# INFINITE MULTIPLE MEMBERSHIP RELATIONAL MODELING FOR COMPLEX NETWORKS

*Morten Mørup, Mikkel N. Schmidt and Lars Kai Hansen*

Section for Cognitive Systems, DTU Informatics, {mm,mns,lkh}@imm.dtu.dk

**ABSTRACT**

Learning latent structure in complex networks has become an important problem fueled by many types of networked data originating from practically all fields of science. In this paper, we propose a new non-parametric Bayesian multiple-membership latent feature model for networks. Contrary to existing multiple-membership models that scale quadratically in the number of vertices the proposed model scales linearly in the number of links admitting multiple-membership analysis in large scale networks. We demonstrate a connection between the single membership relational model and multiple membership models and show on "real" size benchmark network data that accounting for multiple memberships improves the learning of latent structure as measured by link prediction while explicitly accounting for multiple membership result in a more compact representation of the latent structure of networks.

## 1. INTRODUCTION

The analysis of complex networks has become an important challenge spurred by the many types of networked data arising in practically all fields of science. These networks are very different in nature ranging from biology networks such as protein interaction [1, 2] and the connectome of neuronal connectivity [3] to the analysis of interaction between large groups of agents in social and technology networks [4, 3, 5, 6]. Many of the networks exhibit a strong degree of structure; thus, learning this structure facilitates both the understanding of network dynamics, the identification of link density heterogeneities, as well as the prediction of "missing" links.

We will represent a network as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{Y})$ where $\mathcal{V} = \{v_1, \ldots, v_N\}$ is the set of vertices and $\mathcal{Y}$ is the set of observed links and non-links. Let $\boldsymbol{Y} \in \{0, 1, ?\}^{N \times N}$ denote a link (adjacency) matrix where the element $y_{ij} = 1$ if there is a link between vertex $v_i$ and $v_j$, $y_{ij} = 0$ if there is not a link, and $y_{ij} =?$ if the existence of a link is unobserved. Furthermore, let $\mathcal{Y}_1$, $\mathcal{Y}_0$, and $\mathcal{Y}_?$ denote the set of links, non-links, and unobserved links in the graph respectively.

Over the years, a multitude of methods for identifying latent structure in graphs have been proposed, most of which are based on grouping the vertices for the identification of homogeneous regions. Traditionally, this has been based on various community detection approaches where a community is defined as a densely connected subset of vertices that is sparsely linked to the remaining network [7, 8]. These structures have for instance been identified by splitting the graph using spectral approaches, analyzing flows, and through the analysis of the Hamiltonian. Modularity optimization [7] is a special case that measures the deviation of the fraction of links within communities from the expected fraction of such links based on their degree distribution [7, 8]. A drawback, however, for these types of analyses is that they are based on heuristics and do not correspond to an underlying generative process.

### 1.1. Probabilistic generative models:

Recently, generative models for complex networks have been proposed where links are drawn according to conditionally independent Bernoulli densities, such that the probability of observing a link $y_{ij}$ is given by $\pi_{ij}$,

$$p(\boldsymbol{Y}|\boldsymbol{\Pi}) = \prod_{(i,j)\in\mathcal{Y}} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}}. \tag{1.1}$$

In the classical Erdős-Rényi random graph model, each link is included independently with equal probability $\pi_{ij} = \pi_0$; however, more expressive models are needed in order to model complex latent structure of graphs. In the following, we focus on two related methods: latent class and latent feature models.

#### *1.1.1. Latent class models:*

In latent class models, such as the stochastic block model [9], also denoted the relational model (RM), each vertex $v_i$ belongs to a class $c_i$, and the probability, $\pi_{ij}$, of a link between $v_i$ and $v_j$ is determined by the class assignments $c_i$ and $c_j$ as $\pi_{ij} = \rho_{c_i c_j}$. Here, $\rho_{k\ell} \in [0, 1]$ denotes the probability of generating a link between a vertex in class $k$ and a vertex in class $\ell$. Inference in latent class models involves determining the class assignments as well as the class link probabilities. Based on this, communities can be found as (groups of) classes with high internal and low external link probability.

In the model proposed by [10] (HW) the class link probability, $\rho_{k\ell}$, is specified by a within-class probability $\rho_c$ and a between-class probability $\rho_0$. Another intuitive representation, which we refer to as DB, is to have a shared between-class probability but allow for individual within-class probabilities,

$$\boldsymbol{\rho}^{\text{HW}} = \begin{bmatrix} \rho_c & \rho_0 & \cdots & \rho_0 \\ \rho_0 & \rho_c & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_0 \\ \rho_0 & \cdots & \rho_0 & \rho_c \end{bmatrix}, \ \boldsymbol{\rho}^{\text{DB}} = \begin{bmatrix} \rho_1 & \rho_0 & \cdots & \rho_0 \\ \rho_0 & \rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_0 \\ \rho_0 & \cdots & \rho_0 & \rho_K \end{bmatrix}. \tag{1.2}$$

Both of these representations are consistent with the notion of communities with high internal and low external link density, and restricting the number of interaction parameters can facilitate model interpretation compared to the general RM.

Based on the Dirichlet process, [5, 6] propose a non-parametric generalization of the stochastic block model with a potentially infinite number of classes denoted the infinite relational model (IRM) and infinite hidden relational model respectively. The latter generalizing the stochastic block model to simultaneously model potential vertex attributes. Inference in IRM jointly determines the number of latent classes as well as class assignments and class link

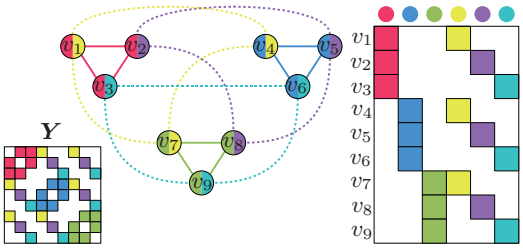


**Fig. 1**. Left: Example of a simple graph where each of the vertices have multiple memberships indicated by colors. Right: The corresponding assignment matrix.

probabilities. This approach readily generalizes to the HW and DB parameterizations of $\boldsymbol{\rho}$.

*1.1.2. Latent feature models:*

In latent feature models, the assumption that each vertex belongs to a single class is relaxed. Instead it is assumed that each vertex $v_i$ has an associated feature $\boldsymbol{z}_i$, and that probabilities of links are determined based on interactions between features. This generalizes the latent class models, which are the special case where the features are binary vectors with exactly one non-zero element.

Many latent feature models support the notion of discrete classes, but allow for mixed or multiple memberships (see Figure 1 for an illustration of a network with multiple class memberships). In the mixed membership stochastic block model (MMSB) [2] the vertices are allowed to have fractional class memberships. In binary matrix factorization [11] multiple memberships are explicitly modeled such that each vertex can be assigned to multiple clusters by an infinite latent feature model based on the Indian buffet process (IBP) [12]. [13] study this approach, for the specific case of a Bernoulli likelihood, Eq. (1.1), and extend the method to include additional side information as covariates in modeling the link probabilities. In their model, the probability of a link $\pi_{ij}$ is specified by $\pi_{ij} = f_\sigma\left(\sum_{k\ell} z_{ik} z_{j\ell} w_{k\ell} + s_{ij}\right)$, where $f_\sigma(\cdot)$ is a function with range $[0, 1]$ such as a sigmoid, and $w_{k\ell}$ are weights that affects the probability of generating a link between vertices in cluster $k$ and $\ell$. The term $s_{ij}$ accounts for bias as well as additional side-information. For example, if covariates $\boldsymbol{\phi}_i$ are available for each vertex $v_i$, [13] suggest including the term $s_{ij} = \beta d(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) + \boldsymbol{\beta}_i^\top \boldsymbol{\phi}_i + \boldsymbol{\beta}_j^\top \boldsymbol{\phi}_j$, where $\beta$, $\boldsymbol{\beta}_i$, and $\boldsymbol{\beta}_j$ are regression parameters, and $d(\cdot, \cdot)$ is some possibly nonlinear function.

In general, the computational cost of the single membership clustering methods mentioned above scales linearly in the number of links in the graph. Unfortunately, existing multiple membership models [2, 11, 13] scale quadratically in the number of vertices, because they require explicit computations for all links and non-links. This renders existing multiple membership modeling approaches infeasible for large networks. Furthermore, determining the multiple membership assignments is a combinatorial challenge as the number of potential states grow as $2^{KN}$ rather than $K^N$ in single membership models. In particular, standard Gibbs sampling approaches tend to get stuck in local suboptimal configurations where single assignment changes are not adequate for the identification of probable alternative configurations [11]. Consequently, there is both a need for computationally efficient models that scale linearly in the number of links as well as reliable inference schemes for modeling multiple memberships.

In this paper, we propose a new non-parametric Bayesian latent feature graph model, denoted the infinite multiple relational model (IMRM), that addresses the challenges mentioned above. Specifically, the contributions in this paper are the following: i) We propose the IMRM in which inference scales linearly in the number of links. ii) We propose a non-conjugate split-merge sampling procedure for parameter inference. iii) We demonstrate how the single membership IRM model implicitly accounts for multiple memberships. iv) We compare existing non-parametric single membership models with our proposed multiple membership counterparts in learning latent structure of a variety of benchmark "real" size networks and demonstrate that explicitly modeling multiple-membership results in more compact representations of latent structure.

## 2. INFINITE MULTIPLE-MEMBERSHIP RELATIONAL MODEL

Given a graph, assume that each vertex $v_i$ has an associated $K$-dimensional binary latent feature vector, $\boldsymbol{z}_i$, with $K_i = |\boldsymbol{z}_i|_1$ assignments. Consider vertex $v_i$ and $v_j$: For all $K_i K_j$ combinations of classes there is an associated probability, $\rho_{k\ell}$, of generating a link. We assume that each of these combinations of classes act independently to generate a link between $v_i$ and $v_j$, such that the total probability, $\pi_{ij}$, of generating a link between $v_i$ and $v_j$ is given by

$$\pi_{ij} = 1 - (1 - \sigma_{ij}) \prod_{k\ell} (1 - \rho_{k\ell})^{z_{ik} z_{j\ell}}, \tag{2.1}$$

where $\sigma_{ij}$ is an optional term that can be used to account for noise or to include further side-information as discussed previously. Under this model, the features act as independent causes of links, and thus if a vertex gets an additional feature it will result in an increased probability of linking to other vertices. In contrast to the model proposed by [13], where negative weights leads to features that inhibit links, our model is more restricted. Although this might result in less power to explain data, we expect that it will be easier to interpret the features in our model because links are directly generated by individual features and not through complex interactions between features. This is analogous to non-negative matrix factorization that is known to form parts-based representation because it does not allow component cancellations [14]. If the latent features $\boldsymbol{z}_i$ have only a single active element and $\sigma_{ij} = 0$, Eq. (2.1) reduces to $\pi_{ij} = \rho_{c_i c_j}$, i.e., the proposed model directly generalizes the IRM model; hence, we denote our model the infinite multiple-membership relational model (IMRM).

The link probability model in Eq. (2.1) has a very attractive computational property. In many real data sets, the number of non-links far exceeds the number of links present in the network. To analyze large scale networks where this holds it is a great advantage to devise algorithms that scale computationally only with the number of links present. As we show in the following, our model has that property. Assuming $\sigma_{ij} = 0$ for simplicity of presentation, we may write Eq. (2.1) more compactly as $\pi_{ij} = 1 - \mathrm{e}^{\boldsymbol{z}_i^\top \boldsymbol{P} \boldsymbol{z}_j}$,

where the elements of the matrix $\boldsymbol{P}$ are $p_{k\ell} = \log(1 - \rho_{kl})$. Inserting this in Eq. (1.1) we have

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{P}) &= \prod_{(i,j)\in\mathcal{Y}} \left(1 - \mathrm{e}^{\boldsymbol{z}_i^\top \boldsymbol{P} \boldsymbol{z}_j}\right)^{y_{ij}} \left(\mathrm{e}^{\boldsymbol{z}_i^\top \boldsymbol{P} \boldsymbol{z}_j}\right)^{1-y_{ij}} \\ &= \Big[\prod_{(i,j)\in\mathcal{Y}_1} (1 - \mathrm{e}^{\boldsymbol{z}_i^\top \boldsymbol{P} \boldsymbol{z}_j})\Big] \exp\Big[\sum_{(i,j)\in\mathcal{Y}_0} \boldsymbol{z}_i^\top \boldsymbol{P} \boldsymbol{z}_j\Big]. \end{aligned} \tag{2.2}$$

The exponent of the second term, which entails a sum over the possibly large set of non-links in the network, can be efficiently computed as

$$\sum_{k\ell} p_{k\ell}\Big(\sum_{i=1}^{N} z_{ik}\sum_{j=1}^{N} z_{j\ell} - \sum_{(i,j)\in\mathcal{Y}_1\cup\mathcal{Y}_?} z_{ik}z_{j\ell}\Big), \tag{2.3}$$

requiring only summation over links and "missing" links. Assuming that the graph is not dominated by "missing" links, the computation of Eq. (2.2) scales linearly in the number of graph links, $|\mathcal{Y}_1|$. We presently consider latent binary features $\boldsymbol{z}_i$, but we note that the model scales linearly for any parameterizations of the latent feature vector $\boldsymbol{z}_i$, as long as $\pi_{ij} = 1 - \mathrm{e}^{\boldsymbol{z}_i^\top \boldsymbol{P}\boldsymbol{z}_j} \in [0;1]$ which holds in general if $\boldsymbol{z}_i$ is non-negative.

As in existing multiple membership models [11, 13] we will assume an unbounded number of latent features. We learn the effective number of features through the Indian buffet process (IBP) representation [12], which defines a distribution over unbounded binary matrices,

$$\boldsymbol{Z} \sim \mathrm{IBP}(\alpha) \propto \frac{\alpha^K}{\prod_{\boldsymbol{h}\in[0,1]^N} K_{\boldsymbol{h}}!}\prod_{k=1}^{K}\frac{(N-m_k)!(m_k-1)!}{N!} \tag{2.4}$$

where $m_k$ is the number of vertices belonging to class $k$ and $K_{\boldsymbol{h}}$ is the number of columns of $\boldsymbol{Z}$ equal to $\boldsymbol{h}$.

As a prior over the class link probabilities we choose independent Beta distributions,

$$\rho_{k\ell}|a_{k\ell}, b_{k\ell} \sim \mathrm{Beta}(a_{k\ell}, b_{k\ell}) \propto \rho_{k\ell}^{a_{k\ell}-1}(1-\rho_{k\ell})^{b_{k\ell}-1}. \tag{2.5}$$

This is a conjugate prior for the single membership models where the parameters $a_{k\ell}$ and $b_{k\ell}$ correspond to pseudo counts of links and non-links respectively between classes $k$ and $\ell$.

## 2.1. Inference

In the following we present a method for inferring the parameters of the model: the infinite binary feature matrix $\boldsymbol{Z}$ and the link probabilities $\rho_{k\ell}$. In the latent class model when only a single feature is active for each vertex, the likelihood in Eq. (2.2) is conjugate to the Beta prior for $\rho_{k\ell}$. In that case, $\boldsymbol{P}$ can be integrated away and a collapsed Gibbs sampling procedure for $\boldsymbol{Z}$ can be used [5]. This is not possible in the IMRM; instead, we propose to sample $\boldsymbol{P} \sim p(\boldsymbol{P}|\boldsymbol{Z},\boldsymbol{Y})$ using Hamiltonian Markov chain Monte Carlo (HMC), and $\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{P},\boldsymbol{Y})$ using Gibbs sampling combined with split-merge moves.

### *2.1.1. HMC for class link probabilities:*

Hamiltonian Markov chain Monte Carlo (HMC) [15] is an auxiliary variable sampling procedure that utilizes the gradient of the log posterior to avoid the random walk behavior of other sampling methods such as Metropolis-Hastings. In the following we do not describe the details of the HMC algorithm, but only derive the required expressions for the gradient. To utilize HMC, the sampled variables must be unconstrained, but since $\rho_{k\ell}$ is a probability we make the following change of variable from $\rho_{k\ell} \in [0,1]$ to $r_{k\ell} \in (-\infty,\infty)$, $\rho_{k\ell} = \frac{1}{1+\exp(-r_{k\ell})}$, $r_{k\ell} = -\log\left(\rho_{k\ell}^{-1} - 1\right)$. Using the change of variables theorem, the prior for the class link probabilities expressed in terms of $r_{k\ell}$ is given by $p(r_{k\ell}|a_{k\ell}, b_{k\ell}) \propto \mathrm{e}^{a_{k\ell}r_{k\ell}}(\mathrm{e}^{r_{k\ell}}+1)^{-(a_{k\ell}+b_{k\ell})}$. With this, the relevant terms of the negative log posterior is given by

$$-\mathcal{L}_{\boldsymbol{P}} = \log p(\boldsymbol{P}|\boldsymbol{Z},\boldsymbol{Y}) = c + \sum_{(i,j)\in\mathcal{Y}_1}\log\left(1-\mathrm{e}^{\boldsymbol{z}_i^\top\boldsymbol{P}\boldsymbol{z}_j}\right)$$
$$+\sum_{(i,j)\in\mathcal{Y}_0}\boldsymbol{z}_i^\top\boldsymbol{P}\boldsymbol{z}_j + \sum_{k\ell} a_{k\ell}r_{k\ell} + (a_{k\ell}+b_{k\ell})\log(\mathrm{e}^{r_{k\ell}}+1), \tag{2.6}$$

where $c$ does not depend on $\boldsymbol{P}$. From this, the required gradient can be computed,

$$\frac{\partial\mathcal{L}_{\boldsymbol{P}}}{\partial r_{k\ell}} = -\sum_{(i,j)\in\mathcal{Y}_1}\frac{\mathrm{e}^{\boldsymbol{z}_i^\top\boldsymbol{P}\boldsymbol{z}_j}}{1-\mathrm{e}^{\boldsymbol{z}_i^\top\boldsymbol{P}\boldsymbol{z}_j}}z_{ik}z_{j\ell}\rho_{k\ell}$$
$$+\sum_{(i,j)\in\mathcal{Y}_0} z_{ik}z_{j\ell}\rho_{k\ell} + (a_{k\ell}+b_{k\ell})\rho_{k\ell} - a_{k\ell}.$$

Again, the possibly large sum over non-links in the second term can be computed efficiently as in Eq. (2.3).

### *2.1.2. Gibbs sampler for binary features:*

Following [12], a Gibbs sampler for the latent binary features $\boldsymbol{Z}$ can be derived. Consider sampling the $k$th feature of vertex $v_i$: If one or more other vertices also possess the feature, i.e., $m_{-ik} = \sum_{j\neq i} z_{jk} > 0$, the posterior marginal is given by

$$p(z_{ik}=1|\boldsymbol{Z}_{-(ik)},\boldsymbol{P},\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{P})\frac{m_{-ik}}{N}. \tag{2.7}$$

When evaluating the likelihood term, only the terms that depend on $z_{ik}$ need be computed and the Gibbs sampler can be implemented efficiently by reusing computation and by up and down dating variables.

In addition to sampling existing features, $K_1^{(i)} = \mathrm{Poisson}\left(\frac{\alpha}{N}\right)$ new features should also be associated with $v_i$. [12] suggest *"… computing probabilities for a range of values of $K_1^{(i)}$ up to some reasonable upper bound…"*; however, following [11] we take another approach and sample the new features by Metropolis-Hastings using the prior as proposal density. The values of $\rho_{kl}$ corresponding to the new features are proposed from the prior in Eq. (2.5).

### *2.1.3. Split-merge move for binary features:*

A drawback of Gibbs sampling procedures is that only a single variable is updated at a time, which makes the sampler prone to get stuck in suboptimal configurations. As a remedy, bolder Metropolis Hasting moves can be considered in which multiple changes of assignments help exploring alternative high probability configurations. How these alternative configurations are proposed is crucial in order to attain reasonable acceptance rates. A popular approach is to split or merge existing classes as proposed in [16] for the Dirichlet process mixture model (DPMM). Split-merge sampling in the IBP has previously been discussed briefly by [11] and [13].

Inspired by the non-conjugate sequential allocation split-merge sampler for the DPMM [17], we propose the following procedure: Draw two non-zero elements of $\boldsymbol{Z}$, $(k_1, i_1)$ and $(k_2, i_2)$. If $k_1 = k_2$ propose a split — otherwise propose to merge classes $k_1$ and $k_2$ into a joint cluster $k_1$. Accept the proposal with the Metropolis-Hastings accept rate, $a^* = \min\left(1, \frac{p(\boldsymbol{Z}^*,\boldsymbol{P}^*|\boldsymbol{Y})q(\boldsymbol{Z}|\boldsymbol{Z}^*)q(\boldsymbol{P}|\boldsymbol{P}^*)}{p(\boldsymbol{Z},\boldsymbol{P}|\boldsymbol{Y})q(\boldsymbol{Z}^*|\boldsymbol{Z})q(\boldsymbol{P}^*|\boldsymbol{P})}\right)$. In case of a merge, we remove $k_2$ and assign all its vertices to $k_1$, and we remove the corresponding row and column of $\boldsymbol{P}$ (this proposal is deterministic and has probability one). For a split, we

remove all vertices except $i_1$ from cluster $k_1 = k_2 = k$ and create a new cluster $k^*$ and assign $i_2$ to it. We then sample a new row and column $\rho^*_{k'\ell'}$ for the new cluster as described below. Next we sequentially allocate [17] the remaining original members of $k$ to either $k$ or $k^*$ or both in a restricted Gibbs sampling sweep, and refine the allocation through $t$ additional restricted Gibbs scans [16].

The proposal density for $\rho^*_{k'\ell'}$ is based on a random walk, $\rho^*_{k'\ell'} \sim \mathrm{Beta}(\bar{a}_{k'\ell'}, \bar{b}_{k'\ell'})$, where

$$\bar{b}_{k'\ell'} = \max\left(1, (1-\bar{\rho}_{k'\ell'})m_k^2 - 1 + \bar{\rho}_{k'\ell'}\right), \quad (2.8)$$

$$\bar{a}_{k'\ell'} = \max\left(1, \frac{\bar{\rho}_{k'\ell'}}{1-\bar{\rho}_{k'\ell'}}\bar{b}_{k'\ell'}\right), \quad (2.9)$$

such that $\rho^*_{k'\ell'}$ has mean $\bar{\rho}_{k'\ell'}$ and variance equal to the empirical variance, $\bar{\rho}_{k'\ell'}(1-\bar{\rho}_{k'\ell'})/m_k^2$. We choose the mean of the random walk as

$$\bar{\rho}_{k'\ell'} = \begin{cases} \rho_{kk} & \ell' = k' = k^* \\ \frac{1}{K-1}\sum_{\ell\neq k}\rho_{k\ell} & k' = k^*, \ell' = k \\ \frac{1}{K-1}\sum_{\ell\neq k}\rho_{\ell k} & k' = k, \ell' = k^* \\ \rho_{k\ell} & \text{otherwise,} \end{cases}$$

such that the new class has a similar within and between class link probabilities as the original class, but such that the class link probability between the original and new cluster is similar to the remaining between class link probabilities. This choice is crucial, since it favors splitting classes into two classes that are no more related than the relation to the remaining classes.

## 3. RESULTS

Based on the HW, DB, and RM parametrization of $\boldsymbol{\rho}$, we compared our proposed IMRM to the corresponding single-membership IRM [5]. We evaluated the models on a range of synthetically generated as well as real world networks. We assessed model performance in terms of ability to predict held-out links and non-links. As performance measure we used the area under curve (AUC) of the receiver operating characteristic (ROC). We also computed the predictive log likelihood (not shown here) which gave similar results. For comparison, we included the performance of several standard non-parametric link prediction approaches based on the following scores, $\gamma^{\mathrm{ComN}}_{i,j} = \boldsymbol{y}_i^\top \boldsymbol{y}_j$, $\gamma^{\mathrm{DegPr}}_{i,j} = k_i k_j$, $\gamma^{\mathrm{Jacc}}_{i,j} = \frac{\boldsymbol{y}_i^\top \boldsymbol{y}_j}{k_i + k_j - \boldsymbol{y}_i^\top \boldsymbol{y}_j}$ and $\gamma^{\mathrm{ShP}}_{i,j} = \frac{1}{\min_p\{(\boldsymbol{Y}^p)_{i,j}>0\}}$, where $k_i = \sum_j y_{ij}$ is the degree of vertex $v_i$.

In all the analyses we removed 2.5% of the links and an equivalent number of non-links for cross-validation. We analyzed a total of five random data splits and all of the analyses were based on 2500 sampling iterations initialized randomly with $K = 50$ clusters. Each iteration was based on split-merge sampling using sequential allocation with $t = 2$ restricted Gibbs scans followed by standard Gibbs sampling. Our implementation of the IRM was based on collapsed Gibbs sampling (i.e. integrating out $\boldsymbol{\rho}$) as proposed in [5] but we also included a conjugate single-membership split-merge step corresponding to the proposed non-conjugate split-merge sampler. The priors were chosen as $\alpha = \log(N)$, $a_{kk} = 5$, $a_{k\ell} = 1 \forall k \neq \ell$ and $b_{kk} = 1$, $b_{k\ell} = 5 \forall k \neq \ell$ which renders the priors practically non-informative.

### 3.1. Synthetic networks

We analyzed a total of six synthetic networks generated according to the HW, DB and RM models based on the vertices having either one or two memberships to the underlying classes. For the single membership models we generated a total of $K = 5$ groups each containing 100 vertices. For the HW generated network we set $\rho_c = 1$ and $\rho_0 = 0$ while for the DB generated network we used a within community densities $\rho_k$ ranging from 0.2 to 1 while $\rho_0 = 0$. The RM generated network had same within community densities as the DB network but included varying degrees of overlap between the communities. The multiple membership models denoted MHW, MDB and MRM were generated from the corresponding single membership models as $\boldsymbol{Y} \vee \boldsymbol{R}\boldsymbol{Y}\boldsymbol{R}^\top$ (where $\vee$ denotes element-wise or and $\boldsymbol{R}$ is a random permutation matrix with diagonal zero), such that each vertex belongs to two classes.

**Table 1**. Summary of the analyzed real networks: $r$ denotes the networks assortativity, $c$ the clustering coefficient [3], $L$ the average shortest path.

| NETWORK | $N$ | $\lvert\mathcal{Y}_1\rvert$ | $r$ | $c$ | $L$ |
|---|---|---|---|---|---|
| Yeast | 2,284 | 6,646 | -0.10 | 0.13 | 4.4 |
| Protein-protein interaction network [1] | | | | | |
| USPower | 4,941 | 6,594 | 0.00 | 0.08 | 19.9 |
| Topology of power grid [3] | | | | | |
| Erdos | 5,534 | 8,472 | -0.04 | 0.08 | 3.9 |
| Erdös 02 collaboration network [18] | | | | | |
| FreeAssoc | 10,299 | 61,677 | -0.07 | 0.12 | 3.9 |
| Word relations in free association [19] | | | | | |
| Reuters911 | 13,314 | 148,038 | -0.11 | 0.37 | 3.1 |
| Word co-occurrence [20] | | | | | |

While the IMRM model explicitly accounts for multiple memberships, the IRM model can also implicitly account for multiple memberships through the between class interactions. To illustrate this, we analyzed the generated HW and MHW data by the IRM model as well as the proposed IMRM model (see Figure 2). When there are only single memberships, the IMRM reduces to the IRM model; however, when the network is generated such that the vertices have multiple memberships the IMRM model correctly identifies the ($2{\cdot}5 = 10$) underlying classes. The IRM model on the other hand extracts a larger number of classes corresponding to all possible ($5^2 = 25$) combinations of classes present in the data. The estimated $\boldsymbol{\rho}$ indicates how these 25 classes combine to form the 10 underlying multiple membership groups in the network. As such, the IRM model has the same expressive power as the proposed multiple membership models but interpreting the results can be difficult when multiple membership community structure is split into several classes with complex patterns of interaction.

Figure 3 shows the link-prediction AUC scores from the analysis of the six generated networks. Results show that all models work well on data generated according to their own model or models which they generalize. We also notice, that the IRM model accounts well for multiple membership structure as discussed and illustrated in Figure 2. The HW and DB models on the other hand fail in modeling networks with multiple memberships.

### 3.2. Real Networks

We finally analyzed five benchmark complex networks summarized in Table 1. The sizes of most of the networks makes it computationally infeasible for us to analyze them using the existing multiple-membership approaches proposed in [2, 11, 13]. For all the networks, multiple memberships are conceivable: In protein

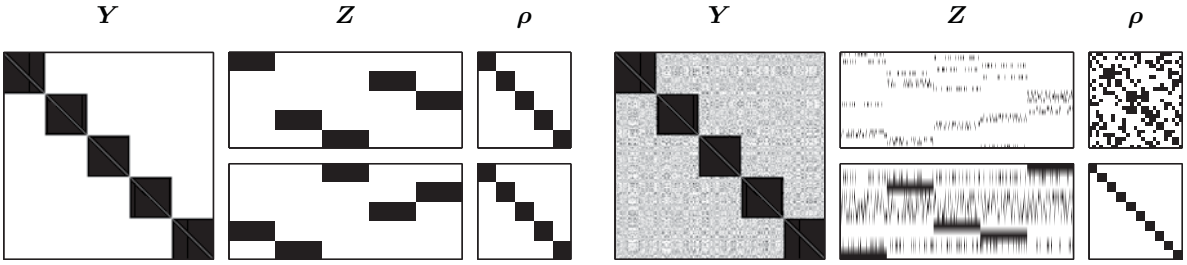


**Fig. 2**. IRM (left) and IMRM (rigth) analysis of single (left) and multiple membership HW network (right). On the single membership data, both models find the correct class assignments. On the multiple membership data, the IMRM finds the correct 10 classes, while IRM extracts 25 classes, which through $\rho$ accounts for all combinations of classes present in the data.

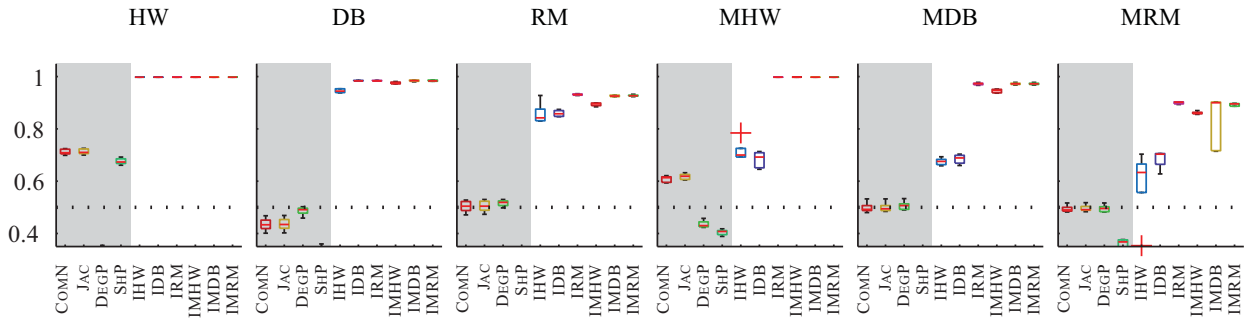


**Fig. 3**. AUC scores for the analysis of the six synthetically generated data sets.

interaction networks such as the Yeast network proteins can be part of multiple functional groups, in social networks such as Erdos scientist collaborate with different groups of people depending on the research topic, and in word relation networks such as FreeAssoc and Reuters911 words can have multiple meanings/contexts. For all these networks explicitly modeling these multiple contexts can potentially improve on the structure identification over the equivalent single membership models.

In Figure 4 the AUC link prediction score is given for the five networks analyzed. As can be seen from the results, modeling multiple memberships significantly improves on predicting links in the network. In particular when considering the IHW and IDB models and the corresponding proposed multiple membership models, the learning of structure is improved substantially for all networks except USPower. Furthermore, it can be seen that the IRM model that can also implicitly account for multiple memberships in general has a similar performance to the multiple membership models. The poor identification of structure in the USPower network might be due to the fact that the average path between vertices are very high rendering it difficult to detect the underlying structure for any but the most simple IHW model. While the IRM and IMRM perform equally well in terms of link prediction it can be seen at the top of table 2 that the average number of extracted components for the IMRM model is significantly smaller than the number of components extracted by the IRM model for all networks except the Reuters911 network where no significant difference is found. As a result, the IMRM model is in general able to extract a more compact representation of the latent structure of networks. At the bottom of table 2 is given the total cpu-time for estimating the 2500 samples for each of the network using the IRM and IMRM showing that the order of magnitude for the computational cost of the two models are the same.

## 4. DISCUSSION

While single membership models based on the IRM indirectly can account for multiple memberships as we have shown, the benefit of the proposed framework is that it allows for these multiple memberships to be modeled explicitly rather than through complex between-group interactions based on a multitude of single membership components. On synthetic and real data we demonstrated that explicitly modeling multiple-membership resulted in a more compact representation of the inherent structure in networks. We further demonstrated that models that can capture multiple memberships (which includes the IRM model) significantly improve on the link prediction relative to models that can only account for single membership structure, i.e., the IHW and IDB models. We presently considered undirected networks but we note that the proposed approach readily generalizes to directed and bipartite graphs. Furthermore, the approach also extends to include side information as proposed in [13] as well as simultaneous modeling of vertex attributes [6]. We note however, that the inclusion of side information requires a linearly scalable parameterization in order for the overall model to remain computationally efficient. An attractive property of the IRM model over the IMRM model is that the IRM model admits the use of collapsed Gibbs sampling which we have found to be more efficient relative to sampling the non-conjugate multiple membership models where additional sampling of the $\rho$ parameter is required. In future research, we envision combining the IRM and IMRM model, using the IRM as initialization for the IMRM or by forming hybrid models.

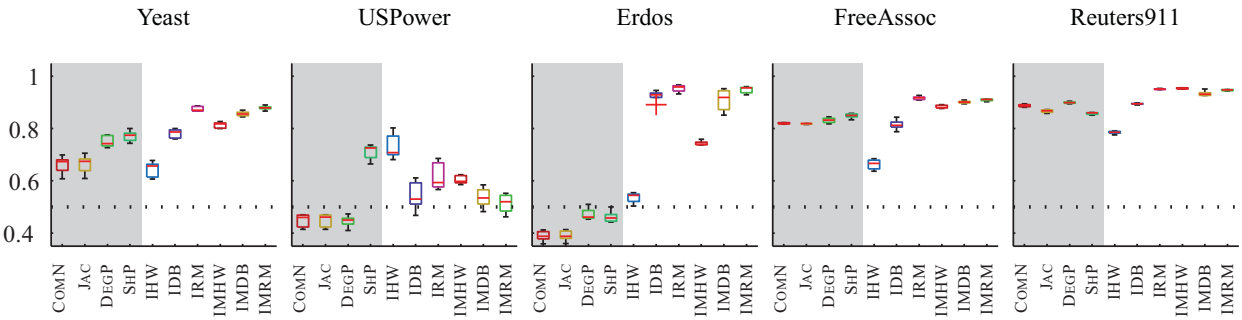


**Fig. 4**. AUC scores for the analysis of the five real networks.

**Table 2**. The number of extracted components and cpu-time for 2500 iterations for the IRM and IMRM models. Bold denotes that the number of components are significantly different between the two models (difference in mean at least two standard deviations apart.)

| | | Yeast | USPower | Erdos | FreeAssoc | Reuters911 |
|---|---|---|---|---|---|---|
| Number of components | IRM | $24.0 \pm 0.8$ | $8.6 \pm 0.4$ | $10.4 \pm 0.3$ | $58.6 \pm 0.7$ | $39.8 \pm 2.1$ |
| | IMRM | $\mathbf{15.4 \pm 0.9}$ | $\mathbf{6.8 \pm 0.5}$ | $\mathbf{6.8 \pm 0.6}$ | $\mathbf{15.6 \pm 0.9}$ | $44.8 \pm 1.0$ |
| CPU-time (hours) | IRM | $2.3 \pm 0.1$ | $4.0 \pm 0.2$ | $14.6 \pm 5.9$ | $30.1 \pm 0.6$ | $32.5 \pm 5.4$ |
| | IMRM | $1.7 \pm 0.1$ | $8.9 \pm 0.8$ | $7.1 \pm 0.5$ | $28.1 \pm 1.9$ | $71.5 \pm 3.2$ |

## 5. REFERENCES


[1] Shiwei Sun, Lunjiang Ling, Nan Zhang, Guojie Li, and Runsheng Chen, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[2] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing, "Mixed membership stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, 2008.

[3] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks.," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.

[4] M. E. J. Newman, "The structure of scientific collaboration networks," *PNAS*, vol. 98, no. 2, pp. 404–409, 2001.

[5] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda, "Learning systems of concepts with an infinite relational model," in *Artificial Intelligence, Proceedings of the National AAAI Conference on*, 2006.

[6] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel, "Learning infinite hidden relational models," *Uncertainity in Artificial Intelligence (UAI2006)*, 2006.

[7] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, pp. 026113–1–15., 2004.

[8] Jorg Reichardt and Stefan Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, 2006.

[9] Krzysztof Nowicki and Tom A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1077–1087, 2001.

[10] Jake M. Hofman and Chris H. Wiggins, "Bayesian approach to network modularity," *Physical Review Letters*, vol. 100, no. 25, Jun 2008.

[11] Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis, "Modeling dyadic data with binary latent factors," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, vol. 19, pp. 977–984.

[12] Thomas L. Griffiths and Zoubin Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Neural Information Processing Systems, Advances in (NIPS)*, 2006, pp. 475–482.

[13] Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan, "Nonparametric latent feature models for link prediction," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1276–1284.

[14] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[15] Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth, "Hybrid monte carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216–222, Sep 1987.

[16] Sonia Jain and Radford M. Neal, "A split-merge markov chain monte carlo procedure for the dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, vol. 13, no. 1, pp. 158–182, 2004.

[17] David B. Dahl, "Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models," Tech. Rep., Texas A&M University, 2005.

[18] Vladimir Batagelj and Andrej Mrvar, "Pajek data sets," 2006.

[19] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, "The university of south florida word association, rhyme, and word fragment norms," 1998.

[20] S. R. Corman, T. Kuhn, R. D. McPhee, and Dooley K. J., "Studying complex discursive systems: Centering resonance analysis of communication," *Human communication research*, vol. 28, no. 2, pp. 157–206, 2002.