Leinhardt, S. (1975). *Sociological Methodology*, chapter Local structure in social networks, pages 1–45. Jossey-Bass.

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

Lanckriet, G. R., Deng, M., Cristianini, N., Jordan, M. I. and Noble, W. S. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing*.

Manton, K. G., Woodbury, M. A. and Tolley, H. D. (1994). *Statistical Applications Using Fuzzy Sets*. Wiley.

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

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

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

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (2002). Genetic structure of human populations. *Science* **298**, 2381–2385.

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

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

Wainwright, M. J. and Jordan, M. I. (2003). Graphical models, exponential

families and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley.

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

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

Xing, E. P., Ng, A. Y., Jordan, M. I. and Russel, S. (2003). Distance metric learning with applications to clustering with side information. In *Advances in Neural Information Processing Systems*, volume 16.

Zelnik-Manor, L. and Perona, P. (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608.

1. initialize $\boldsymbol{\gamma}_{ig}^0 = \frac{2N}{K}$ for all i, g
2. **repeat**
3. **for** $i = 1$ to N
4. **for** $j = 1$ to N
5. get **variational** $\boldsymbol{\phi}_{ij1}^{t+1}$ and $\boldsymbol{\phi}_{ij2}^{t+1} = f(r_{ij}, \boldsymbol{\gamma}_i^t, \boldsymbol{\gamma}_j^t, \boldsymbol{\eta}^t)$
6. partially update $\boldsymbol{\gamma}_{\boldsymbol{i}}^{\boldsymbol{t+1}}$, $\boldsymbol{\gamma}_{\boldsymbol{j}}^{\boldsymbol{t+1}}$ and $\boldsymbol{\eta}^{\boldsymbol{t+1}}$
7. **until** convergence

1. initialize $\phi_{ij1g}^0 = \phi_{ij2h}^0 = \frac{1}{K}$ for all g, h
2. **repeat**
3. **for** $g = 1$ to K
4. update $\phi_{ij1g}^{s+1} \propto f_1(\boldsymbol{\phi}_{ij2}^s, \boldsymbol{\gamma}, \boldsymbol{\eta})$
5. normalize $\boldsymbol{\phi}_{ij1}^{s+1}$ to sum to 1
6. **for** $h = 1$ to K
7. update $\phi_{ij2h}^{s+1} \propto f_2(\boldsymbol{\phi}_{ij1}^s, \boldsymbol{\gamma}, \boldsymbol{\eta})$
8. normalize $\boldsymbol{\phi}_{ij2}^{s+1}$ to sum to 1
9. **until** convergence

Figure 1. Top: The two-layered variational inference for $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$. The inner layer consists of Step 5. The function f is described in details in the bottom panel. **Bottom:** Inference for the variational parameters $(\boldsymbol{\phi}_{ij1}, \boldsymbol{\phi}_{ij2})$ corresponding to the basic observation $r_{i,j}$. This is the detailed description of Step 5. in the top panel. The functions f_1 and f_2 are updates for ϕ_{ij1g} and ϕ_{ij2h} described in the text of Section 3.1.

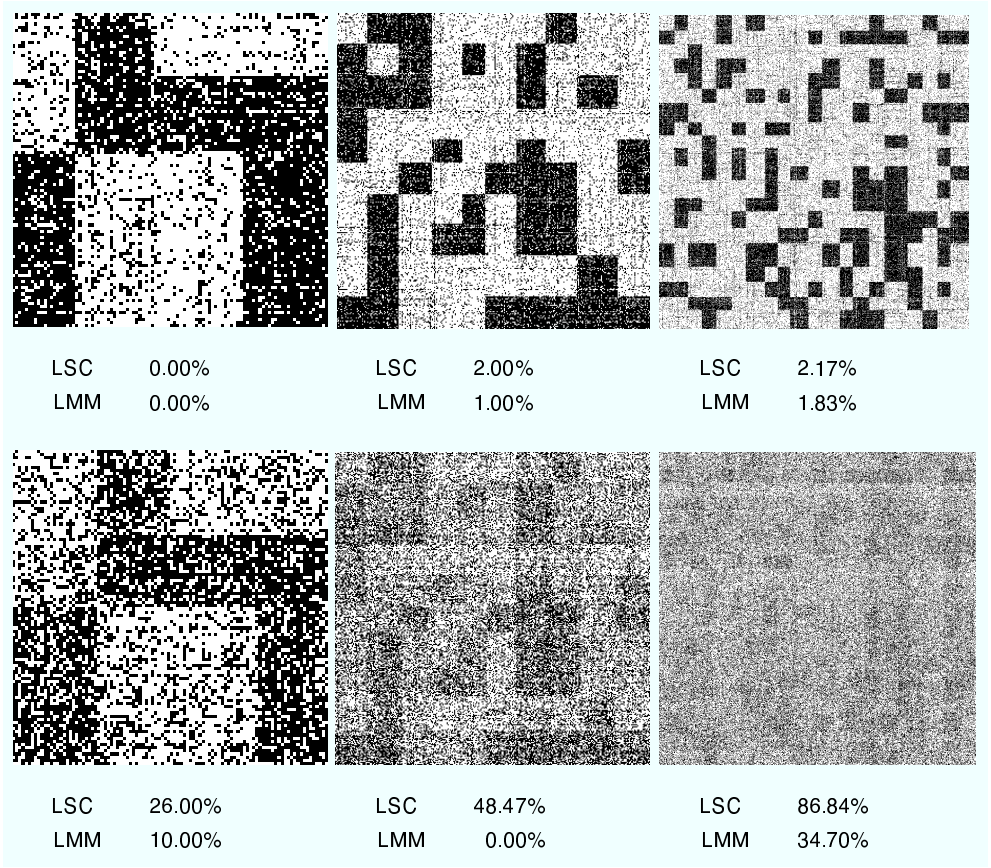

Figure 2. Error rates on simulated protein-protein interaction networks, the lower the better, for spectral clustering with local scaling (LSC) versus mixed-membership stochastic block model (MMSB). From left to right: the adjacency matrices contain 100, 300 and 600 proteins and 4, 10 and 20 latent functional groups, respectively. From top to bottom: the matrices were generated using Dirichlet parameter $\alpha = 0.05$ (stringent membership), 0.25 (more diffused membership), respectively. The proteins are re-ordered to make explicit the structure of the group interactions. The number of proteins per cluster averages 30 over all matrices. The Bernoulli probabilities in $\boldsymbol{\eta}$ are either 0.9 or 0.1. Random guesses about single-membership of proteins to clusters correspond to error rates of $0.75, 0.9$ and 0.95, respectively.

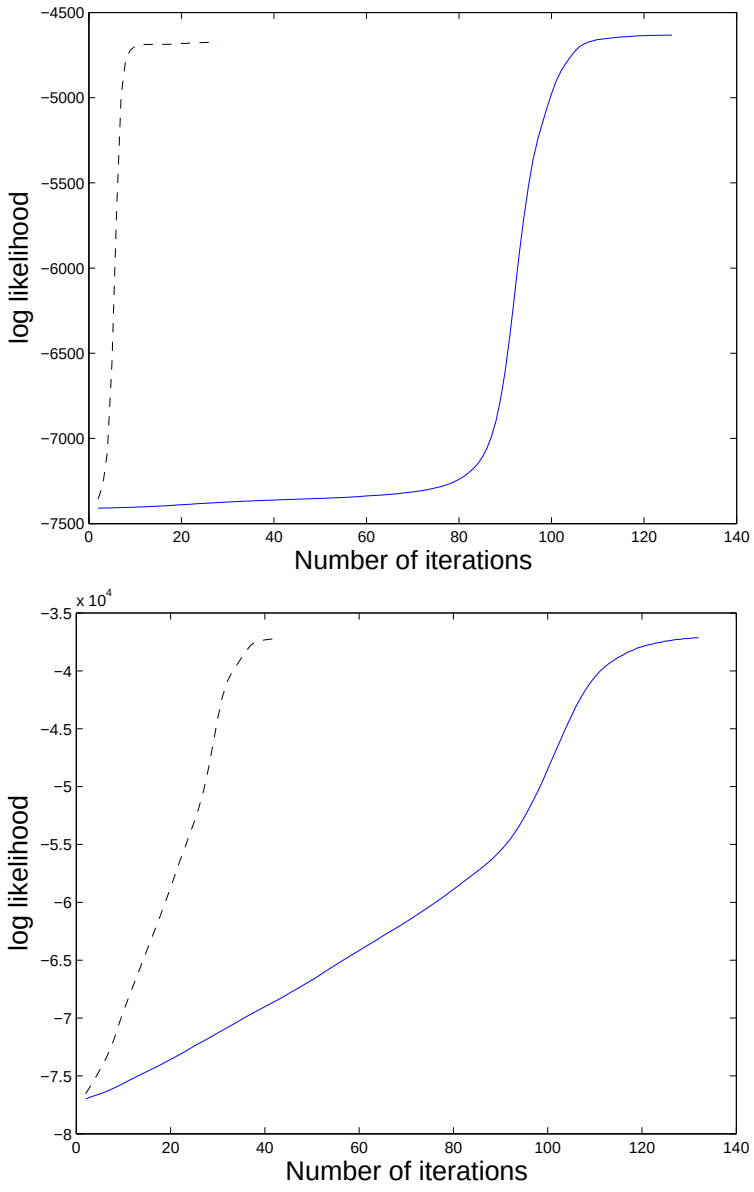

Figure 3. We compare the running time of the naïve variational inference (solid line) against the running time of our enhanced (nested) variational inference algorithm (dashed line), in two experiments. We measure the number of iterations on the X axis and the log-likelihood on the Y axis. The two profiles (iterations/log-likelihood) in each panel correspond to the same initial values for the parameters. Both algorithms reach the same plateau in terms of log-likelihood, which correspond to the same parameter estimates.

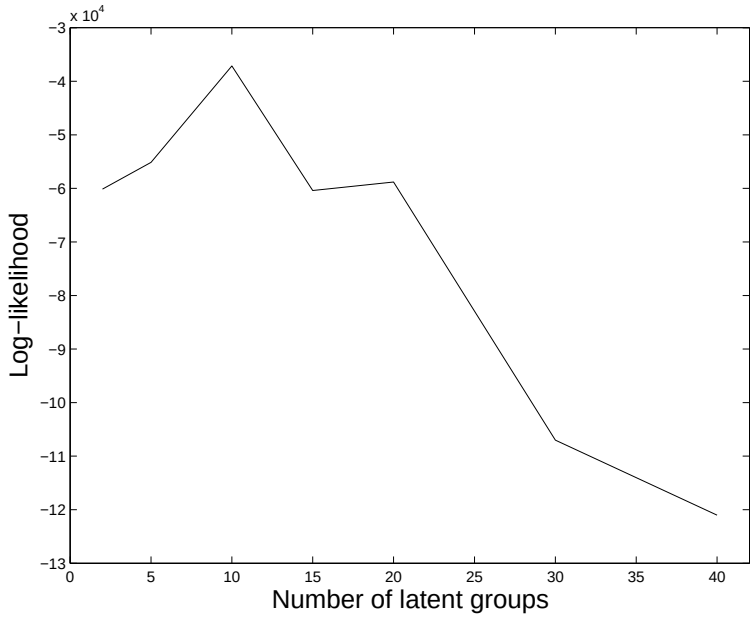

Figure 4. The log likelihood is indicative of the true number of latent functions, on simulated data. We measure the number of latent functions on the X axis and the log-likelihood on a test set on the Y axis. In the example shown the peak corresponds to the correct number of functions.

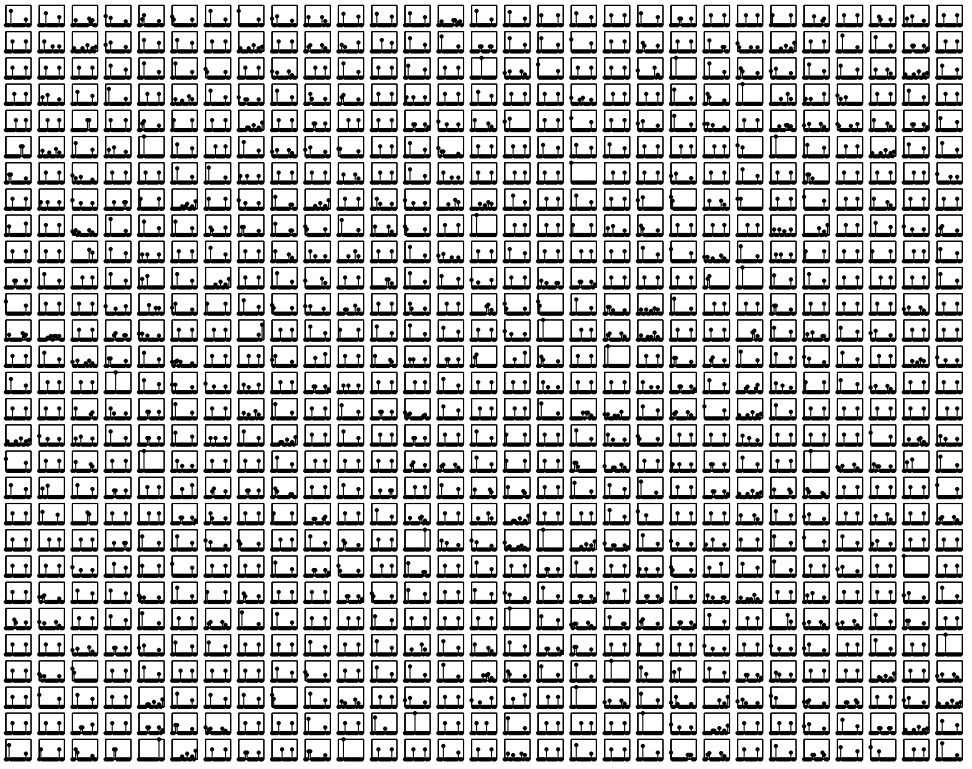

Figure 5. Manually curated functional annotations for 841 proteins in our data set: most proteins participate in at least two functions. Each panel corresponds to a protein. The values on the X axis range from 1 to 15, and are mapped to functions as in Table 1. The values on the Y axis correspond to normalized frequencies of participation of each protein in sub-processes of the 15 high-level functions.

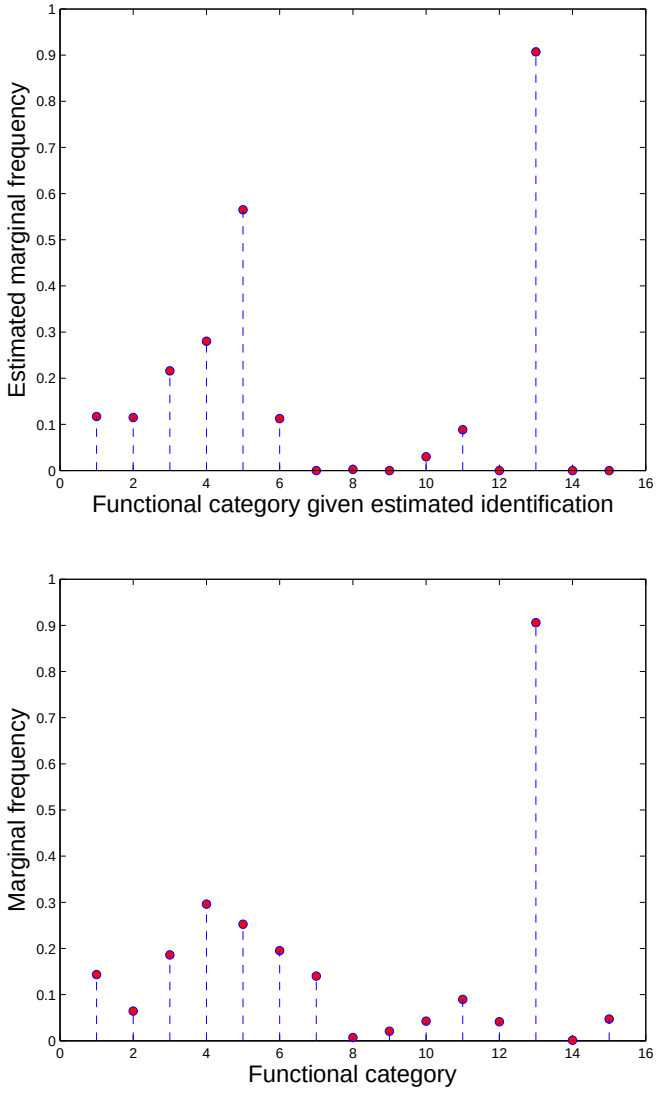

Figure 6. We estimate the mapping of latent groups to functions. The two plots show the marginal frequencies of membership of proteins to true functions (bottom) and to identified functions (top), in the cross-validation experiment. The mapping is selected to maximize the accuracy of the predictions on the training set, in the cross-validation experiment, and to minimize the divergence between marginal true and predicted frequencies if no training data is available.

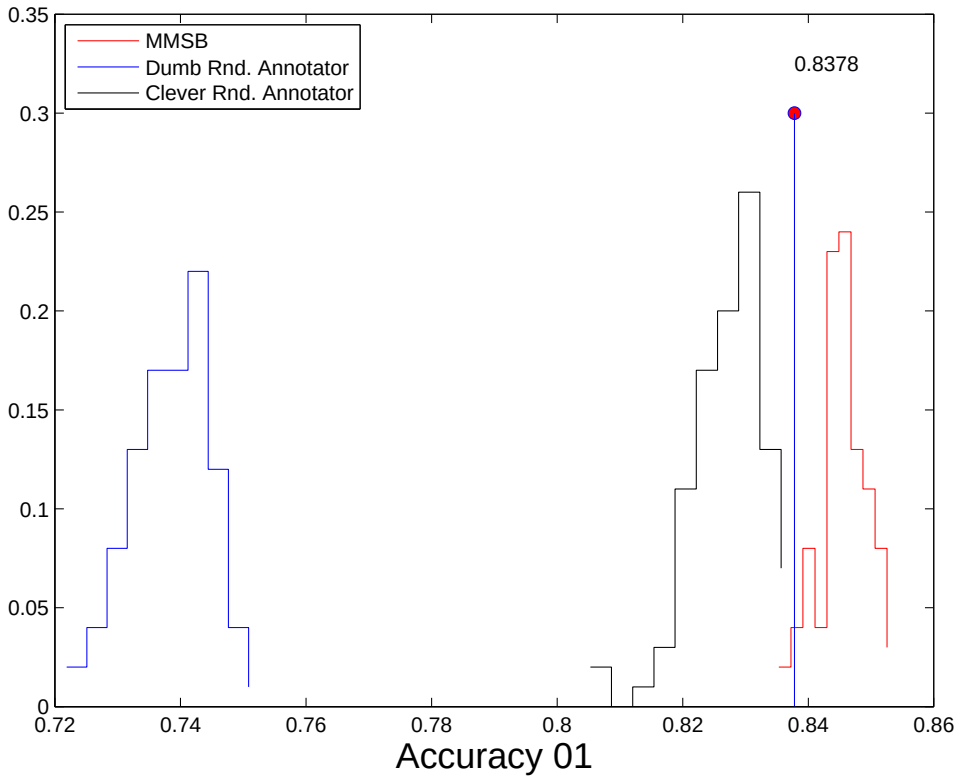

Figure 7. The accuracy of the predictions obtained with MMSB is 85% on average. Our method significantly outperforms both the dumb random annotator, 74% accurate on average, and the clever random annotator, 83% accurate on average. This suggests that our method leverages the signal in the data, and that the identification of latent groups to functions is somewhat feasible in this application.

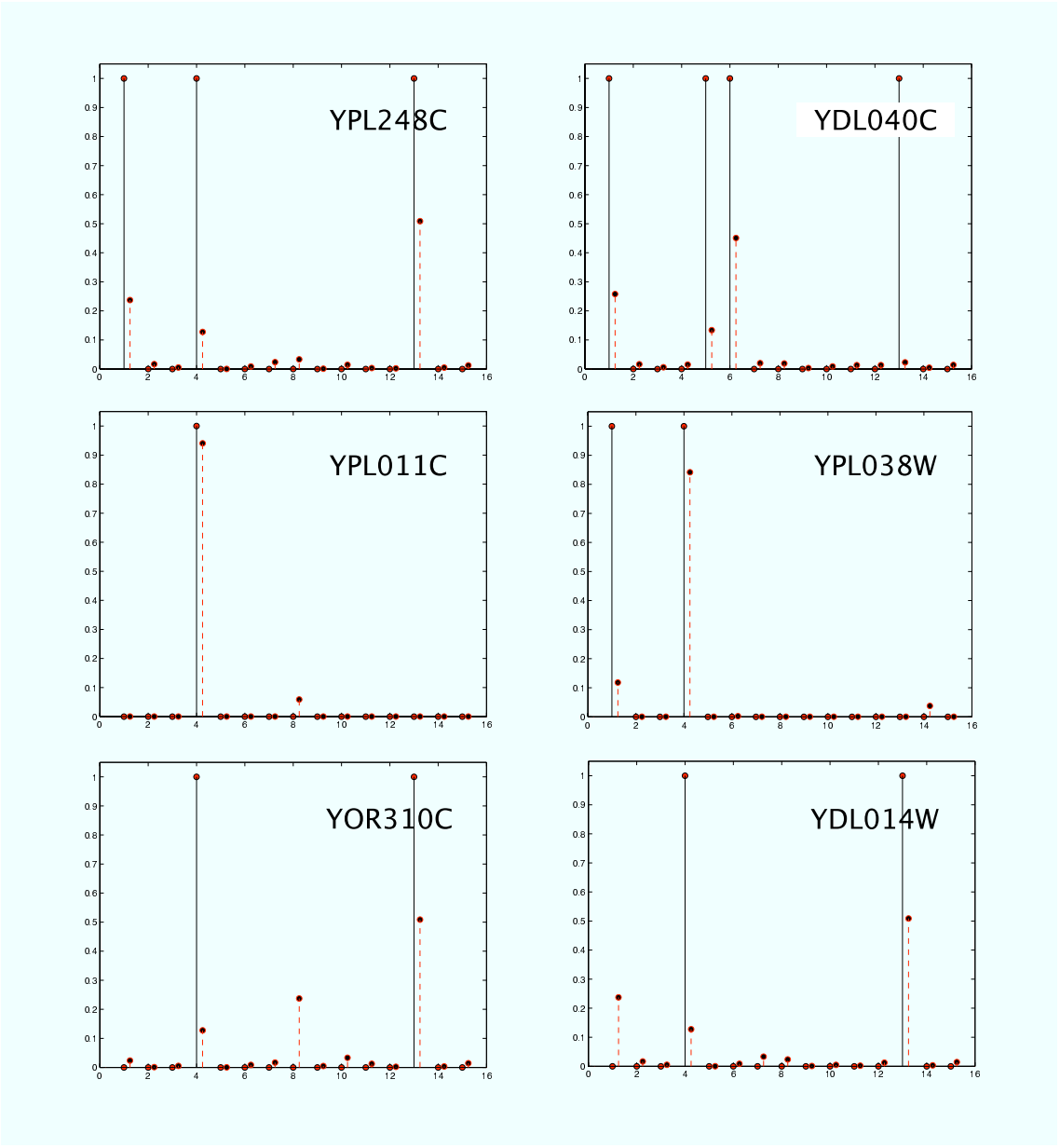

Figure 8. Predicted mixed-membership probabilities (dashed, red lines) versus binary manually curated functional annotations (solid, black lines) for 6 example proteins. The identification of latent groups to functions is estimated, and it is discussed in Figure 6.

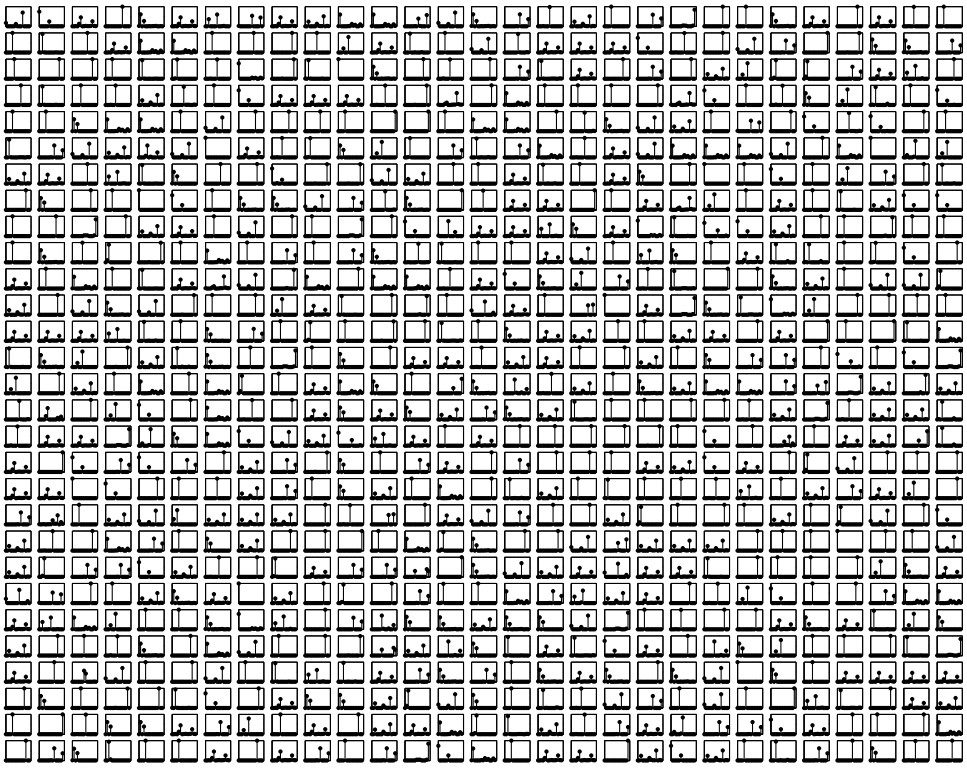

Figure 9. Predicted mixed membership scores for 841 proteins using our mixed-membership stochastic block model (MMSB): most proteins are predicted as participating in at least two functions. The predicted degree of membership is reasonably good, and the estimated Dirichlet parameter is $\hat{\alpha} = 0.417$. The values on the X axis range from 1 to 15, and are mapped to functions as in Table 1. The values on the Y axis correspond to normalized frequencies of participation of each protein in sub-processes of the 15 high-level functions.

Table 1

Functional Categories. In the table we report the functions proteins in the MIPS collection participate in. Most proteins participate in more than one function (≈ 2.4 on average) and, in the table, we added one count for each function each protein participates in.

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