Combining Stochastic Block Models and Mixed Membership for Statistical Network Analysis

Edoardo M. Airoldi[1,⋆], David M. Blei[2], Stephen E. Fienberg[1,3], and Eric P. Xing[1]

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213 USA
[2] Department of Computer Science, Princeton University, Princeton NJ 08540 USA
[3] Department of Statistics, Carnegie Mellon University, Pittsburgh PA 15213 USA
eairoldi@cs.cmu.edu

Abstract. Data in the form of multiple matrices of relations among objects of a single type, representable as a collection of unipartite graphs, arise in a variety of biological settings, with collections of author-recipient email, and in social networks. Clustering the objects of study or situating them in a low dimensional space (e.g., a simplex) is only one of the goals of the analysis of such data; being able to estimate relational structures among the clusters themselves may be important. In [1], we introduced the family of *stochastic block models of mixed membership* to support such integrated data analyses. Our models combine features of mixed-membership models and block models for relational data in a hierarchical Bayesian framework. Here we present a *nested* variational inference scheme for this class of models, which is necessary to successfully perform fast approximate posterior inference, and we use the models and the estimation scheme to examine two data sets. (1) a collection of sociometric relations among monks is used to investigate the crisis that took place in a monastery [2], and (2) data from a school-based longitudinal study of the health-related behaviors of adolescents. Both data sets have recently been reanalyzed in [3] using a latent position clustering model and we compare our analyses with those presented there.

1 Introduction

Relational information arise in a variety of settings, e.g., in scientific literature papers are connected by citation, in the word wide web the webpages are connected by hyperlinks, and in cellular systems the proteins are often related by physical protein-protein interactions revealed in yeast-two-hybrid experiments. These types of relational data violate the assumptions of independence or exchangeability of objects adopted in many conventional analyses. In fact, the relationships themselves between objects are often of interest in addition to the

⋆ To whom correspondence should be addressed, edo@cmu.edu.

E.M. Airoldi et al. (Eds.): ICML 2006 Ws, LNCS 4503, pp. 57–74, 2007.

object attributes. For example, one may be interested in predicting the citations of newly written papers or the likely links of a web-page, or in clustering cellular proteins based on patterns of interactions between them.

In many such applications, clustering the objects of study or projecting them in a low dimensional space (e.g., a simplex) is only one of the goals of the analysis. Being able to estimate the relational structures among the clusters themselves is often as important as object clustering. For example, from observations about email communications of a study population, one may be not only interested in identifying groups of people of common characteristics or social states, but also at the same time exploring how the overall communication volume or pattern among these groups can reveal the organizational structures of the population. A popular class of probabilistic models for relational data analysis are based on the stochastic block model (SBM) formalism for psychometric and sociological analysis pioneered by Holland and Leinhardt [4], and later extended in various contexts [5,6,7,8,9]. In machine learning, Markov random networks have been used for link prediction [10] and the traditional block models have been extended to include nonparametric Bayesian priors [11,12] and to integrate relations and text [13]. Typically, these models posit that every node in a study network is characterized by a unary *latent aspect* that accounts for its interaction patterns to peers in the networks; and conditioning on the observed network topology one can reason about these *latent aspects* of nodes via posterior inference.

Largely disjoint from the network analysis literature, methodologies for latent aspect modeling have also been widely investigated in the contexts of different informational retrieval problems concerning modeling the high-dimensional non-relational attributes such as text content or genetic-allele profile. In many of these domains, variants of a mixed membership formalism have been proposed to capture a more realistic assumption about the observed attributes, that the observations are resulted from contributions from multiple latent aspects rather than a unary aspects as assumed in most extant network models such as SBM. The mixed membership models have emerged as a powerful and popular analytical tool for analyzing large databases involving text [14], text and references [15,16], text and images [17], multiple disability measures [18,19], and genetics information [20,21,22]. These models often employ a simple generative model, such as a bag-of-words model or a naive Bayes, embedded in a hierarchical Bayesian framework involving a latent variable structure that combines multiples latents aspects. This scheme induces dependencies among the objects' relational behaviors in the form of probabilistic constraints over the estimation of what might otherwise be an extremely large set of parameters.

In modern network analysis tasks described above, it is desirable to also relax the unary-aspect assumption on each node imposed by extant models. We have proposed a new class of stochastic network models based the principle of *stochastic block models of mixed membership* [1], which combines features of the mixed-membership models [18] and the block models [23,24,25,9] via a hierarchical Bayesian framework, and offers a flexible machinery to capture rich semantic aspects of various network data. In this paper, we describe an instantiation of

this class of model, referred to as *admixture of latent blocks* (ALB) [26] to reasons to be explained shortly, for analyzing networks of objects with multiple latent roles (e.g., social activities in case the objects refer to people, or biological functions in case the objects refer to proteins). As mentioned above, classical network models such as the stochastic block models only allow each nodes to bear a single role. Our model alleviates this constraint, and furthermore posits that each nodes can adopt different roles when interacting with different other nodes.

Here is an outline of the rest of the paper. In Sections 2 we present the statistical formulation of the *Admixture of Latent Blocks* model (ALB). Then, in Section 3, we describe a variational inference algorithm for latent role inference and approximate maximum likelihood parameter estimation. In Section 4, we apply our model to two social networks widely studied in the literature, and we compare results of our analysis with that from a latent space model recently developed by Handcock, et al. [3].

2 The Statistical Model

We concern ourselves with modeling data represented as a collection of directed unipartite graphs. A unipartite graph is a graph whose nodes are of a single type, e.g., individual human beings in case of a person-to-person communication network, as opposed to bipartite and multipartite graphs, where the nodes are of two or multiple types (e.g., genes-to-experiments [14,27] or employees-to-tasks-to-resources [28]).

Let $G = (N, R)$ denote a graph with edge set R on node set N. We consider situations where we observe a collection of M unipartite graphs, $\mathcal{G} = \{G_m : r = 1, \ldots, M\}$ defined on a common set of nodes $\mathcal{N}$, of which the presence or absence of edges between node-pair i and j in graph G_m is denoted by variable $R_m(p, q)$. For example, in our experiment presented in the sequel, $\mathcal{N}$ corresponds a group of monks in a monestary [2], and $\{R_m(p, q)\}$ correspond to the relationships measured among these monks over a period. We observe typically asymmetric binary relations such as "Do you like X?", over a sequence of time.

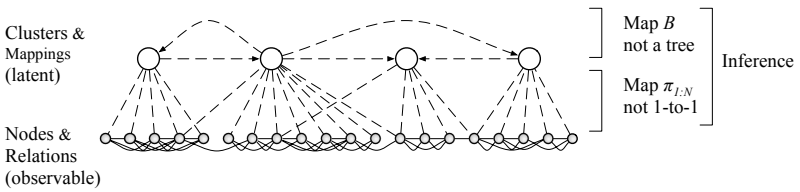

Fig. 1. The scientific problem at a glance. The goal of the analysis is to make inference on two mappings; nodes-to-clusters (via $\boldsymbol{\pi}_{1:N}$) and clusters-to-clusters (via B). The facts that B does not necessarily encode a tree, and that $\boldsymbol{\pi}_{1:N}$ is not necessarily one-to-one distinguish our formulation from typical hierarchical and hard clustering.

The analysis of such data typically focuses on the following objectives: (1) identifying clustering of nodes; (2) determining the number of clusters; and (3) estimating the probability distribution of interactions among actors within and between clusters. Back to the example of the monestary social network, objective 1 translates to identifying the solid factions among monks, In addition one wants to determine how many factions are likely to exist in the monastery, and how the factions relate to one another.

2.1 The Model

Our approach detailed below employs a hierarchical Bayesian formalism that encodes statistical assumptions underlying a network generative process. This process generates the observed networks according to the latent distribution of the hypothetical group-involvement of each monk, as specified by a mixed-mememberships multinomial vector $\pi := [\pi_1, \ldots, \pi_K]'$ where π_i denote the probability of a monk belonging to group i; and the probabilities of having interactions between different groups, as defined by a matrix of Bernoulli rates $B_{(K\times K)} = \{B_{ij}\}$ where B_{ij} represents the probability of having a link between a monk from group i and a monk from group j. Each monk is associated with a unique π, meaning that he can be simultaneously belonging to multiple groups, and the degree of involvements in different groups is unique for each monk; and π of different monks independently follow a Dirichlet distribution parameterized by α.

More generally, for graph m and each node, let indicator vector [1] $\boldsymbol{z}^m_{p\to q}$ denote the group membership of node p when it is to approach with node q; let $\boldsymbol{z}^m_{p\leftarrow q}$ denote the group membership of node q when it is approached by node p; let $N := |\mathcal{N}|$ denote the number of nodes in the graph; and let K denote the number of distinct groups a node can belong to. An admixture of latent blocks (ALB) model posit that a sequence of M networks can be instantiated according to the following procedure:

- For each node $p = 1, \ldots, N$:
  - $\boldsymbol{\pi}_p \sim \text{Dirichlet}\left(\alpha\right)$ sample a K dimensional *mixed membership* vector;
- for each network G_m, and each pair of nodes $(p, q) \in [1, N] \times [1, N]$ (denote p as the *initiator* and q as the *receiver*) in G_m:
  - $z^m_{p\to q} \sim \text{Multinomial}\left(\boldsymbol{\pi}_p\right)$ sample membership indicator for the initiator,
  - $z^m_{p\leftarrow q} \sim \text{Multinomial}\left(\boldsymbol{\pi}_q\right)$ sample membership indicator for the receiver,
  - $R_m(p, q) \sim \text{Bernoulli}\left(\boldsymbol{z}^{m\top}_{p\to q} B\, \boldsymbol{z}^m_{p\leftarrow q}\right)$ sample the value of their interaction.

It is noteworthy that in the above model, the group membership of each node is *context dependent*, that is, each nodes can assume different membership when interacting to or being interacted by different peers. Therefore, each node is statistically an admixture of group-specific interactions, and we denote the two sets of latent group indicators corresponding to the *m-th* observed network by

[1] An indicator vector of memberships in one of the K groups is defined as a K-dimensional vector of which only one element whose index corresponds to the id of the group to be indicated equals to one, and all other elements equal to zero.

$\{\boldsymbol{z}^m_{p\to q} : p, q \in \mathcal{N}\} =: Z^{\rightarrow}_m$ and $\{\boldsymbol{z}^m_{p\leftarrow q} : p, q \in \mathcal{N}\} =: Z^{\leftarrow}_m$. Marginalizing out the latent group indicators, it is easy to show that the probability of observing an interaction between node p and q across the M networks is $\bar{\sigma}_{pq} = \boldsymbol{\pi}_p{}^\top B\ \boldsymbol{\pi}_q$.

Under an ALB model outlined above, the joint probability distribution of the data, $R_{1:M}$, and the latent variables $(\boldsymbol{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M})$ can be written in the following factored form:

$$\begin{aligned} &p(R_{1:M}, \boldsymbol{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M} | \boldsymbol{\alpha}, B) \qquad (1)\\ &= \prod_m \prod_{p,q} P(R_m(p,q) | \boldsymbol{z}^m_{p\to q}, \boldsymbol{z}^m_{p\leftarrow q}, B) P(\boldsymbol{z}^m_{p\to q} | \boldsymbol{\pi}_p) P(\boldsymbol{z}^m_{p\leftarrow q} | \boldsymbol{\pi}_q) \prod_p p_3(\boldsymbol{\pi}_p | \boldsymbol{\alpha}). \end{aligned}$$

To compute the likelihood of the observed networks, one needs to marginalize out the hidden variables $\boldsymbol{\pi}$ and Z for all notes, which is intractable for even for small graphs. In §3, we describe a variational scheme to approximate this likelihood for parameter estimation.

2.2 Dealing with Sparsity

Most networks in real world are sparse, meaning that most pairs of nodes do not have edges connecting them. But in many network analyses, observations about interactions and non-interactions are equally important in terms of their contributions to model fitness. In other words, they would compete for a statistical explanation in terms of estimates for parameters $(\boldsymbol{\alpha}, B)$, and would both influence the distribution of latent variables such as $\boldsymbol{\pi}_{1:N}$. A non desirable consequence of this, in scenarios where interactions are rare, is that parameter estimation and posterior inference would explain patterns of non-interaction rather than patterns of interaction.

In order to be able to calibrate the importance of rare interactions, we introduce the sparsity parameter $\rho \in [0, 1]$, which models how often a non-interaction is due to measurement noise (which is common in certain experimentally derived networks such as the protein-protein interaction networks) and how often it carries information about the group memberships of the nodes. This leads to a small extension of the generative process outlined in the last subsection. Specifically, instead of drawing an edge directly from a Bernoulli with rate $\boldsymbol{z}^{m\ \top}_{p\to q} B\ \boldsymbol{z}^m_{p\leftarrow q}$, now we sample an interaction with probability $\sigma^m_{pq} = (1-\rho) \cdot \boldsymbol{z}^{m\ \top}_{p\to q} B\ \boldsymbol{z}^m_{p\leftarrow q}$; therefore the probability of having no interaction this pair of nodes is $1 - \sigma^m_{pq} = (1-\rho) \cdot \boldsymbol{z}^{m\ \top}_{p\to q}(1-B)\ \boldsymbol{z}^m_{p\leftarrow q} + \rho$. This is equivalent to re-parameterizing the interaction matrix B. During estimation and inference, a large value of ρ would cause the interactions in the matrix to be weighted more than non-interactions in determining the estimates of $(\boldsymbol{\alpha}, B, \boldsymbol{\pi}_{1:N})$.

3 Parameter Estimation and Posterior Inference

We use an empirical Bayes framework for estimating the parameters $(\boldsymbol{\alpha}, B)$, and employ a mean-field approximation scheme [29] for posterior inference of the

(latent) mixed-membership vectors, $\boldsymbol{\pi}_{1:N}$. Model selection can be performed to determine the plausible value of K—the number of groups of nodes—based on a strategy described in [30].

In order to estimate $(\boldsymbol{\alpha}, B)$ and infer the posterior distributions of $\boldsymbol{\pi}_{1:N}$ we need to be able to evaluate the likelihood, which involves the non-tractable integral over Z and $\boldsymbol{\pi}_{1:N}$ in Equation 1. Given the large amount of data available for most networks, we focus on approximate posterior inference strategies in the context of variational methods, and we find a tractable lower bound for the likelihood that can be used as a surrogate for inference purposes. This leads to approximate MLEs for the hyper-parameters and approximate posterior distributions for the (latent) mixed-membership vectors.

3.1 Lower Bound for the Likelihood

According to the mean-field theory [29,31], one can approximate an intractable distribution such as the one defined by Equation (1) by a fully factored distribution $q(\boldsymbol{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M})$ defined as follows:

$$\begin{aligned} & q(\boldsymbol{\pi}_{1:N}, Z^{\rightarrow}_{1:M}, Z^{\leftarrow}_{1:M} | \boldsymbol{\gamma}_{1:N}, \Phi^{\rightarrow}_{1:M}, \Phi^{\leftarrow}_{1:M}) \\ = & \prod_{p} q_1(\boldsymbol{\pi}_p|\boldsymbol{\gamma}_p) \prod_{m} \prod_{p,q} \Big(q_2(\boldsymbol{z}^m_{p\rightarrow q}|\boldsymbol{\phi}^m_{p\rightarrow q}, 1)\ q_2(\boldsymbol{z}^m_{p\leftarrow q}|\boldsymbol{\phi}^m_{p\leftarrow q}, 1) \Big), \end{aligned} \tag{2}$$

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

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

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

Working on the single expectations leads to the following expression,

$$\begin{aligned} \mathcal{L}_{\Delta}(q, \Theta) = & \sum_{m} \sum_{p,q} \sum_{g,h} \phi^m_{p\rightarrow q,g} \phi^m_{p\leftarrow q,h} \cdot f\ \big(R_m(p,q), B(g,h) \big) \\ & + \sum_{m} \sum_{p,q} \sum_{g} \phi^m_{p\rightarrow q,g} \left[\psi(\gamma_{p,g}) - \psi(\sum_{g} \gamma_{p,g}) \right] \\ & + \sum_{m} \sum_{p,q} \sum_{h} \phi^m_{p\leftarrow q,h} \left[\psi(\gamma_{p,h}) - \psi(\sum_{h} \gamma_{p,h}) \right] \end{aligned}$$

$$+ \sum_p \log \Gamma(\sum_k \alpha_k) - \sum_{p,k} \log \Gamma(\alpha_k) + \sum_{p,k} (\alpha_k - 1) \left[\psi(\gamma_{p,k}) - \psi(\sum_k \gamma_{p,k}) \right]$$
$$- \sum_p \log \Gamma(\sum_k \gamma_{p,k}) + \sum_{p,k} \log \Gamma(\gamma_{p,k}) - \sum_{p,k} (\gamma_{p,k} - 1) \left[\psi(\gamma_{p,k}) - \psi(\sum_k \gamma_{p,k}) \right]$$
$$- \sum_m \sum_{p,q} \sum_g \phi^m_{p \to q,g} \log \phi^m_{p \to q,g} - \sum_m \sum_{p,q} \sum_h \phi^m_{p \leftarrow q,h} \log \phi^m_{p \leftarrow q,h}$$

where

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

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, $\frac{d \log \Gamma(x)}{dx}$.

3.2 The Expected Value of the Log of a Dirichlet Random Vector

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

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

with natural parameters $\theta_k(\boldsymbol{\alpha}) = (\alpha_k - 1)$ and natural sufficient statistics $t_k(\boldsymbol{\pi}) = \log(\pi_k)$. Let $c'(\boldsymbol{\theta}) = c(\alpha_1(\boldsymbol{\theta}), \ldots, \alpha_K(\boldsymbol{\theta}))$; using a well known property of the exponential family distributions [32] we find that

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

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

3.3 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 [33].

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 $\boldsymbol{g}^m_{p \to q}$ and $\boldsymbol{g}^m_{p \leftarrow q}$ corresponding to the

natural sufficient statistics $\log(\boldsymbol{z}^m_{p\to q})$ and $\log(\boldsymbol{z}^m_{p\leftarrow q})$ are functions of the other latent variables and the observations. We find that

$$g^m_{p\to q,g} = \log \pi_{p,g} + \sum_h z^m_{p\leftarrow q,h} \cdot f\ \big(\ R_m(p,q), B(g,h)\ \big),$$
$$g^m_{p\leftarrow q,h} = \log \pi_{q,h} + \sum_g z^m_{p\to q,g} \cdot f\ \big(\ R_m(p,q), B(g,h)\ \big),$$

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

$$f\ \big(\ R_m(p,q), B(g,h)\ \big) = R_m(p,q)\log B(g,h) + \big(\ 1 - R_m(p,q)\ \big)\ \log\ \big(\ 1 - B(g,h)\ \big).$$

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

$$\begin{aligned}
\hat{\phi}^m_{p\to q,g} &\propto e^{\ \mathbb{E}_q\left[g^m_{p\to q,g}\right]} \qquad (3)\\
&= e^{\ \mathbb{E}_q\left[\log \pi_{p,g}\right]} \cdot e^{\ \sum_h \phi^m_{p\leftarrow q,h}\cdot\ \mathbb{E}_q\left[f\left(R_m(p,q),B(g,h)\right)\right]}\\
&= e^{\ \mathbb{E}_q\left[\log \pi_{p,g}\right]} \cdot \prod_h \left(\ B(g,h)^{R_m(p,q)} \cdot\ \big(\ 1 - B(g,h)\ \big)^{1-R_m(p,q)}\right)^{\phi^m_{p\leftarrow q,h}}\\
\hat{\phi}^m_{p\leftarrow q,h} &\propto e^{\ \mathbb{E}_q\left[g^m_{p\leftarrow q,h}\right]} \qquad (4)\\
&= e^{\ \mathbb{E}_q\left[\log \pi_{q,h}\right]} \cdot e^{\ \sum_g \phi^m_{p\to q,g}\cdot\ \mathbb{E}_q\left[f\left(R_m(p,q),B(g,h)\right)\right]}\\
&= e^{\ \mathbb{E}_q\left[\log \pi_{q,h}\right]} \cdot \prod_g \left(\ B(g,h)^{R_m(p,q)} \cdot\ \big(\ 1 - B(g,h)\ \big)^{1-R_m(p,q)}\right)^{\phi^m_{p\to q,g}}
\end{aligned}$$

for $g,h = 1,\ldots,K$. These estimates of the parameters underlying the distribution of the nodes' group indicators $\boldsymbol{\phi}^m_{p\to q}$ and $\boldsymbol{\phi}^m_{p\leftarrow q}$ 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:

$$\hat{\gamma}_{p,k} = \alpha_k + \sum_m \sum_q \phi^m_{p\to q,k} + \sum_m \sum_q \phi^m_{p\leftarrow q,k}, \qquad (5)$$

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 $\Theta^{(t-1)} = (\boldsymbol{\alpha}^{(t-1)}, B^{(t-1)})$, and finds the optimal approximate lower bound for the likelihood $\mathcal{L}_{\Delta^*}(q,\Theta^{(t-1)})$.

3.4 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.[2] 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 $\boldsymbol{\alpha}$ we obtain $\mathcal{L}_{\boldsymbol{\alpha}}(q, \Theta)$. Unfortunately, a closed form solution for the approximate maximum likelihood estimate of $\boldsymbol{\alpha}$ does not exist [14]. We can produce a Newton-Raphson method that is linear in time, where the gradient and Hessian for the bound $\mathcal{L}_{\boldsymbol{\alpha}}$ are

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

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

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

$$\hat{B}(g,h) = \frac{1}{M} \sum_m \left(\frac{\sum_{p,q} R_m(p,q) \cdot \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh}}{\sum_{p,q} \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh}} \right), \tag{6}$$

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

In Section 2.2 we introduced an extra parameter, ρ, to control the relative importance of presence and absence of interactions in likelihood, i.e., 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

$$\hat{\rho} = \frac{1}{M} \sum_m \left(\frac{\sum_{p,q} \left(1 - R_m(p,q) \right) \cdot \left(\sum_{g,h} \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh} \right)}{\sum_{p,q} \sum_{g,h} \phi^m_{p \to qg} \, \phi^m_{p \leftarrow qh}} \right). \tag{7}$$

Alternatively, we can fix ρ prior to the analysis; the density of the interaction matrix is estimated with $\hat{d} = \sum_{m,p,q} R_m(p,q)/(N^2 M)$, 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 $\boldsymbol{\pi}_{1:N}$. It does however provide a quick recipe to reduce the computational burden during exploratory analyses.[3]

3.5 Smoothing

In problems where the number of clusters is deemed to be likely large a-priori, we can smooth the (consequently large number of) cluster-to-cluster relation probabilities encoded in the block model B by positing that all the elements $B(g,h)$ of the block model are non-observable samples from a common (prior) distribution. In the admixture of latent blocks model we posit that $p(B|\boldsymbol{\lambda})$ is a collection non-symmetric beta distributions, with a pair of hyper-parameters $\boldsymbol{\lambda}$ common to all elements of B.

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

[3] 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.

3.6 The Nested Variational EM Algorithm

The complete algorithm to perform variational inference in the model is described in detail in Figure 2. To achieve fast convergence, we employed a highly effective *nested* variational inference scheme based on a non-trivial scheduling of variational parameters updating. The resulting algorithm is also parallelizable on a computer cluster.

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

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

Fig. 2. Top: The two-layered variational inference for $(\boldsymbol{\gamma}, \phi_{p \to q,g}, \phi_{p \leftarrow q,h})$ 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 $\boldsymbol{\gamma}$ and B refer to Equation 5 of Section 3.3 and Equation 6 of Section 3.4, respectively. **Bottom:** Inference for the variational parameters $(\boldsymbol{\phi}_{p \to q}, \boldsymbol{\phi}_{p \leftarrow q})$ corresponding to the basic observation $R(p,q)$. 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 3 and 4 of Section 3.3.

In a naïve iteration scheme for variational inference, one would initialize the variational Dirichlet parameters $\boldsymbol{\gamma}_{1:N}$ and the variational multinomial parameters $(\boldsymbol{\phi}_{p \to q}, \boldsymbol{\phi}_{p \leftarrow q})$ to non-informative values, and then iterate until convergence the following two steps: (i) update $\boldsymbol{\phi}_{p \to q}$ and $\phi_{p \leftarrow q}$ for all edges (p, q), and (ii) update $\boldsymbol{\gamma}_p$ 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 [1] the naïve variational algorithm often failed to converge, or converged after a large number of iterations. We attribute this behavior to a dependence that our two main

assumptions (block model and mixed membership) induce between $\boldsymbol{\gamma}_{1:N}$ and B, which is not satisfied by the naïve algorithm. Some intuition about why this may happen follows. From a purely algorithmic perspective, the naïve variational EM algorithm instantiates a large coordinate ascent algorithm, where the parameters can be semantically divided into coherent blocks. Blocks are processed in a specific order, and the parameters within each block get all updated each time.[4] At every new iteration the naïve algorithm sets all the elements of $\boldsymbol{\gamma}_{1:N}^{t+1}$ equal to the same constant. This dampens the likelihood by suddenly breaking the dependence between the estimates of parameters in $\widehat{\boldsymbol{\gamma}}_{1:N}^{t}$ 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 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, $(\boldsymbol{\phi}_{p\to q}, \boldsymbol{\phi}_{p\leftarrow q})$, optimized given the other variational parameters. Note that these parameters are involved in the updates of parameters in $\boldsymbol{\gamma}_{1:N}$ and in B, thus providing us with a channel to maintain some of the dependence among them, i.e., by keeping them at their optimal value given the data. Further, the nested algorithm has the advantage that it trades time for space thus allowing us to deal with large graphs; at each variational cycle we need to allocate $NK + 2K$ scalars 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. This algorithm is also better than MCMC variations (i.e., blocked and collapsed Gibbs samplers) in terms of memory requirements and convergence rates.

4 Experiments: Applications to Social Networks

We illustrate our model and algorithm in the context of two examples that have recently been reanalyzed in [3] using a *latent position clustering model* and [34].

4.1 Example 1: Crisis in a Cloister

Sampson [2] 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. Sampson's original analysis strongly suggested the existence of tight factions among the novices, and the events that took place during his stay at the monastery supported his observations. Briefly, novices of one faction left the monastery or were expelled over religious differences. The factions identified by Sampson provide a credible gold standard, to which we compare our results.

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

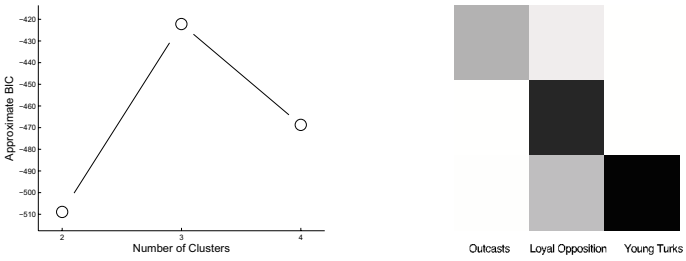

Fig. 3. The approximate BIC (left panel) suggests the relations among monks are best explained by a model with three factions. The faction-to-faction estimated relational patterns $\hat{B}$ (right panel) suggest that the Outcasts are an isolated faction, whereas Young Turks *like* members of the Loyal Opposition, although the sentiment is not reciprocated.

We consider Breiger's collation of Sampson's data [35]. Briefly, for each of the four sociometric relations above, only the top three choices of each novice were recorded as positive relations—the edges in the graph. We use the following approximation to BIC for model selection:

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

where $|\boldsymbol{\alpha}, B|$ is the number of hyper-parameters in the model, and $|R|$ is the number of positive relations observed—following arguments in [3]. The approximate BIC value suggests that the relations among monks in the monastery studied by Sampson are best explained by a model with three factions, independently of the number of hyper-parameters in the ALB model we fit. Hence we fixed $\hat{K} = 3$ in subsequent analyses, which involved ALB models with increasing degree of complexity. In the left panel of Figure 3 we show the approximate BIC for a model with a single hyper-parameter, α scalar. In the right panel of Figure 3 we show the estimated faction-to-faction block model, $\widehat{B}$, corresponds to a full model (i.e., no constraints on B). This estimate suggests that the Outcasts are an isolated faction, whereas Young Turks *like* members of the Loyal Opposition, although the sentiment is not reciprocated. In Figure 5 we investigate the the posterior means of the mixed membership scores, $\mathbb{E}[\boldsymbol{\pi}|R]$, for the 18 monks in the monastery ($\alpha = 0.058$ scalar, $B := \mathbb{I}_3$). We have a panel for each monk, and the subscripts associated with the names of the monks specify the order according to which they left the monastery, e.g., John left first. The three factions on the X axis are the Outcast, the Young Turks , and the Loyal Opposition (from left to right); and on the Y axis we measure the degree of membership of monks to factions. From these panels, the centrality of the role played by John and Greg, first to leave the monastery, as well as the uncertain affiliations of Romul, and Victor to a minor extent, unequivocally emerge. The mixed membership vectors, $\boldsymbol{\pi}_{1:18}$, provide us with low-dimensional representations of monks. In Figure 6 we plot them in their natural space, that

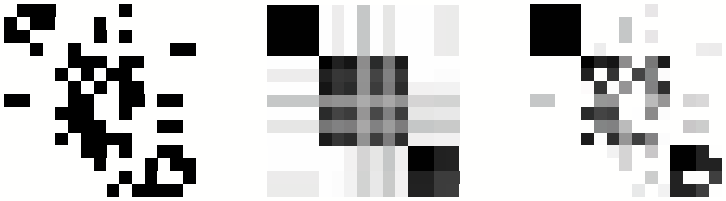

Fig. 4. Original matrix of sociometric relations (left), and estimated relations obtained by thresholding the posterior expectations $\boldsymbol{\pi}_p{}' B\, \boldsymbol{\pi}_q | R$ (center), and $\boldsymbol{\phi}_p{}' B\, \boldsymbol{\phi}_q | R$ (right)

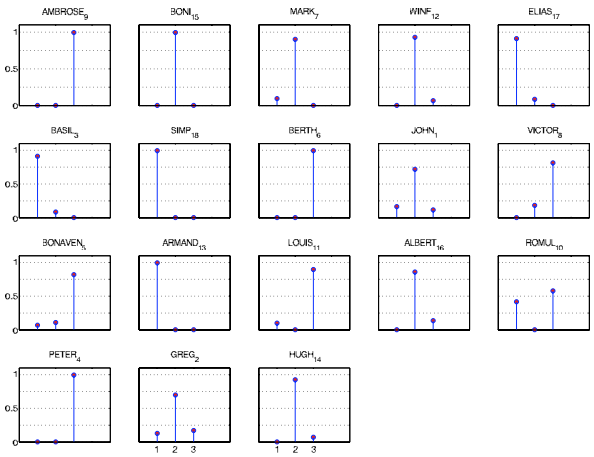

Fig. 5. The posterior mixed membership scores, $\boldsymbol{\pi}$, for the 18 monks in the monastery. Each panel correspond to a monk; on the Y axis we measure the grade of membership, corresponding to the Outcast (left bar), to the Young Turks (center bar), and to the Loyal Opposition (right bar), on the X axis. The subscripts associated with the names of the monks specify the order according to which they left the monastery.

is, the(3-dimensional) simplex. Dots correspond to monks; the red circles were obtained by fixing $B = \mathbb{I}_3$ and $\alpha = 0.01$, whereas the blue triangles correspond to fixing $B := \mathbb{I}_3$, but estimating $\hat{\alpha} = 0.058$.

4.2 Example 2: Health-Related Behaviors of Adolescents

The National Longitudinal Study of Adolescent Health [36,37] includes questionnaire administered to a sample of students, who were allowed to nominate up to 10 friends. Following [3], we focus on friendship nominations collected among 71 students in grades 7 to 12 at one school. Two students did not nominate any friends, so we analyzed the network of (binary, asymmetric) friendship relations

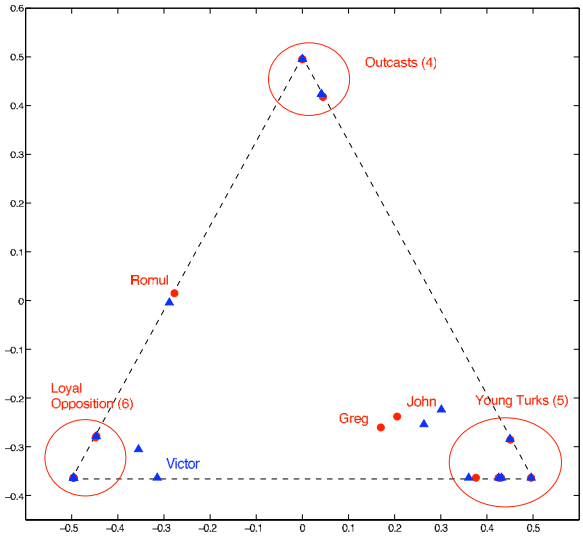

Fig. 6. Mixed membership vectors, $\boldsymbol{\pi}_{1:18}$, plotted in the simplex. Points correspond to monks; the red circles correspond to an ALB model with ($B = \mathbb{I}_3, \alpha = 0.01$), whereas the blue triangles correspond to an ALB model with ($B := \mathbb{I}_3, \hat{\alpha} = 0.058$).

among the remaining 69 students. The left panel of Figure 8 shows the raw relations, and we contrast this to the estimated networks in the central and right panels based on our model estimates using the full model. We proceeded with the analysis as in the previous study, but we fitted a full model in this case. Salient features of the analysis are: (i) the posterior mixed membership of the 69 students—shown in Figure 7; (ii) the correspondence of latent clusters to student grade levels—shown in Table 1; and (iii) the hyper-parameters were estimated with an empirical Bayes strategy; we obtained $\hat{\alpha} = 0.0487$, $\hat{\rho} = 0.936$, and a practically diagonal matrix that encodes the cluster-to-cluster relations,

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

4.3 Discussion

There is a tight relationship between ALB and the latent space models in [8,3]. In the latent space models, the latent vectors are drawn from Gaussian distributions and the interaction data is drawn from a Gaussian with mean $\boldsymbol{\pi}_p{}' \mathbb{I} \boldsymbol{\pi}_q$. In ALB, the marginal probability of an interaction takes a similar form, $\boldsymbol{\pi}_p{}' B \boldsymbol{\pi}_q$,

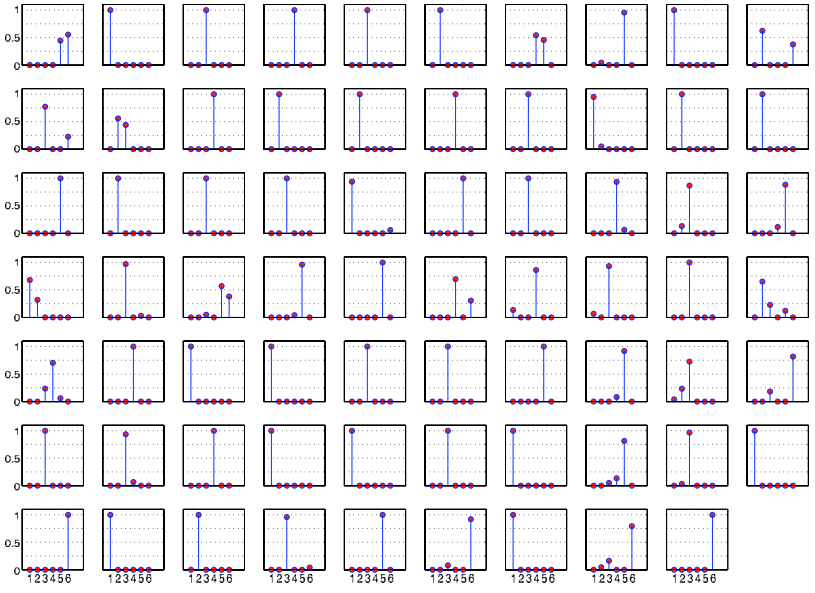

Fig. 7. The posterior mixed membership scores, $\boldsymbol{\pi}$, for the 69 students in a school. Each panel correspond to a student; on the Y axis we measure the grade of membership, corresponding to the six grade levels from 7 to 12, on the X axis.

Fig. 8. Original matrix of friensdhip relations (left), and estimated relations obtained by thresholding the posterior expectations $\boldsymbol{\pi}_p{}' B\, \boldsymbol{\pi}_q | R$ (center), and $\boldsymbol{\phi}_p{}' B\, \boldsymbol{\phi}_q | R$ (right)

where B is the matrix of probabilities of interactions for each pair of latent factions. In contrast to the latent space model, the relations can be modeled by an arbitrary distribution, in our model. With binary relations we can use a collection of Bernoulli parameters; with continuous relations, we can use a collection of Gaussian parameters. While more flexible, ALB does not subsume latent space models; they make different assumptions about the data.

Table 1. Grade levels versus (highest) expected posterior membership

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

Acknowledgments

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

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Stochastic block models of mixed membership. Manuscript under review (2006)
2. Sampson, F.S.: A Novitiate in a period of change: An experimental and case study of social relationships. PhD thesis, Cornell University (1968)
3. Handcock, M.S., Raftery, A.E., Tantrum, J.M.: Model-based clustering for social networks. Journal of the Royal Statistical Society, Series A **170** (2007) 1–22
4. Holland, P.W., Leinhardt, S.: Local structure in social networks. In Heise, D., ed.: Sociological Methodology, Jossey-Bass (1975) 1–45
5. Fienberg, S.E., Meyer, M.M., Wasserman, S.: Statistical analysis of multiple sociometric relations. Journal of the American Statistical Association **80** (1985) 51–67
6. Wasserman, S., Pattison, P.: Logit models and logistic regression for social networks: I. an introduction to markov graphs and p^*. Psychometrika **61** (1996) 401–425
7. Snijders, T.A.B.: Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure (2002)
8. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. Journal of the American Statistical Association **97** (2002) 1090–1098
9. Doreian, P., Batagelj, V., Ferligoj, A.: Generalized Blockmodeling. Cambridge University Press (2004)
10. Taskar, B., Wong, M.F., Abbeel, P., Koller, D.: Link prediction in relational data. In: Neural Information Processing Systems 15. (2003)
11. Kemp, C., Griffiths, T.L., Tenenbaum, J.B.: Discovering latent classes in relational data. Technical Report AI Memo 2004-019, MIT (2004)
12. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: Proceedings of the 21st National Conference on Artificial Intelligence. (2006)

13. McCallum, A., Wang, X., Mohanty, N.: 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)
14. Blei, D.M., Ng, A., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
15. Cohn, D., Hofmann, T.: The missing link—A probabilistic model of document content and hypertext connectivity. In: Advances in Neural Information Processing Systems 13. (2001)
16. Erosheva, E.A., Fienberg, S.E., Lafferty, J.: Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences **97**(22) (2004) 11885–11892
17. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research **3** (2003) 1107–1135
18. Erosheva, E.A., Fienberg, S.E.: Bayesian mixed membership models for soft clustering and classification. In Weihs, C., Gaul, W., eds.: Classification—The Ubiquitous Challenge. Springer-Verlag (2005) 11–26
19. Manton, K.G., Woodbury, M.A., Tolley, H.D.: Statistical Applications Using Fuzzy Sets. Wiley (1994)
20. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W.: Genetic structure of human populations. Science **298** (2002) 2381–2385
21. Pritchard, J., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics **155** (2000) 945–959
22. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with applications to clustering with side information. In: Advances in Neural Information Processing Systems. Volume 16. (2003)
23. Holland, P., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: Some first steps. Social Networks **5** (1983) 109–137
24. Anderson, C.J., Wasserman, S., Faust, K.: Building stochastic blockmodels. Social Networks **14** (1992) 137–161
25. Nowicki, K., Snijders, T.A.B.: Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association **96** (2001) 1077–1087
26. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Admixtures of latent blocks with application to protein interaction networks. Manuscript under review (2006)
27. Airoldi, E.M., Fienberg, S.E., Xing, E.P.: Latent aspects analysis for gene expression data. Manuscript under review (2006)
28. Carley, K.M.: Smart agents and organizations of the future. In Lievrouw, L., Livingstone, S., eds.: The Handbook of New Media. (2002) 206–220
29. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: Introduction to variational methods for graphical models. Machine Learning **37** (1999) 183–233
30. Airoldi, E.M., Fienberg, S.E., Joutard, C., Love, T.M.: Discovering latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. Technical Report CMU-ML-06-101, School of Computer Science, Carnegie Mellon University (2006)
31. Xing, E.P., Jordan, M.I., Russell, S.: A generalized mean field algorithm for variational inference in exponential families. In: Uncertainty in Artificial Intelligence. Volume 19. (2003)
32. Schervish, M.J.: Theory of Statistics. Springer (1995)
33. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley (2003)

34. David, G.B., Carley, K.M.: Clearing the FOG: Fuzzy, overlapping groups for social networks. Manuscript under review (2006)
35. Breiger, R.L., Boorman, S.A., Arabie, P.: An algorithm for clustering relational data with applications to social network analysis and comparison to multidimensional scaling. Journal of Mathematical Psychology **12** (1975) 328–383
36. Harris, K.M., Florey, F., Tabor, J., Bearman, P.S., Jones, J., Udry, R.J.: The national longitudinal study of adolescent health: research design. Technical report, Caorlina Population Center, University of North Carolina, Chapel Hill (2003)
37. Udry, R.J.: The national longitudinal study of adolescent health: (add health) waves i and ii, 1994–1996; wave iii 2001–2002. Technical report, Carolina Population Center, University of North Carolina, Chapel Hill (2003)