UCRL-TR-230743

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Unsupervised Group Discovery and LInk Prediction in Relational Datasets: a nonparametric Bayesian approach

P.S. Koutsourelakis

May 7, 2007

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Unsupervised Group Discovery and Link Prediction in Relational Datasets: a nonparametric Bayesian approach

Predictive Knowledge Systems Initiative

P.S. Koutsourelakis

May 9, 2007

Contents

1 Introduction

Clustering represents one of the most common statistical procedures and a standard tool for pattern discovery and dimension reduction. Most often the objects to be clustered are described by a set of measurements or observables e.g. the coordinates of the vectors, the attributes of people. In a lot of cases however the available observations appear in the form of links or connections (e.g. communication or transaction networks). This data contains valuable information that can in general be exploited in order to discover groups and better understand the structure of the dataset. Since in most real-world datasets, several of these links are missing, it is also useful to develop procedures that can predict those unobserved connections.

In this report we address the problem of unsupervised group discovery in relational datasets. A fundamental issue in all clustering problems is that the actual number of clusters is unknown a priori. In most cases this is addressed by running the model several times assuming a different number of clusters each time and selecting the value that provides the best fit based on some criterion (i.e. Bayes factor in the case of Bayesian techniques). It is easily understood that it would be preferable to develop techniques that are able to number of clusters is essentially learned from that data along with the rest of model parameters. For that purpose, we adopt a nonparametric Bayesian framework which provides a very flexible modeling environment in which the size of the model i.e. the number of clusters, can adapt to the available data and readily accommodate outliers. The latter is particularly important since several groups of interest might consist of a small number of members and would most likely be smeared out by traditional modeling techniques. Finally, the proposed framework combines all the advantages of standard Bayesian techniques such as integration of prior knowledge in a principled manner, seamless accommodation of missing data, quantification of confidence in the output etc.

In the first section of this report, we review the Infinite Relational Model (IRM) which serves as the basis for further developments. The IRM assumes that each object belongs to a single group. In subsequent sections we discuss two mixed-membership models i.e. models which can account for the fact that an object can belong to several groups simultaneously in which case we are also interested in the degree of membership. For that purpose it is perhaps more natural to talk with respect to identities rather than groups. In particular we assume that each object has an unknown identity which can consist of one or more components. **The terms groups and identities would therefore be considered equivalent in subsequent sections.** A su-section is also devoted to variational techniques which have the potential of accelerating the inference process. Finally we discuss possible extensions to dynamic settings in which the available data includes timestamps and the goal is to find how group sizes and group memberships evolve in time. Even though the majority of the presentation is restricted to objects of a single type (domain) and pairwise, binary links of a single type, it is shown that the framework proposed can be extended to links of various types between several domains.

2 The Infinite Relational Model

Consider a dataset which contains observations about links/connections between objects of various types. These links can be of various types and take binary, categorical or real values. Furthermore they might relate two or more objects at the time. For illustration purposes and without loss of generality we will restrict the presentation to pairwise, binary links $R_{i,j}$ in a single domain (i.e. *person i to person j*) and follow the formalism introduced in ([15]) for the Infinite Relational Model (IRM). We present examples with two domains in ubsequent sections. The goal is to group the objects based on those observables. For that purpose we define a generative model which postulates that the likelihood of any link between a pair of objects i and j depends exclusively on the identities I_i and I_j. In that respect it is identical to the stochastic block-model [19] which is based however on a fixed, a priori determined number of clusters and shares a lot of common characteristics with other latent variable models ([11, 13, 12]). Formally this leads to the following decomposition of the likelihood:

$$p(\boldsymbol{R} \mid \textit{identities } \boldsymbol{I}) = \prod_{i,j} p(R_{i,j} \mid I_i, I_j) \tag{1}$$

The product above is over all pairs of nodes between which links (with value 0 or 1) have been observed (missing links are omitted). In a Bayesian setting the individual likelihoods can be modeled with a Bernoulli distribution with a hyper-parameter $\eta(I_i, I_j)$. Furthermore a beta distribution $Beta(\beta_1, \beta_2)$ can be used as a hyper-prior for each η ([15]). In fact the η's can be readily integrated out which leads to a simpler expression for the likelihood $p(\boldsymbol{R} \mid \textit{identities } \boldsymbol{I})$ that depends only on the counts $m_0(I, J)$, $m_1(I, J)$ of 0 and 1 links respectively between each pair of identities $I,\ J$:

$$p(\boldsymbol{R} \mid \textit{identities } \boldsymbol{I}) = \prod_{I,J} \frac{beta(m_0(I,J) + \beta_1, m_1(I,J) + \beta_2)}{beta(\beta_1, \beta_2)} \tag{2}$$

where $beta(\ ,\ )$ denotes the beta function.

Extensions to real-valued links can be readily obtained by using an appropriate prior for $p(R_{i,j} \mid I_i, I_j)$ (i.e exponential, gamma etc). Furthermore if a vector of attributes

$\boldsymbol{x^{(i)}}$ is also observed at each object i then the likelihood can be augmented as follows:

$$p(\boldsymbol{R}, \boldsymbol{x} \mid \textit{identities } \boldsymbol{I}) = \prod_{i,j} p(R_{i,j} \mid I_i, I_j) \prod_{i} p(\boldsymbol{x^{(i)}} \mid I_i) \tag{3}$$

and an appropriate prior can be defined for the individual likelihoods $p(\boldsymbol{x^{(i)}} \mid I_i)$.

Since the number of identities is unknown a priori (a problem discussed in the introduction) we adopt a nonparametric prior for I_i's. In particular we use a distribution over partitions induced by a Chinese Restaurant Process (CRP) ([3, 8, 20]). Of the several mathematical interpretations that have appeared perhaps the simplest is the one in which the CRP arises as the infinite limit of a Dirichlet distribution on the K-dimensional simplex as $K \to \infty$ ([17]). A fundamental characteristic is of the CRP is exchangeability which simply implies that the probability associated to a certain partition is independent of the order in which objects are assigned to groups. Under the CRP, customers (which in our case correspond to objects) enter a Chinese restaurant sequentially and are assigned to tables (which in our case correspond to groups) according to the following conditional:

$$p(I_N = t \mid \boldsymbol{I_{-N}} = t) = \begin{cases} \frac{n_t}{N-1+a} & \text{if } n_t > 0 \\ \frac{a}{N-1+a} & \text{if } n_t = 0 \end{cases} \tag{4}$$

where I_N and $\boldsymbol{I_{-N}}$ are the group indicator variables of object N and $1, 2, \ldots, N-1$ respectively and n_t is the number of objects already assigned to group t. Hence the N^{th} object can be assigned to an existing group or to a new group. The number of groups can therefore vary and the parameter a controls the propensity of the model to create new groups. Typically a gamma prior is adopted which leads to a simple expression for the conditional posterior that can then be used in Gibbs sampling ([26]). Posterior inference with respect to the latent variables I_i can also be performed using Gibbs sampling ([7, 6, 18, 27]). This simply makes use of the prior conditionals (Equation (4)) and the likelihood function (Equation (2)).

The IRM is a flexible and lightweight model for group discovery. An important disadvantage has to do with the computational effort involved particularly in datasets where a large number of objects is present in which case Gibbs sampling can become inefficient as it affects only a single latent variable at each iteration. Significant acceleration can be achieved by employing split-merge techniques recently developed for nonparametric models ([14]). With respect to its modeling capabilities, the most significant deficiency is that each object can adopt a single identity (i.e. belong to a single group) and all the links it participates in arise as a result of that identity. This assumption can be too restrictive as in general the identity of each object does not consist of a single component but rather of several components which co-exist at different proportions. For example if the links are phone-calls and the objects are people then a person might communicate with other people as a co-worker or as a friend etc. This is particularly noticeable if several link types are simultaneously considered such as phone-calls, emails and letters where depending on the type, each person participates with different identities. This issue is addressed in detail in section 3.

2.1 Posterior Inference with respect to missing links

As mentioned earlier a task that is of interest in real-world data is to predict unobserved links. If $\boldsymbol{R}$ represents the observed data and $R_{i,j}$ a missing link between objects i and j, we wish to calculate the (posterior) probability $p(R_{i,j} \mid \boldsymbol{R})$. In oder to facilitate the exposition and without loss of generality, we assume that the hyper-parameters β_1, β_2 and a are fixed and omit them for the expressions that follow. Hence:

$$p(R_{i,j} \mid \boldsymbol{R}) = \int_{\boldsymbol{I}} p(R_{i,j} \mid \boldsymbol{I}, \boldsymbol{R}) p(\boldsymbol{I} \mid \boldsymbol{R}) \, d\boldsymbol{I} \tag{5}$$

where the vector $\boldsymbol{I}$ represents the identities of all the objects and $p(\boldsymbol{I} \mid \boldsymbol{R})$ the posterior distribution. As the latter is not known explicitly, the integration above is carried out using Monte Carlo and the posterior samples drawn by Gibbs sampling as mentioned above. As for the first term in the integrand:

$$p(R_{i,j} \mid \boldsymbol{I}, \boldsymbol{R}) = \frac{p(R_{i,j} \text{ and } \boldsymbol{R} \mid \boldsymbol{I})}{p(\boldsymbol{R} \mid \boldsymbol{I})} \tag{6}$$

For $R_{i,j} = 1$, according to Equation (2) and the definition of the beta function this ratio becomes:

$$\frac{p(R_{i,j} \text{ and } \boldsymbol{R} \mid \boldsymbol{I})}{p(\boldsymbol{R} \mid \boldsymbol{I})}) = \frac{m_1(I_i, I_j) + \beta_2 + 1}{m_0(I_i, I_j) + \beta_1 + m_1(I_i, I_j) + \beta_2 + 1} \tag{7}$$

where $m_0(I_i, Ij_{)}$ and $m_1(I_i, I_j)$ are the counts of 0 and 1 links in the observable data $\boldsymbol{R}$ between objects assigned to identities I_i and I_j respectively.

3 Mixed Membership Model

Mixed-membership models have been introduced to account for the fact that objects can exhibit several distinct identities in their relational patterns ([2, 1]). Posed differently, an object can establish links as a member of multiple groups. This aspect is particularly important in real-world datasets where relational data can be used for detection of an anomalous behavior/identity. It is unlikely that the objects of interest will exhibit this identity in all their relations. It is of interest therefore to find all the different identities exhibited but also the degree to which these are present in each object's overall identity. These components can be shared among the objects in the same domain but the proportions can vary from one to another.

In order to capture that effect we alter the aforementioned model by introducing a latent variable for each object and for each observable link that this object participates in. Let $R_{i,j}^m$ be an observable link between objects i and j where m is an index over all available links. We introduce therefore the latent variables $I_{i,m}$ which denote the identity exhibited by object i in link m (The index m is redundant with respect to the definition of the link as the participating objects i and j suffice, but is used herein to facilitate the notation for the latent identity variables). Similarly to the IRM (Equation (1)) we assume that the likelihood can be decomposed as:

$$p(\boldsymbol{R} \mid \ \boldsymbol{I}) = \prod_m p(R_{i,j}^m \mid I_{i,m}, I_{j,m}) \tag{8}$$

Hence, (in general) there are several latent variables, say m_i, associated with each object i. Chinese restaurant process priors can be used for each object with a parameter a_i. Although this would produce groupings for each object, these groups will not be shared across objects and therefore would not be relevant with respect to group discovery in the whole domain. For that purpose we adopt a hierarchical prior, namely the Chinese Restaurant Franchise (CRF) which was first presented in [22]. Based on the restaurant analog customers enter several restaurants belonging to a franchise and share the same menu. Their group assignment is based on the dish they end up eating which is determined in a two-step process. Firstly, the customers in each restaurant are seated based on independent CRP's. Therefore the table assignment $t_{i,m}$ of customer m in restaurant i is defined by:

$$p(t_{i,m} = t \mid \boldsymbol{t_{i,-m}}) = \begin{cases} \frac{n_{i,t}}{m_i - 1 + a_i} & \text{if } n_{i,t} > 0 \\ \frac{a_i}{m_i - 1 + a_i} & \text{if } n_{i,t} = 0 \end{cases} \tag{9}$$

where $n_{i,t}$ is the number of customers seated at table t in restaurant i and a_i the parameter of the CRP pertinent to restaurant i. Once the seating has taken place, each table in each restaurant orders sequentially a dish (common for all the occupants of the table) from the common menu. The probabilities are again independent of the order in which this process takes place and are determined by a base CRP with parameter a_0 (denoted by CRP_0):

$$p(d_{i,t} = d \mid \boldsymbol{d_{-(i,t)}}) = \begin{cases} \frac{s_d}{M - 1 + a_0} & \text{if } s_d > 0 \\ \frac{a_0}{M - 1 + a_0} & \text{if } s_d = 0 \end{cases} \tag{10}$$

where $d_{i,t}$ is the dish served at table t of restaurant i, s_k is the number of tables (over all restaurants) that have ordered dish d and M is the total number of tables (over all restaurants). Based on the notation introduced the group assignment $I_{i,m}$ is equal to $d_{i,t_{i,m}}$ i.e. the dish served at table $t_{i,m}$ where the customer m of restaurant i was seated. It becomes apparent that the CRPs at the restaurant level express the mixed-membership effect while the base CRP accounts for the groups/identities associated with all the objects. The model is summarized below:

$$\begin{aligned} CRP_0 \mid a_0 &\sim CRP(a_0) \\ I_{i,m} \mid a_i &\sim CRP(a_i, CRP_0) \\ \eta(I_1, I_2) \mid \beta_1, \beta_2 &\sim Beta(\beta_1, \beta_2) \\ R_{i,m} \mid I_{i,m}, I_{j,m}, \boldsymbol{\eta} &\sim Bernoulli\left(\eta(I_{i,m}, I_{j,m})\right) \end{aligned} \tag{11}$$

Equations 9 and 10 readily imply how Gibbs sampling can be performed for posterior inference with respect to the latent variables $I_{i,m}$. The latter are not directly sampled but instead we first sample $t_{i,m}$ and then for $d_{i,t}$. Further details are contained in ([22]). It should finally be noted that the posterior is a distribution on partitions and therefore exchangeable. If for example we have three objects with a single latent variable associated with each one and two groups, then the group assignment (1, 2, 1) is equivalent (in the sense that the posterior likelihood is the same) to (2, 1, 2). This complicates matters in the sense that posterior inference across several samples cannot be performed with respect to specific groups (as their labels might change from sample to sample). We can however look at the maximum likelihood (or maximum posterior) configuration and calculate degrees of membership as described below.

3.1 Quantifying Degree of Membership

Consider a specific configuration drawn from the posterior in which all latent variables $I_{i,m}$ (the customers in our CRF analog) have been associated with tables and dishes (i.e. identities). We wish to calculate the degree of membership of each object to each of the identities, say K, that have been found. Posed differently, if a new customer $m_i + 1$ arrived at restaurant i what would the probability be that he ends up eating one of the K dishes?

If we consider a dish k then this probability can be decomposed into the sum two terms: a) probability that he eats k while seated to one of the existing tables, and b) probability that he eats k while being seated to a new table in restaurant i which was created to accommodate only him. If T_i is the number of existing tables at restaurant i then the first term p_a would be:

$$p_a = \sum_{t=1}^{T} p(t_{i,m_{i+1}} = t)\ p(d_{i,t} = k) \tag{12}$$

The second term in that sum would either be 0 or 1 since all the existing tables have already been assigned one of the K dishes. The first term depends on the $CRP(a_i)$ and can be calculated based on Equation (9).

Returning to the probability of the second component, p_b which corresponds to the event that the new customer is being seated at a new table $T_i + 1$ and is served dish k, then this can be expressed as:

$$p_b = p(t_{i,m_{i+1}} = T_i + 1)\ p(d_{i,T_i+1} = k) \tag{13}$$

The first term above is given by Equation (9) and the second from Equation (10) as it depends on the number of tables already assigned to dish k.

3.2 Link Prediction

Prediction of missing links is also a task of interest that can be readily performed based on what was already mentioned regarding the same task for IRM and the section above on quantifying the degrees of membership to the various groups. It should be noted that since it is a "relative quantity", i.e. it does not depend on the labels of the groups, this probability can be calculated across several samples of the posterior by essentially averaging over the samples drawn from MCMC as implied in Equation (5). For that purpose we calculate here only the first term in that integrand i.e. $p(R_{i,j} \mid \boldsymbol{R}, \boldsymbol{I})$ i.e. the probability of an unobserved link $R_{i,j}$ between objects i and j given the observed links $\boldsymbol{R}$ and the latent variables $\boldsymbol{I}$. It is also assumed without loss of generality and for notational economy that the hyper-parameters β_1, β_2, a_0, a_i $\forall i$ are constant.

In order to calculate the probability of the new link we have to introduce two new latent variables I_{i,m_i+1}, I_{j,m_j+1} for each of the participating objects i and j which are essentially the $m_i + 1$ and $m_j + 1$ customer that arrive in the respective restaurants. Hence:

$$p(R_{i,j} \mid \boldsymbol{R}, \boldsymbol{I}) = \sum_{I_{i,m_i+1}=1, I_{j,m_j+1}} p(R_{i,j} \mid I_{i,m_i+1}, I_{j,m_j+1}, \boldsymbol{R}, \boldsymbol{I}) p(I_{i,m_i+1}, I_{j,m_j+1} \mid \boldsymbol{R}, \boldsymbol{I}) \tag{14}$$

where the summation is over all possible values that these variables can take which is equal to the number of existing dishes $K + 2$. The addition of 2 reflects the fact that the new customers, i.e. I_{i,m_i+1}, I_{j,m_j+1} can be assigned to new tables and new dishes based on the CRP priors adopted. Due to exchangeability and the fact that the new latent variables do not affect the likelihood, the second term in the above sum over all possible values k_1 and k_2 can be expressed as:

$$p(I_{i,m_i+1} = k_1, I_{j,m_j+1} = k_2 \mid \boldsymbol{R}, \boldsymbol{I}) = p(I_{j,m_j+1} = k_2 \mid I_{i,m_i+1} = k_1, \boldsymbol{I}) p(I_{i,m_i+1} = k_1 \mid \boldsymbol{I}) \tag{15}$$

where each of the factors can be calculated as explained in the previous section.

Furthermore, for $R_{i,j} = 1$ the first term in the sum of Equation (14) can be calculated as follows (see also Equation (7)):

$$\begin{aligned} p(R_{i,j} = 1 \mid I_{i,m_i+1} = k_1, I_{j,m_j+1} = k_2, \boldsymbol{R}, \boldsymbol{I}) &= \frac{p(R_{i,j} = 1, \boldsymbol{R} \mid I_{i,m_i+1} = k_1, I_{j,m_j+1} = k_2, \boldsymbol{I})}{p(\boldsymbol{R} \mid I_{i,m_i+1}, I_{j,m_j+1}, \boldsymbol{I})} \\ &= \frac{m_1(k_1, k_2) + \beta_2 + 1}{m_0(k_1, k_2) + \beta_1 + m_1(k_1, k_2) + \beta_2 + 1} \end{aligned} \tag{16}$$

where $m_0(k_1, k_2)$ and $m_1(k_1, k_2)$ are the counts of 0 and 1 links in the observable data $\boldsymbol{R}$ between groups k_1 and k_2 respectively. It should be noted that if k_1 or k_2 are greater than K (i.e. the number of existing dishes/groups) then these counts are zero.

3.3 Non-identifiability

Before we embark with the presentation of numerical examples we discuss the issue of non-identifiability of the model, meaning its inability in certain cases where artificial data is used to find the original structure i.e. actual identities and degrees of membership. This is a problematic feature of the model but also a testament to its versatility and expressibility.

Non-identifiability arises from the fact that the observables (i.e. the links) can be generated by several different configurations if one allows the degree of membership and the number of groups to vary. To illustrate this consider a dataset which consists of two objects and two groups/identities. The first belongs exclusively to group 1 whereas the second by 50% to group 1 and 50% to group 2. Assume also that the probability of a link is 1 between objects in the same group and 0 otherwise. In order to generate values for the two possible links (i.e. 1 → 2 and 2 → 1) we first sample the associated

Number of					Log-Likelihood
Groups/Identities	$I_{1,1}$	$I_{1,2}$	$I_{2,1}$	$I_{2,2}$	
4	1	2	3	4	-1.39
3	1	2	3	3	-1.39
	1	3	2	3	-1.39
	3	1	2	3	-1.39
	3	3	1	2	-1.39
	1	3	3	2	-1.39
	3	1	3	2	-1.39
2	1	1	2	2	-1.39
	1	2	1	2	-1.39
	1	2	2	1	-3.18
	1	2	2	2	-1.39
	1	1	1	2	-1.39
	1	1	2	1	-1.39
	1	2	1	1	-1.39
	2	1	1	1	-1.39
1	1	1	1	1	-3.18

Table 1: Possible configurations

latent variables based on the degrees of membership above and get $I_{1,1} = I_{1,2} = 1$ for the first object and $I_{2,1} = 1,\ I_{2,2} = 2$ for the second. This implies $R^1_{1,2} = 1$ and $R^2_{2,1} = 0$. In table 1 we enumerate all the possible group/identity allocations and the respective log-likelihood based on Equation (2) (for $\beta_1 = \beta_2 = 0.1$). As it can be seen, the actual configuration $(1, 1, 1, 2)$ is equivalent with 11 others which correspond to 2, 3 and 4 groups/identities. Even the ones with 2 groups might lead to degrees of membership which are different from the ground truth.

This non-identifiability can also be explained by the the multiple ways of defining

	G1	G2	G1’
Object 1	0.5	0.5	1.0
Object 2	0.5	0.5	1.0

Table 2: Degree of Membership Matrix

the identity components. To illustrate this we consider an example that might seem trivial but nevertheless reveals the various possibilities that exist. We consider again a dataset with two objects and two identity components, say G1 and G2 with identical degrees of memberships as summarized in the left part of Table 2. It is obvious that we can instead define a new identity G1’ as G1’=0.5G1+0.5G2 i.e. that consists 50% of the identity component G1 and 50% of the identity component G2. In this case, both objects will belong exclusively to G1’ (see Table 2) and the model will adjust the $\boldsymbol{\eta}$ matrix in order to reflect the observed data. Naturally, the opposite scenario can also take place i.e. starting with one identity component and finding two. An intermediate solution would be to impose a restrictive prior on the β_1 and β_2 parameters in order to favor particular group structures (for example small βs imply rather clearly separated groups that have either one or zero probability of a link). This however would imply that prior beliefs have a larger weight than the data in inferring the underlying structure. Instead we set in advance the identities $\{I_{i,m}\}_{m=1}^{m_i}$ of an arbitrarily selected object i equal to 1. This definitely alleviates the identifiability issues described above but might not be sufficient. Naturally if prior knowledge about group assignments exists, this can be utilized at this step.

3.4 Numerical Examples

In the following several numerical examples are presented on synthetic and real-world data. In all cases the degree of membership is calculated for the maximum likelihood configuration The latter was not found by an optimization algorithm but was selected from 10 independent runs with $20,000$ MCMC iterations each. A standard version of

Object Set	Identity 1	Identity 2	Identity 3	Identity 4
Set 1 (Objects 1-4)	1.0	0.0	0.0	0.0
Set 2 (Objects 5-8)	0.2	0.8	0.0	0.0
Set 3 (Objects 9-12)	0.1	0.1	0.8	0.0
Set 4 (Objects 13-16)	0.1	0.1	0.0	0.8

Table 3: Degree of Membership Matrix

simulated annealing was used in order to avoid local modes with an initial temperature of 100 and reduction factor 0.995. In all cases the following hyper-priors were used:

- for β_1, β_2 : independent $Poisson(0.1)$
- for a_0 and a_i's: independent $Gamma(0.5, 0.5)$

3.4.1 Example 1: Artificial Data

An artificial dataset consisting of 16 objects and 4 identities (groups) was constructed. These were divided into four sets (set1 through set4) each consisting of 4 nodes. The degree of membership of each set to the 4 groups can be seen in Table 3. A matrix of probabilities of links between any pair of identities was also generated from a $Beta(0.1, 0.1)$ and links were drawn. The full adjacency matrix was then given to the model.

In Figure 1 the posterior on the number of identities is compared with the result from the IRM model. It can be readily seen that the mixed membership model correctly assigns a higher probability to the true value 4. Furthermore the mode of the posterior for IRM is located at 3. This is to be expected as IRM is not capable of accounting for the mixed-membership of the participating objects.

The maximum likelihood configuration correctly identified four groups and the degrees of membership of each of the 16 to the 4 groups is depicted in Figure 2. It can be clearly seen that it correctly identifies that objects $1-4$ belong (almost) exclusively to group 1. Furthermore the results show good agreement with Table 3 with

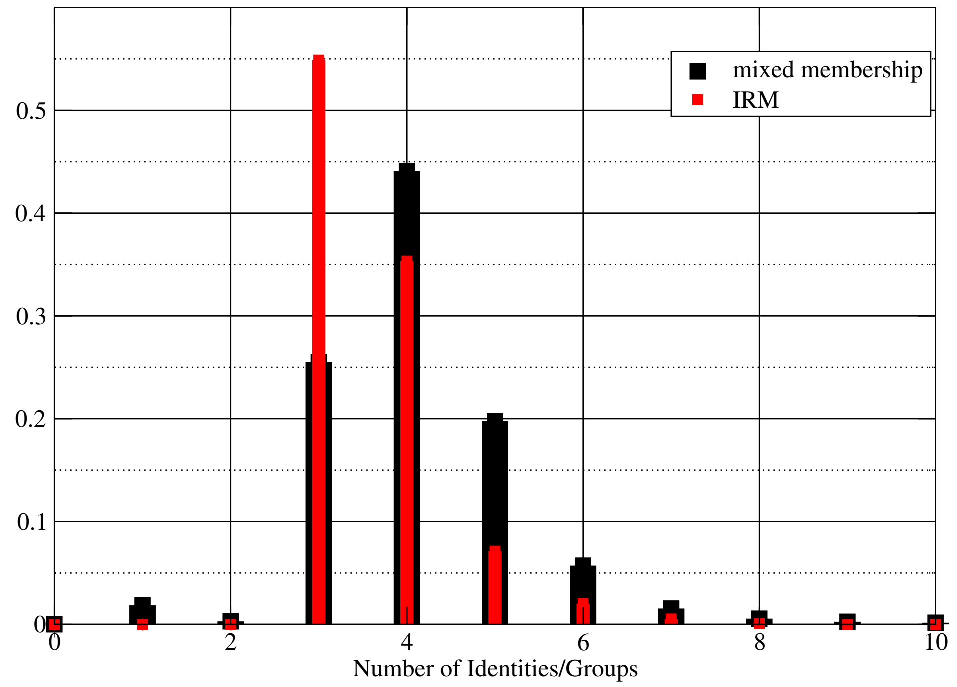


Figure 1: Posterior on the number of identities/groups

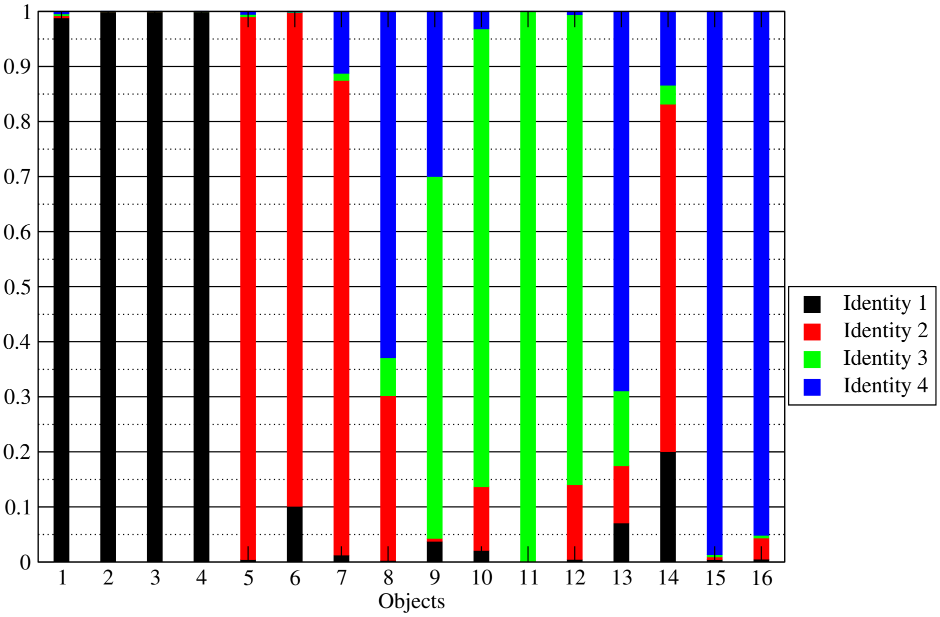


Figure 2: Posterior on the number of identities/groups

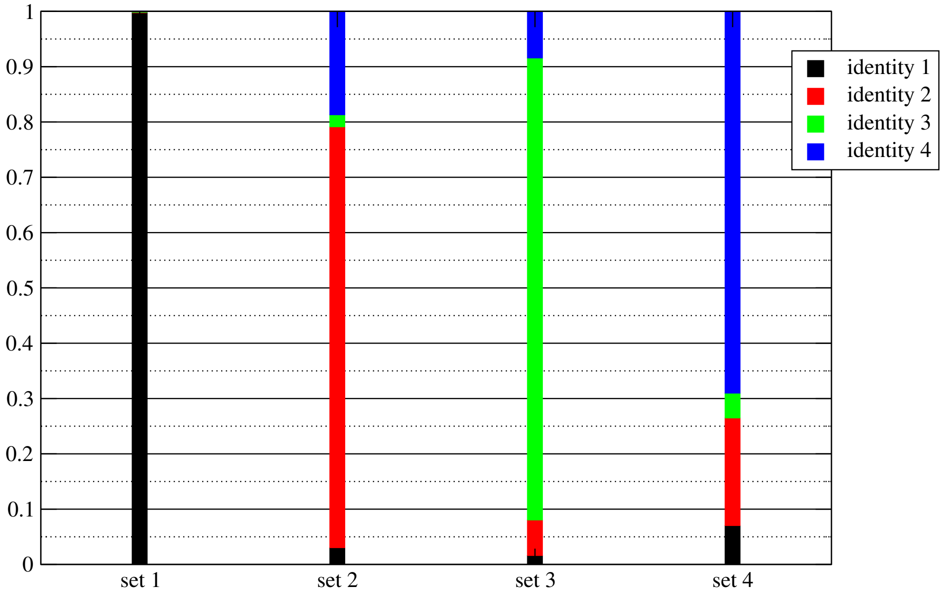


Figure 3: Posterior on the number of identities/groups

the exception of objects 8 (membership to group 2 is underestimated and to groups 3 and 4 overestimated), 13 (membership to group 3 overestimated), 15 (membership to group 4 underestimated and to groups 1 and 2 overestimated). This can also be seen in Figure 3 where the degrees of membership have been averaged over all objects belonging to each of the four sets. The discrepancy can be attributed to the fact the actual max. likelihood configuration is not the one found by the algorithm.

We are also able to calculate from posterior samples the probabilities that any pair of objects belong to the same group. These can be ordered on a 16 × 16 matrix $\boldsymbol{P_{mm}}$ and compared with the actual probabilities $\boldsymbol{P_0}$ based on Table 3. For example the probability that an object from set 2 is in the same group with an object from set 3 is $0.2 \times 0.1 + 0.8 \times 0.1 + 0.0 \times 0.8 = 0.1$. The absolute value of the deviation between exact probabilities and the ones calculated from the model are depicted in Figure 4 for all pairs of objects, i.e. $\mid P_{mm}^{i,j} - P_0^{i,j} \mid$ and are compared with the error from the IRM model $\mid P_{IRM}^{i,j} - P_0^{i,j} \mid$. The errors are much smaller for the mixed-membership model. In fact ratio of the error norms is $\frac{||\boldsymbol{P_{mm}}-\boldsymbol{P_0}||}{||\boldsymbol{P_{IRM}}-\boldsymbol{P_0}||} \approx 0.49$.

The same problem was examined with different degrees of membership following

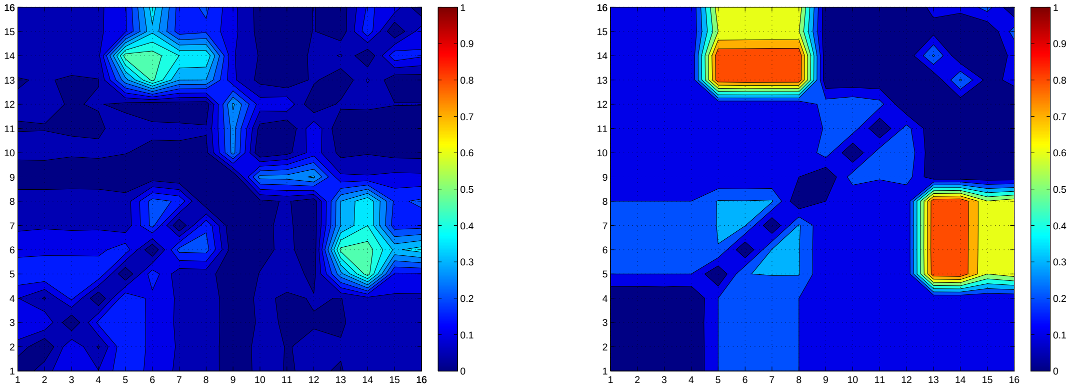


Figure 4: $| P^{i,j}_{mm} - P^{i,j}_0 |$ (left) and $| P^{i,j}_{IRM} - P^{i,j}_0 |$ (right) for all pairs of objects i, j

Object Set	Identity 1	Identity 2	Identity 3	Identity 4
Set 1 (Objects 1-4)	1.0	0.0	0.0	0.0
Set 2 (Objects 5-8)	0.5	0.5	0.0	0.0
Set 3 (Objects 9-12)	0.25	0.25	0.5	0.0
Set 4 (Objects 13-16)	0.25	0.25	0.0	0.5

Table 4: Degree of Membership Matrix

Table 4. The error in the predictions between the mixed-membership and the IRM model are depicted in Figure 5 in a manner identical to that of Figure 4. Again, the mixed membership model produces smaller deviations from the ground truth and the ratio of error norms is now $\frac{||\boldsymbol{P_{mm}}-\boldsymbol{P_0}||}{||\boldsymbol{P_{IRM}}-\boldsymbol{P_0}||} \approx 0.66$

In most cases of practical interest, a significant portion of the possible links is unobserved. It is important therefore to examine the robustness of the model in terms of discovering the underlying structure and also its ability to predict those missing links. In order to explore this issue we consider the first case of the aforementioned dataset (Table 3) and hide some of the links. In particular each generated link value is hidden independently with probability 12.5% and the remaining links are the observables that the model is given. Table 5 contains the average error norm with respect to $\boldsymbol{P_0}$ as obtained by averaging over 5 independent tests. Furthermore we also present aver-

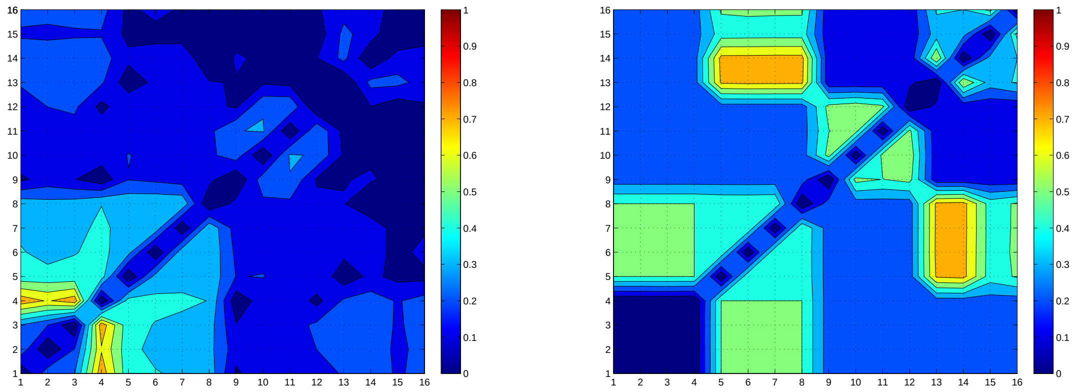


Figure 5: $| P_{mm}^{i,j} - P_0^{i,j} |$ (left) and $| P_{IRM}^{i,j} - P_0^{i,j} |$ (right) for all pairs of objects i, j

	Error Norm	AUC
IRM	0.58	0.85
Mixed Membership	0.60	0.83

Table 5: Comparison of performance when 12.5% of the links are missing

aged values of the AUC metric for the predictions regarding the hidden links which is calculated from the ROC curve as in classification models.

As it can be readily the IRM outperforms the mixed membership model with respect to both metrics. The same outcome is observed in Table 6 which depicts results obtained when 25% of the links are missing.

We repeated the experiment on a larger dataset that consisted of 40 objects broken up in 4 sets of 10 objects each that exhibit the same membership characteristics as in the dataset with the 16 objects (see Table 7). The averaged results over 5 independent runs are summarized in Table 8). The IRM seems to perform slightly better in terms of the AUC metric but significantly worse in terms of the error norm in the $\boldsymbol{P_0}$ matrix. The same conclusions can be drawn from Table 7 which contains results with 50% of the links hidden.

	Error Norm	AUC
IRM	0.63	0.83
Mixed Membership	0.72	0.81

Table 6: Comparison of performance when 25.0% of the links are missing

Object Set	Identity 1	Identity 2	Identity 3	Identity 4
Set 1 (Objects 1-10)	1.0	0.0	0.0	0.0
Set 2 (Objects 11-20)	0.2	0.8	0.0	0.0
Set 3 (Objects 21-30)	0.1	0.1	0.8	0.0
Set 4 (Objects 31-40)	0.1	0.1	0.0	0.8

Table 7: Degree of Membership Matrix

	Error Norm	AUC
IRM	0.50	0.88
Mixed Membership	0.40	0.87

Table 8: Comparison of performance when 25.0% of the links are missing in the dataset based on Table 7

	Error Norm	AUC
IRM	0.52	0.88
Mixed Membership	0.42	0.87

Table 9: Comparison of performance when 50.0% of the links are missing in the dataset based on Table 7

3.4.2 Example 2 : Monk Data

In the second example we consider a well-studied social network consisting of 18 monk residing in the same cloister ([21, 11]). Various sociometric relationship were recorded at different times (such as like, dislike, esteem, disesteem etc) by asking each member to rank only his top three choices for each relation type. Herein we present results based on the "like" data collected at the last period. We assigned a 1 link from monk i to monk j if the latter was in the top three choices of the former and a 0 link otherwise. Sampson ([21]) identified three basic groups: a) the Young Turks consisting of monks 1,2, 7, 12, 14, 15 and 16, b) the Loyal Opposition consisting of monks 4, 5, 6, 9, 11 and c) the Outcasts consisting of monks 3, 17 and 18. Monks 8, 10, 13 seem to waver between the Young Turks and the Loyal Opposition which he described as being in intense conflict.

Figure 6 depicts the probability that any pair of monks belong to the same group averaged over the samples from the posterior. As it can be seen the model has identified three basic groups consisting of the following monks: a) 1, 2, 7, 12, 14, 15, 16 (all Young Turks) b) 3, 17, 18 (all Outcasts) and c) 4 ,5, 6, 8, 9, 10, 11 (all the members of Loyal Opposition and two Waverers). Also monk 13 (Waverer) does not seem to belong to any particular group.

3.4.3 Example 3 : Zachary's Karate Club

We consider Zachary's karate club, a well studied social network which is based on the data collected and analyzed in [29]. It consists of 34 individuals which initially belonged to the same club but due to a disagreement between the administrator (object 34) and the instructor (object 1) ended up splitting in two as illustrated in Figure 7 ([10]). The members that aligned with the instructor are marked with squares (group 1) and the members that favored the administrator are marked with circles (group 2). Furthermore Figure 7 give us an idea about the topology of the network that was estimated using some friendship measures ([10]). For example even though individuals

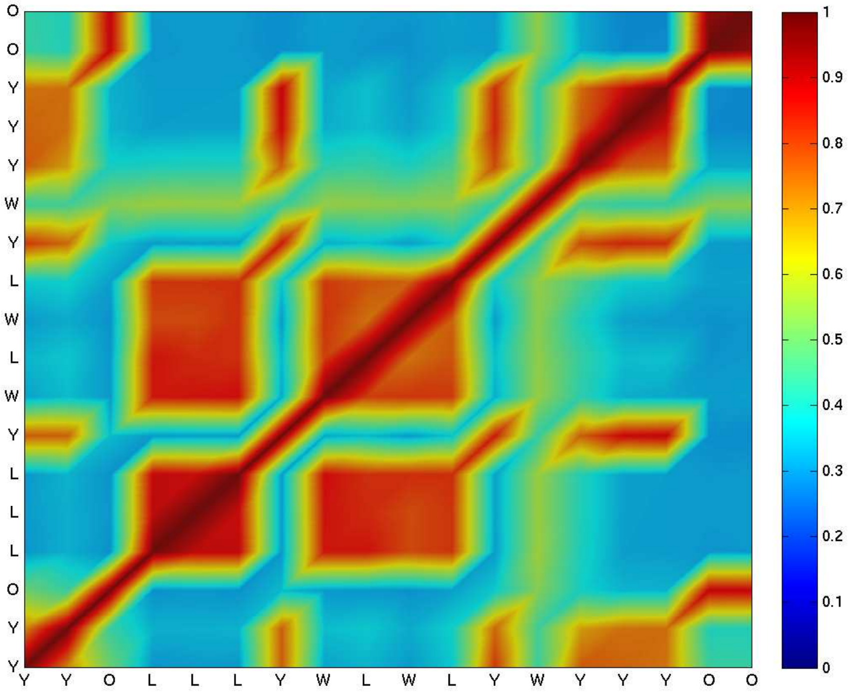


Figure 6: Probability that any pair of monks belong to the same group averaged over the samples from the posterior (Labels Y: Young Turks, O: Outcasts, L: Loyal Opposition and W:Waverer)

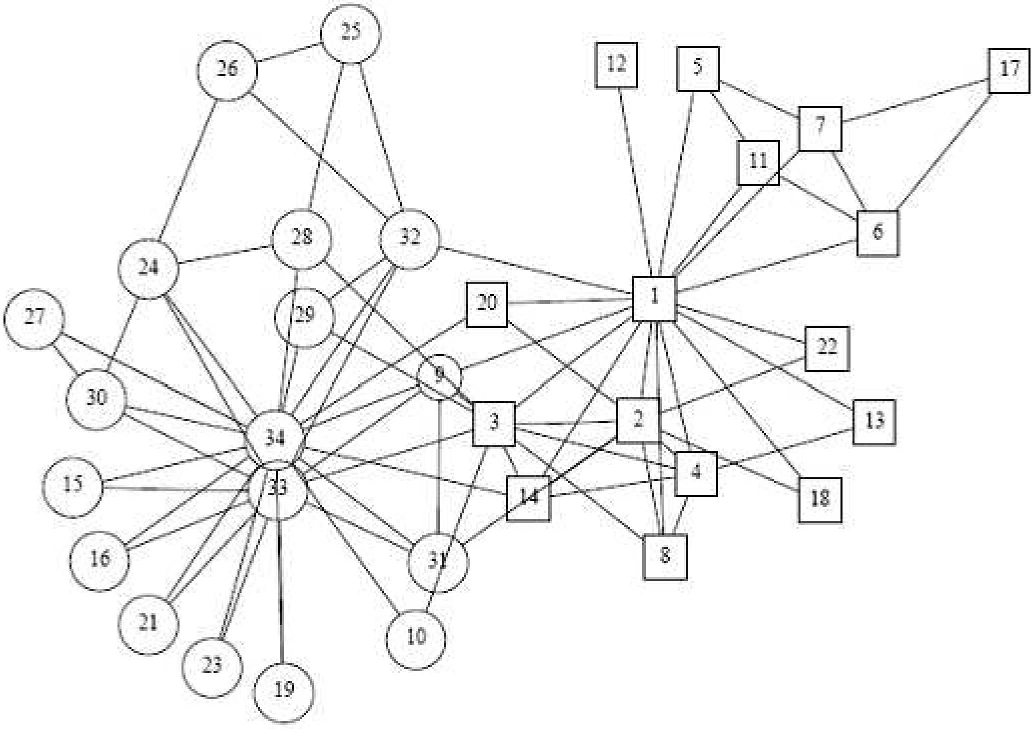


Figure 7: Friendship network for Zachary's karate club. Figure taken from [10].

3 and 13 belong to the same group, the former seems closer to group 2 than the latter.

We used a binary version of the friendship links as observables for our mixed-membership model. The maximum likelihood configuration identified 5 groups/identities. This is expected as apart from the 2 main groups/identities there are several smaller sub-groups as it can also be seen in Figure 7. The degrees of membership for each object are depicted in Figure 8. It is clear however that identities 1 and 2 are mostly associated with group 1 whereas identities 3, 4 and 5 with group 2. If we depict each set of those identities with a single color as in Figure 9 one observes that the partitioning corresponds to the actual two groups. In fact if we assign each individual to a group based on whether they are mostly red or blue (Figure 9) then we will recover *exactly* the two groups in Figure 7. The mixed-membership model provides however far more information. It identifies hidden identities and quantifies the degree to which

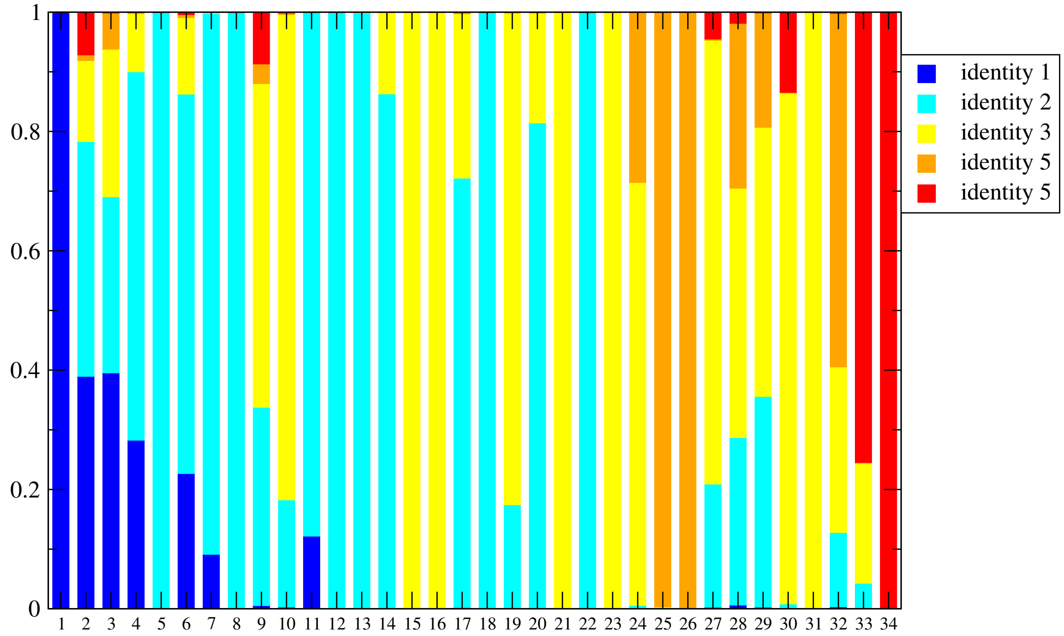


Figure 8: Degrees of Membership for maximum likelihood configuration

these are exhibited in each individual. For example, as observed earlier individuals 3 and 13 belong to the group 1, but 3 appears to have closer ties with group 2 than 13 (Figure 7). This is reflected also in Figure 9 where it can be seen that 3 belongs up to 25% to group 2 whereas for 13 this percentage is 0. Several other individuals such as 3, 14, 20 from group 1 and 9, 10, 32 from group 2 which appear to lie somewhere in the middle in Figure 7, also exhibit significant percentages of the other group in their identity (Figure 9). In the same manner, individuals 7,8 11, 12, 13, 18 from group 1 and 15, 16 21, 23, 24, 25, 26 from group 2, which seem to have no friendships with members of the other group, appear to have a single component in their identity. It is also worth mentioning that the pivotal individuals 1 and 34 exhibit exclusively their respective groups identities.

Finally, Figure 10 depicts the probability that any pair of people belongs to the same group averaged over the samples from the posterior. As it can be seen the model has identified 7 basic groups consisting of the following people: a) 1 (group 1), b) 2, 3, 4, 8, 14 (group 1), c) 5, 6, 7, 11, 17 (group 1), d) 9, 10, 15, 16, 19, 21, 23, 27, 28, 29, 30, 31 (group 2), e) 12, 13, 18, 20, 22 (group 1) f) 25, 26 (group 2), g) 33, 34 (group

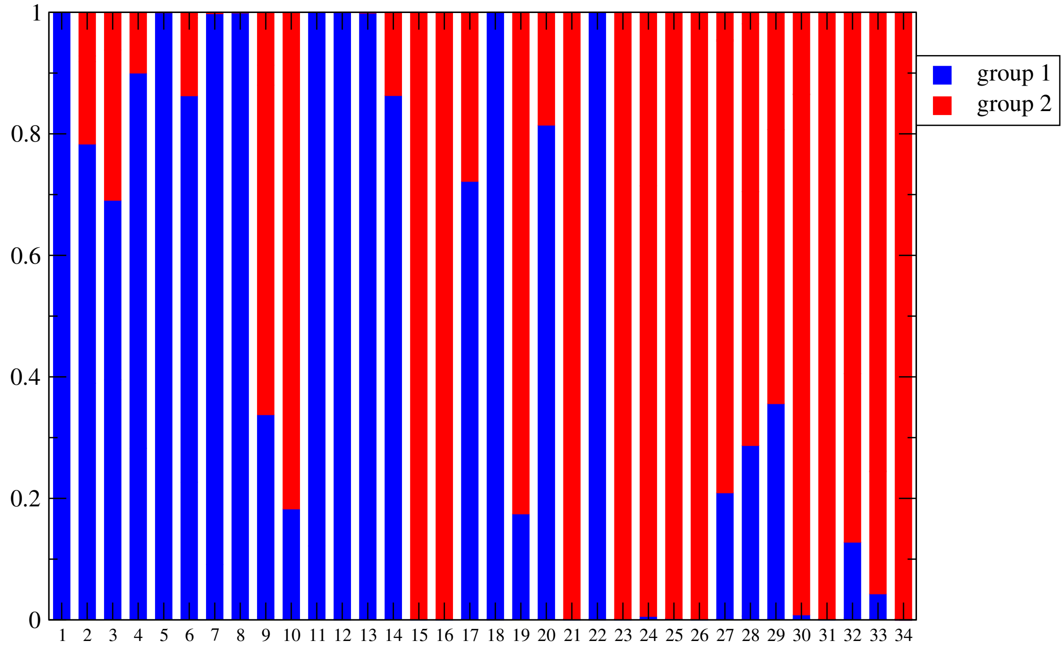


Figure 9: Degrees of Membership for maximum likelihood configuration

2). Also objects 24 and 32 seem to be by themselves. It is worth pointing out that none of the aforementioned groups is mixed i.e. none contains members from the two fractions. It is also significant that the analysis provides provides quantitative evidence on the strength of the bond between any pair of people.

3.4.4 Example 4: Animal-Feature Dataset

In this subsection we consider an example consisting of two domains, i.e. objects of two different types. In particular the first domain is made up of 16 animals and the second by 13 features. Binary links have been established based on whether the animal has the particular feature and can be seen in Figure 11. The degrees of membership for the maximum likelihood configuration are depicted in Figures 12 and 14. The model identifies 4 and 8 groups/identities for the animal and feature domains respectively. It is clear that in the animal domain, identity 1 is primarily associated with the birds whereas identities 2, 3 and 4 with the 4-legged mammals. If we depict each set of those identities with a single color as in Figure 12 one observes that the partitioning corre-

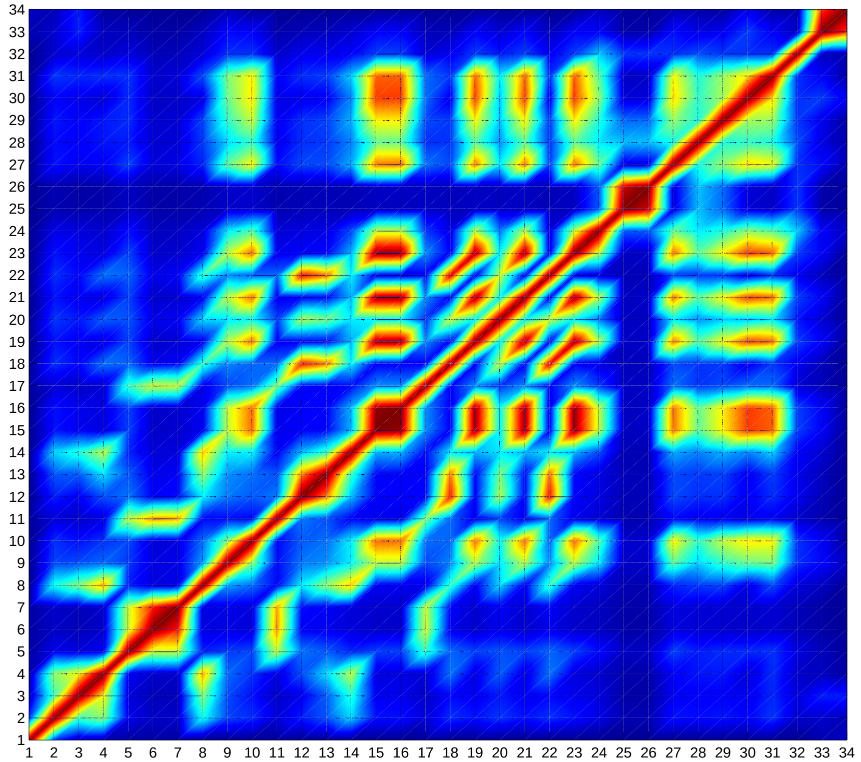
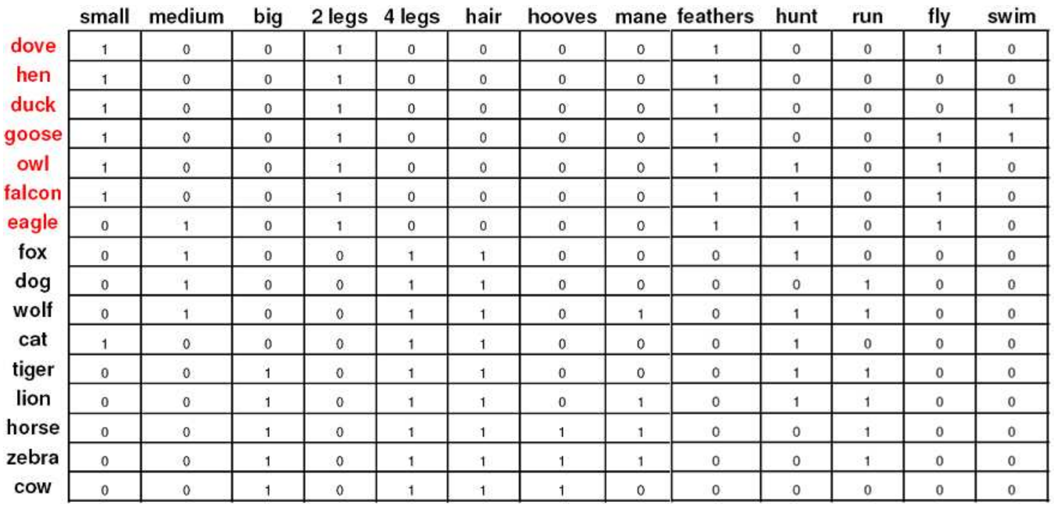


Figure 10: Probability that any pair of people belong to the same group averaged over the samples from the posterior



	small	medium	big	2 legs	4 legs	hair	hooves	mane	feathers	hunt	run	fly	swim
dove	1	0	0	1	0	0	0	0	1	0	0	1	0
hen	1	0	0	1	0	0	0	0	1	0	0	0	0
duck	1	0	0	1	0	0	0	0	1	0	0	0	1
goose	1	0	0	1	0	0	0	0	1	0	0	1	1
owl	1	0	0	1	0	0	0	0	1	1	0	1	0
falcon	1	0	0	1	0	0	0	0	1	1	0	1	0
eagle	0	1	0	1	0	0	0	0	1	1	0	1	0
fox	0	1	0	0	1	1	0	0	0	1	0	0	0
dog	0	1	0	0	1	1	0	0	0	0	1	0	0
wolf	0	1	0	0	1	1	0	1	0	1	1	0	0
cat	1	0	0	0	1	1	0	0	0	1	0	0	0
tiger	0	0	1	0	1	1	0	0	0	1	1	0	0
lion	0	0	1	0	1	1	0	1	0	1	1	0	0
horse	0	0	1	0	1	1	1	1	0	0	1	0	0
zebra	0	0	1	0	1	1	1	1	0	0	1	0	0
cow	0	0	1	0	1	1	1	0	0	0	0	0	0

Figure 11: Animal-Feature Dataset

sponds to the basic two groups i.e. the birds and the 4-legged mammals. Furthermore, the model is able to identify mixed-membership effects for the eagle (which has the basic feature of birds, but shares some common features with mammals such medium size, hunt, not swim), the cat (which has the basic feature of mammals, but shares some common features with the birds such as small size, not run).

Similarly in the feature domain (Figure 14) the model identifies to basic identities namely, identity 4 which consists of small, 2-legs, feathers and fly (to a certain extent) that are features shared by the birds and identity 7 which consists of 4-legs, hair, big and to a lesser extent medium, hooves, hunt and run that are shares mostly by the mammals. It also identifies a number of other less prevalent identities that encapsulate the nuances in the dataset.

3.5 Variational Inference

As mentionned earlier Gibbs sampling can be ineffective as the number of latent variables is proportional to the observed links. Advanced MCMC schemes ([14]) could potentially alleviate this problem but have not been explored thus far. It is unlikely though that the impovement in effiency would be such to allow processing large datasets

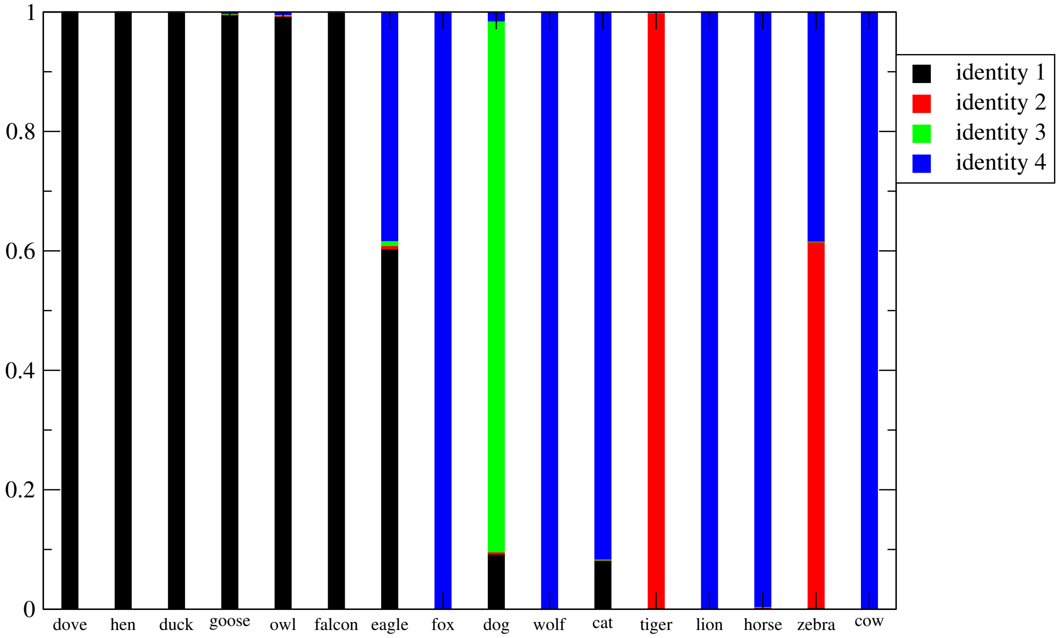


Figure 12: Animal Domain - Degrees of Membership for maximum likelihood configuration

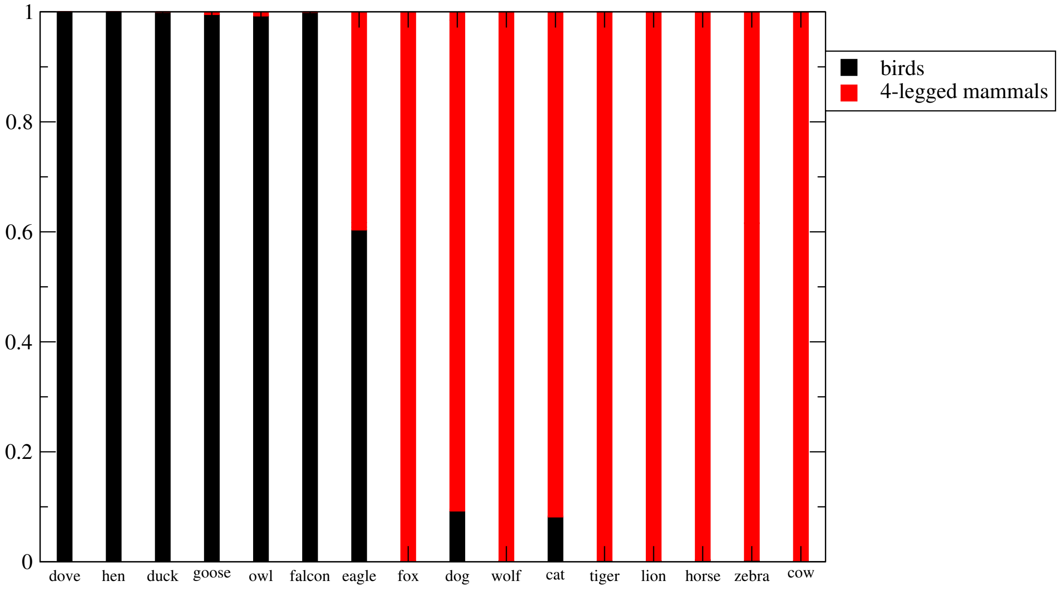


Figure 13: Animal Domain - Degrees of Membership for maximum likelihood configuration

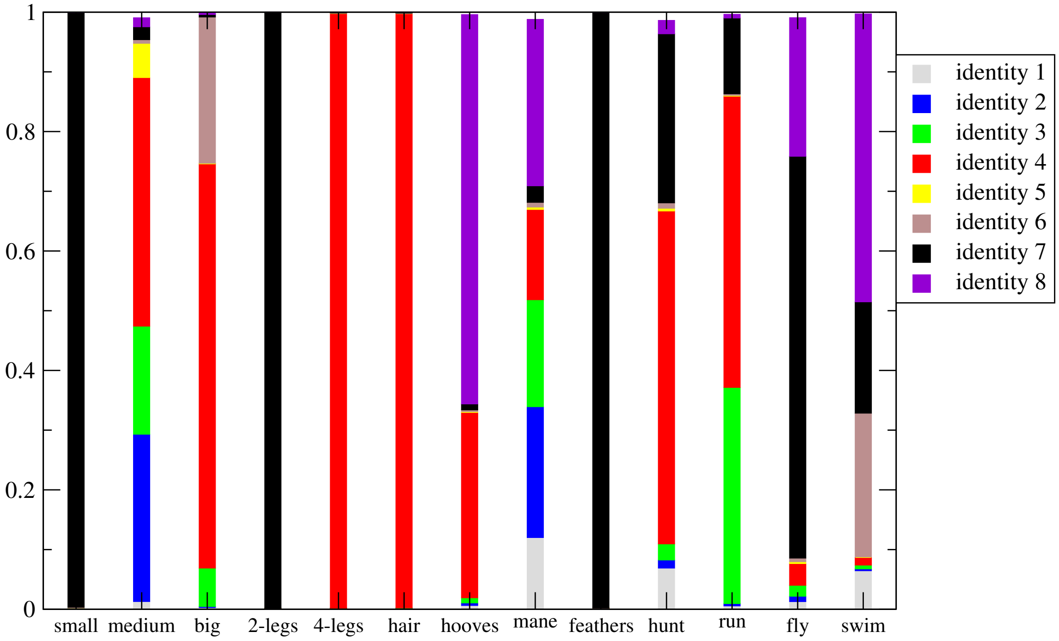


Figure 14: Features Domain - Degrees of Membership for maximum likelihood configuration

(> 1000 objects) in reasonable times. For that purpose we explore here the possibility of approximate variational inference which has proven successful in several models ([23]). The basic idea is to approximate the posterior with a simple distribution depending on a finite number of parameters which can then be determined by minimizing the Kullback-Leibler divergence of the two distributions. It can also be shown that this approach can readily provide a lower bound on the marginal log-likelihood of the data ([9]).

Although variational methods have been fairly well established for traditional Bayesian models, their implementation in nonparametric models, let alone hierarchical ones, is still at its infancy. In the present formulation we define an explicit truncation level, following Blei and Jordan ([5]), but do not include the stick breaking weights in the latent variables. One one hand this leads to more complex expressions which can be adequately approximated as it will be seen in the sequence. On the other hand, incorporating those weights would have required an unrealistic assumption for their variational distribution which would have compromized the accuracy of the overall formulation.

Following the exposition for the CRF and ignoring for the time-being the hyperparameters of the model, we identify three sets of parameters, namely the table assignements $\boldsymbol{t} = \{t_{i,m}\}$ (Equation (9)), the dish assignments $\boldsymbol{d} = \{d_{i,t}\}$ (Equation (10)) and the pairwise probabilities of links between groups $\boldsymbol{\eta} = \{\eta(I,J)\}$ (section 2). The posterior $p(\boldsymbol{Z} \mid \boldsymbol{R})$ where $\boldsymbol{Z} = (\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{\eta})$ is proportional to:

$$p(\boldsymbol{Z} \mid \boldsymbol{R}) \propto p(\boldsymbol{R} \mid \boldsymbol{Z}) p(\boldsymbol{t}, \boldsymbol{d}) p(\boldsymbol{\eta}) \tag{17}$$

where (the influence of the hyperparameters has been supressed) and $p(\boldsymbol{R} \mid \boldsymbol{Z}) = \prod_{I,J} \eta(i,J)^{m_0(I,J)}(1-\eta(I,J))^{m_1(I,J)}$ is the likelihood function (where the product is taken over all pairs of groups/identities). The prior on $\boldsymbol{\eta}$ is simply the product of $Beta's$, i.e. $p(\boldsymbol{\eta}) = \prod_{I,J} \frac{\eta(i,J)^{\beta_1-1}(1-\eta(I,J))^{\beta_2-1}}{\beta(\beta_1,\beta_2)}$ (section 2). Finally, even though the prior $p(\boldsymbol{t}, \boldsymbol{d}) = p(\boldsymbol{t})p(\boldsymbol{d})$ cannot in general be written explicitly, we only make use of the conditionals as those in Equations (9) and (10) as it will be seen in the sequence.

The approximating mean-field variational approximation $Q(\boldsymbol{Z})$ is assumed to have the following form:

$$Q(\boldsymbol{Z}) = Q(\boldsymbol{t}) Q(\boldsymbol{d}) Q(\boldsymbol{\eta}) \tag{18}$$

where:

- $Q(\boldsymbol{t}) = \prod_i \prod_{m=1}^{m_i} q_{i,m}(t_{i,m})$
- $Q(\boldsymbol{d}) = \prod_i \prod_{t=1}^{T} q_{i,t}(d_{i,t})$
- $Q(\boldsymbol{\eta}) = \prod_{I,J=1}^{K} q_{I,J}(\eta_{I,J})$

As it can be seen in the equations above we have truncated the base abd restaurant-level CRPs to finite K and T components assuming the have zero power beyond these values. Minimizing the Kullback-Leibler divergence between the posterior $p(\boldsymbol{Z} \mid \boldsymbol{R})$ and $Q(\boldsymbol{Z})$ requires minimization of the following quantity:

$$\begin{aligned} B(Q) &= E_Q\left[\log \frac{Q}{p}\right] \\ &= E_Q[\log Q(\boldsymbol{t})] + E_Q[\log Q(\boldsymbol{d})] + E_Q[\log Q(\boldsymbol{\eta})] \\ &- E_Q[\log p(\boldsymbol{R} \mid \boldsymbol{Z})] - E_Q[\log p(\boldsymbol{t})] - E_Q[\log p(\boldsymbol{d})] - E_Q[\log \ p(\boldsymbol{\eta})] \end{aligned} \tag{19}$$

Minimization is performed by taking the derivarives w.r.t the unknown distributions Q.

a) Update Equations for $Q(\boldsymbol{\eta})$

In particular, for $Q(\boldsymbol{\eta})$ the following three terms are relevant:

- $$\begin{aligned} E_Q[\log p(\boldsymbol{R} \mid \boldsymbol{Z})] &= \sum_{I,J=1}^{K} E_Q[m_1(I,J)\log\eta(I,J) + m_0(I,J)\log(1-\eta(I,J))] \\ &= \sum_{I,J=1}^{K} \big(E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_1(I,J)] E_{q_{I,J}}[\log\eta(I,J)] \\ &+ E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_0(I,J)] E_{q_{I,J}}[\log(1-\eta(I,J))] \big) \end{aligned} \tag{20}$$

- $$E_Q[\log\ p(\boldsymbol{\eta})] = \sum_{I,J=1}^{K} \big((\beta_1 - 1)\ E_{q_{I,J}}[\log\eta(I,J)] + (\beta_2 - 1)\ E_{q_{I,J}}[\log(1-\eta(I,J))] \big) \tag{21}$$

- $$E_Q[\log Q(\boldsymbol{\eta})] = \sum_{I,J=1}^{K} E_{q_{I,J}}[\log q_{I,J}] \tag{22}$$

Differentiation w.r.t. $q_{I,J}(\eta)$ leads to a *Beta* distribution:

$$q_{I,J}(\eta) \propto \eta^{E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_1(I,J)]+\beta_1-1}\ (1-\eta)^{E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_0(I,J)]+\beta_2-1} \tag{23}$$

In order to calculate the expectations under Q appearing in the exponents, we express them as:

$$m_1(I,J) = \sum_{m:R_{i,j}^m=1} 1_{d_{i,t_{i,m}}=I} 1_{d_{j,t_{j,m}}=J} \tag{24}$$

where the summation is over all 1 links and the indicator functions become one when the objects participating in the link are assigned to groups I and J which is equivalent to the respective customers being served dishes I and J. Under Q these two terms are independent and therefore the expectation is equal to sum of the products of the following probabilities:

$$u_{i,m}(I) = Pr[1_{d_{i,t_{i,m}}=I} = 1] = \sum_{t=1,T} q_{i,m}(t_{i,m} = t)q_{i,t}(d_{i,t} = I) \tag{25}$$

From Equation (24):

$$E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_1(I,J)] = \sum_{m:R^m_{i,j}=1} u_{i,m}(I)u_{j,m}(J) \tag{26}$$

and similarly:

$$E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_0(I,J)] = \sum_{m:R^m_{i,j}=0} u_{i,m}(I)u_{j,m}(J) \tag{27}$$

b) Update Equations for $Q(\boldsymbol{t})$

Taking derivatives w.r.t. $Q(\boldsymbol{t})$ involves the first, fourth and fith term of Equation (19). In order to derive an update equation for $q_{i,m}(t_0)$, without loss of generality we can assume that $R^m_{i,j} = 1$. From Equations (25) and (26) we get that:

$$\frac{\partial E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_1(I,J)]}{\partial q_{i,m}(t_0)} = q_{i,t_0}(I)u_{j,m}(J) \tag{28}$$

which, based on Equation (20) implies that:

$$\frac{\partial \log p(\boldsymbol{R} \mid \boldsymbol{Z})}{\partial q_{i,m}(t_0)} = \sum_{I,J=1}^{K} q_{i,t_0}(I)u_{j,m}(J)E_{q_{I,J}}[\log \eta(I,J)] \tag{29}$$

Trivially for the first term of Equation (19) we get:

$$\frac{\partial E_Q[\log Q(\boldsymbol{t})]}{\partial q_{i,m}(t_0)} = \log q_{i,m}(t_0) + 1 \tag{30}$$

Finally, we can express $p(\boldsymbol{T})$ that appears in Equation (19) as $p(\boldsymbol{t}) = p(t_{i,m} \mid \boldsymbol{t_{-(i,m)}})p(\boldsymbol{t_{-(i,m)}})$ where $\boldsymbol{t_{-(i,m)}}$ represents all the entries in $\boldsymbol{t}$ except for $t_{i,m}$. Hence $E_Q[\log p(\boldsymbol{t})] = E_Q[\log p(t_{i,m} \mid \boldsymbol{t_{-(i,m)}})] + E_Q[\log p(\boldsymbol{t_{-(i,m)}})]$ of which only the first term involves $q_{i,m}$. In order to make expression tractable we adopt a finite symmet4ric dirichlet representation with T clusters according to which:

$$\frac{\partial E_Q[\log p(t_{i,m} \mid \boldsymbol{t_{-(i,m)}})]}{\partial q_{i,m}(t_0)} = E_{Q_{\boldsymbol{t}_{-(i,m)}}}\left[\log \frac{n_{i,t_0+a_i/T}}{m_i - 1 + a_i}\right] \tag{31}$$

It should be noted that the counts n_{i,t_0} above are based on $\boldsymbol{t_{-(i,m)}}$. Collecting the results from Equations (29), (30) and (31) we can conclude that:

$$q_{i,m}(t_0) \propto \exp\left[\sum_{I,J=1}^{K} q_{i,t_0}(I)u_{j,m}(J)E_{q_{I,J}}[\log\eta(I,J)] + E_{Q_{\boldsymbol{t}_{-(i,m)}}}\left[\log\frac{n_{i,t_0}+a_i/T}{m_i-1+a_i}\right]\right] \tag{32}$$

The expectation w.r.t. $Q_{\boldsymbol{t}_{-(i,m)}}$ appears intractable due the large dimensionality of vector $\boldsymbol{t_{-(i,m)}}$. for that purpose we employ a Gaussian approximation in the same manner described in ([16]). According to this the expectation of a functuion $f(x)$ can be approximated as:

$$E[f(x)] \approx f(E[x]) + \frac{1}{2}f''(E[x])Var[x] \tag{33}$$

which is nothing more than a 2^{nd} order Taylor series expansion around $E[x]$. This approximation is valid if higher order moments of x and/or derivatives of f are negligible. In order to apply it to Equation (32) we need to calculate the expectation and variance of $n_{i,t0}$. Since $n_{i,t_0} = \sum_{j\neq m} 1_{t_{i,j}=t_0}$, i.e. a sum of independent variables under $Q_{\boldsymbol{t}_{-(i,m)}}$ these can be expressed as:

$$E[n_{i,t_0}] = \sum_{j\neq m} q_{i,j}(t_0) \tag{34}$$

and:

$$Var[n_{i,t_0}] = \sum_{j\neq m} q_{i,j}(t_0)(1-q_{i,j}(t_0)) \tag{35}$$

c) Update Equations for $Q(\boldsymbol{d})$

Taking derivatives w.r.t. $Q(\boldsymbol{t})$ involves the second, fourth and sixth term of Equation (19). From Equations (25) and (26) we get that:

$$\frac{\partial E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_1(I,J)]}{\partial q_{i,t}(d_0)} = \sum_{m:R_{i,j}^m=1} \frac{\partial u_{i,m}(I)}{\partial q_{i,t}(d_0)} u_{j,m}(J) \tag{36}$$

But from Equation (20) we have:

$$\frac{\partial u_{i,m}(I)}{\partial q_{i,t}(d_0)} = q_{i,m}(t)\delta_{I,d_0} \tag{37}$$

where δ is the Kronecker delta function. This implies that:

$$\frac{\partial E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_1(I,J)]}{\partial q_{i,t}(d_0)} = \delta_{I,d_0} \sum_{m:R^m_{i,j}=1} q_{i,m}(t) u_{j,m}(J) \tag{38}$$

Similarly it can be shown that:

$$\frac{\partial E_{Q(\boldsymbol{t},\boldsymbol{d})}[m_0(I,J)]}{\partial q_{i,t}(d_0)} = \delta_{I,d_0} \sum_{m:R^m_{i,j}=0} q_{i,m}(t) u_{j,m}(J) \tag{39}$$

In summary, from Equation (20):

$$\begin{aligned}\frac{\partial \log p(\boldsymbol{R} \mid \boldsymbol{Z})}{\partial q_{i,t}(d_0)} &= \sum_{J=1}^{K} \Bigg(\sum_{m:R^m_{i,j}=1} q_{i,m}(t) u_{j,m}(J) E_{q_{d_0,J}}[\log \eta(d_0,J)] \\ &+ \sum_{m:R^m_{i,j}=0} q_{i,m}(t) u_{j,m}(J) + E_{q_{d_0,J}}[\log 1 - \eta(d_0,J)] \Bigg) \end{aligned} \tag{40}$$

The derivatives w.r.t. the second term of Equation (19) are trivial. Furthermore, for the effect of the sixth term of Equation (19) we can proceed in a similar manner as in Equation (32), to finnaly arrive at:

$$\begin{aligned} q_{i,m}(t_0) \propto \exp & \Bigg[\sum_{J=1}^{K} \Bigg(\sum_{m:R^m_{i,j}=1} q_{i,m}(t) u_{j,m}(J) E_{q_{d_0,J}}[\log \eta(d_0,J)] \\ &+ \sum_{m:R^m_{i,j}=0} q_{i,m}(t) u_{j,m}(J) + E_{q_{d_0,J}}[\log 1 - \eta(d_0,J)] \Bigg) \\ & E_{Q_{\boldsymbol{d}_{-(i,t)}}} \left[\log \frac{s_{d_0} + a_0/K}{M - 1 + a_0} \right] \Bigg] \end{aligned} \tag{41}$$

where the base CRP has been approximated by a symmetric Dirichlet with K clusters. where $d_{-(i,t)}$ denotes all the entries of $\boldsymbol{d}$ except for $d_{i,t}$ (Equation (10)). In order to calculate the expectation w.r.t. $Q_{\boldsymbol{d}_{-(i,t)}}$ we employ the Gaussian approximation in Equation (33) where:

$$E[s_{d_0}] = \sum_i \sum_t q_{i,t}(d_0) \tag{42}$$

and :

$$Var[s_{d_0}] = \sum_i \sum_t q_{i,t}(d_0)(1 - q_{i,t}(d_0)) \tag{43}$$

4 Hierarchical Mixed Membership Model

In this section we pursue further the mixed-membership modeling formulation by adopting a hierarchical model. For that purpose we assume that the identity of each object is a mixture of a finite number of $L + 1$ components and a vector $\boldsymbol{\theta_l}$ expresses the prevalence of each of those components in the object's identity. As before these components can be shared amongst the objects in the same domain but the proportions can vary from one to another. In order to represent the hierarchical structure of the dataset, objects are associated with branches in an $L + 1$-level tree. At level 0 there is always a single root node. Each of the nodes in the tree represents a different identity/group and an object associated with a branch can belong to any of the identities in that branch. We assume that its identity is a mixture of these groups and the proportions $\boldsymbol{\theta_l}$ specify the degree of membership to each group.

In order to represent this hierarchy we define a novel nonparametric prior on trees with a fixed number of levels but with a potentially infinite number of nodes at each level. Hence the number of nodes at each level is not fixed and can be learned from the data. This results in a very flexible model and a reasonable prior over all $L + 1$-level trees. It should be noted that a similar prior under the name nested CRP has been developed by Blei et al. ([4]). Herein we adopt a different construction which starts at the bottom of the tree i.e. the L^{th} level and moves upwards. Specifically the objects are grouped at the bottom level based on a CRP with parameter a_L. In order to move to the $L - 1$ level, we assume that the groups of level L are a new set of customers which are grouped based on a new CRP with parameter a_{L-1}. Hence the conditional probabilities at level $L-1$ depend now on the number of groups of level L. This process continues until the level 0 is reached where it is assumed that a single group exists. Sampling of configurations from this hierarchical CRP can be readily performed by partitioning first at the bottom level based on Equation (4) and moving upwards. The number of levels $L + 1$ will always be predefined and essentially specifies the level of detail in which we want to decompose the dataset.

Based on this nonparametric prior we propose the following generative model that gives rise to the observed links $\boldsymbol{R}$:

a) Draw an L-level tree from the hierarchical CRP (denoted by hCRP) defined in the previously. This associates each object with a branch. Hence each object is associated with the identity components $z_l^{(i)}$ (where i corresponds to the node and l corresponds to the level) that appear in this branch.

b) For each object i, draw a proportion vector $\boldsymbol{\theta^{(i)}}$ from an $L+1$ dimensional Dirichlet $Dir(\gamma)$.

c) For each link $R_{i,j}^m$ between objects i and j:

   - Select the identity component $z_{l_{i,m}}^{(i)}$ for object i by drawing $l_{i,m}$ from the $Discrete(\boldsymbol{\theta^{(i)}})$. Thus $I_{i,m} = z_{l_{i,m}}^{(i)}$.
   - Select the identity component $z_{l_{j,m}}^{(j)}$ for object j by drawing $l_{j,m}$ from $Discrete(\boldsymbol{\theta^{(j)}})$.. Thus $I_{j,m} = z_{l_{j,m}}^{(j)}$
   - Draw $R_{i,j}^m$ from a $Bernoulli(\eta(I_{i,m}, I_{j,m})$ where η (an in the previous models) expresses the probability of a link between any pair of groups.

The proposed model is also summarized below:

$$\begin{aligned}
z_l^{(i)} \mid \boldsymbol{a} = (a_0, a_1, \ldots, a_L) &\sim hCRP(\boldsymbol{a}) \\
\boldsymbol{\theta^{(i)}} \mid \gamma &\sim Dirichlet(\gamma) \\
l_{i,m} \mid \boldsymbol{\theta^{(i)}} &\sim Discrete(\theta_0^{(i)}, \ldots, \theta_L^{(i)}) \\
I_{i,m} \mid z_l^{(i)}, l_{i,m} &\sim I_{i,m} = z_{l_{i,m}}^{(i)} \\
\eta(I_1, I_2) \mid \beta_1, \beta_2 &\sim Beta(\beta_1, \beta_2) \\
R_{i,m} \mid I_{i,m}, I_{j,m}, \boldsymbol{\eta} &\sim Bernoulli\left(\eta(I_{i,m}, I_{j,m})\right)
\end{aligned} \tag{44}$$

As it can be seen if a certain identity component is present in all the objects in the domain, it will appear at the top of the tree which is the only node that is shared by all branches. More obscure identity components which are pertinent to a few objects will in turn appear close to the leaves. In addition to the IRM, the aforementioned model is able to simultaneously learn up to $L + 1$ distinct identities exhibited by each object and their relative proportions. Expressed differently, we are able to learn all the groups that exist in the domain and the degree of membership of each object (mixed membership). Furthermore it can identify commonalities and idiosyncrasies in the dataset as those are reflected in the hierarchical order of identity components. The use of the nonparametric prior can readily accommodate outliers in the sense that if an object exhibits different relational patterns a new path can be generated in the tree which will be occupied by this node. It should finally be noted that if multiple domains are present then a separate tree is constructed for each domain.

Inference in this nonparametric Bayesian model is analytically intractable and for that purpose is performed using MCMC and in particular component-wise Gibbs sampling as explained below. This is the obvious choice in non-parametric models as conditional probabilities (as in Equation (4)) are readily available. It should be noted that intelligent and faster mixing, block Gibbs sampling techniques which recently have appeared in the literature ([14]) or approximate variational methods ([23]) could provide a better alternative but have not been employed in this study.

In order to facilitate the exposition, we assume for the moment that hyper-parameters are fixed. These consist of the a_l associated with the hierarchical CRP, the β_1, β_2 associated with the Beta distribution that appears in the likelihood, or γ associated with the Dirichlet distribution that appears in the proportions $\theta^{(i)}$. The variables that need to be sampled consist of $z_l^{(i)}$ which define the path that each object is associated with in the tree and the level variables l_i (Note that the number of level variables for each node is equal to the number of links that this node participates in). Although they both affect the likelihood in the manner described in Equation (4), the respective priors are independent and Gibbs sampling can be performed independently for each set.

In sampling $z_l^{(i)}$ we have to consider all the possible paths in the tree. This consist of all the existing paths plus any possible new paths that arise by adding new nodes in the levels of the tree, starting from the bottom. If M_l is the number of nodes at each level$l$ in the tree, the total number of paths that need to be considered is $1 + \sum_{l=1}^{L} M_l$ (since at the 0-level there is always a single node). The prior probabilities can be easily calculated using 4 and proceeding from level L upwards.

As for the level variables $\boldsymbol{l_i} = \{l_{i,m}\}_{m=1}^{m_i}$ where m_i is the number of links in which object i participates, we can arrive to simple expressions for the conditional prior probabilities by integrating the Dirichlet distributed, $L + 1$ dimensional, vector of proportions $\boldsymbol{\theta^{(i)}}$. In particular this leads to the following:

$$p(l_{i,m} = l \mid \boldsymbol{l_{i,-m}}) = \frac{m_l^{(i)} + \gamma}{m_i - 1 + (L+1)\gamma} \tag{45}$$

where $m_l^{(i)}$ is the number of level variables of object i assigned to level l (excluding l_i^c) and $m_i = \sum_l m_l^{(i)} + 1$.

4.1 Numerical Examples

In order to address the issue of identifiability as described in section 3 that can arise in mixed-membership models, we again fix all the identity components of an arbitrarily selected object to the top node of the tree. We examine one synthetic and two real-life datasets. In all cases the following hyper-priors were used:

- for β_1, β_2 : independent $Uniform(0, 5)$
- for $\{a_l\}_{l=0}^{L}$: independent $Gamma(1., 0.1)$

In general we report the maximum likelihood configurations found by performing 10 independent runs with $10,000$ MCMC iterations each.

4.1.1 Example 1: Artificial Data

We utilize the dataset summarized in Table 7 which consists of 40 objects. As it can be seen, identity 1 is present in all objects and should therefore appear at the top of any

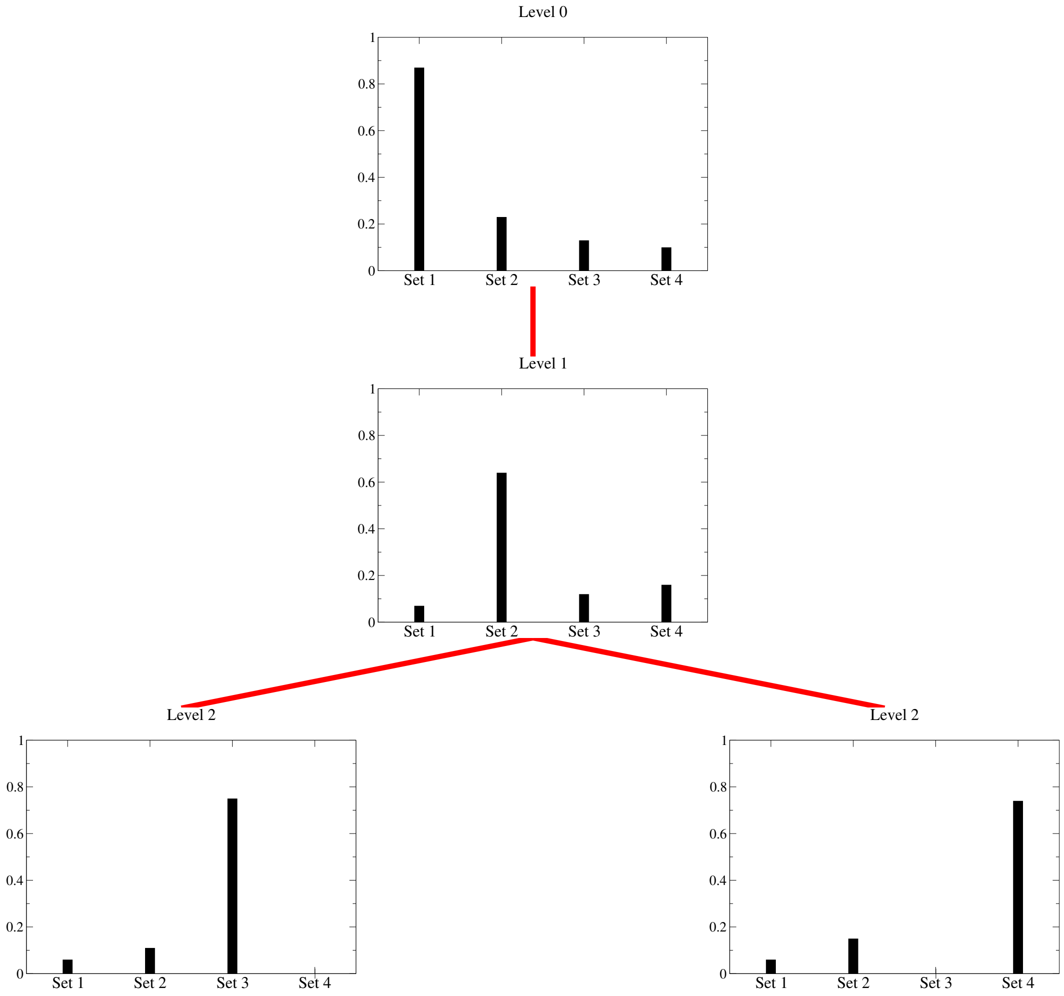


Figure 15: Average probability of membership for each of the four sets of objects described in Table 7

tree. Furthermore identities 3 and 4 are pertinent only to two subsets of the objects (namely objects 21-30 and 31-40) and should therefore appear at the leaves of the tree. We run the hierarchical mixed-membership model with $L = 2$ and Figure 15 depicts the maximum-likelihood configuration which consists of one node/group at level 1 and two at level 2. At each level we plot the average probability of membership for each of the four sets of objects. This is consistent with Table 7. It should also be noted that the model can correctly identify the separation of sets 3 and 4 at the lowest level.

4.1.2 Example 2: Political Books

This dataset consists of 105 political books sold by the online bookseller Amazon.com. the data was assembled by V. Krebs (http://www.orgnet.com/). Links were created based on frequent co-purchasing of books by the same buyers. The books have been assigned 3 labels, namely liberal, conservative and neutral by M. Newman (http://www-personal.umich.edu/ mejn/) based on the reviews and descriptions of the books as posted on Amazon.com.

Figure 16 depicts the maximum likelihood configuration found by the hierarchical mixed membership model for $L = 2$. In particular, each book was assigned to the node for which it had the highest probability of membership. It is worth pointing out that at lower levels the groups identified tend to be cleaner as they tend to consist of books with the same label. Furthermore, the algorithm identifies two basic groups at level 1 of which the left one seems to be associated with liberal books.

4.1.3 Example 3: Reality Mining

This dataset utilizes proximity data for 97 individuals collected over a single week during the academic year 2004-2005 (http://reality.media.mit.edu/). Each person was equipped with a cell phone with a bluetooth device that registered other bluetooth devices that were in close proximity. The individuals participating in this experiment were broadly categorized into 4 groups, namely Sloan business school students, faculty and staff, students of the MIT media lab and others. In the latter category we have added a single object to represent all outsiders.

Figure 17 depicts the maximum likelihood configuration found by the hierarchical mixed membership model for $L = 3$. Each individual was assigned to the node for which it had the highest probability of membership. Although proximity is not necessarily indicative of ones identity, it is worth pointing out that at lower levels the groups identified tend to be cleaner as they tend to consist of individuals with the same label.

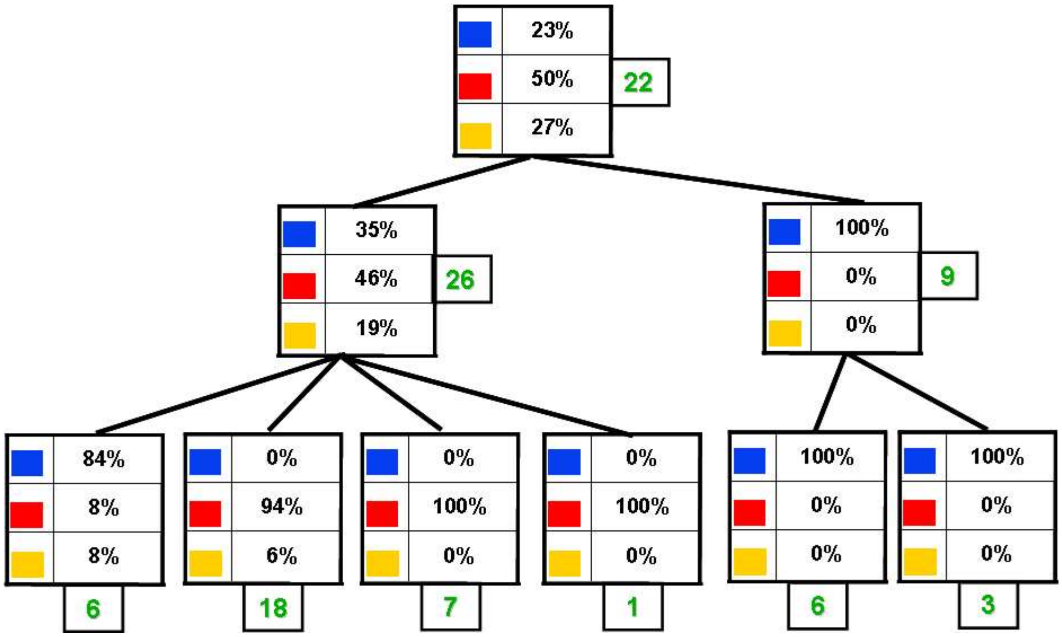


Figure 16: Maximum likelihood configuration for political books dataset. Each book has been assigned to the group for which it had the highest degree of membership. Each box describes the percentage of liberal (blue), conservative (red) and neutral (yellow) books that it consists of. The total number of books in each group is indicated with green

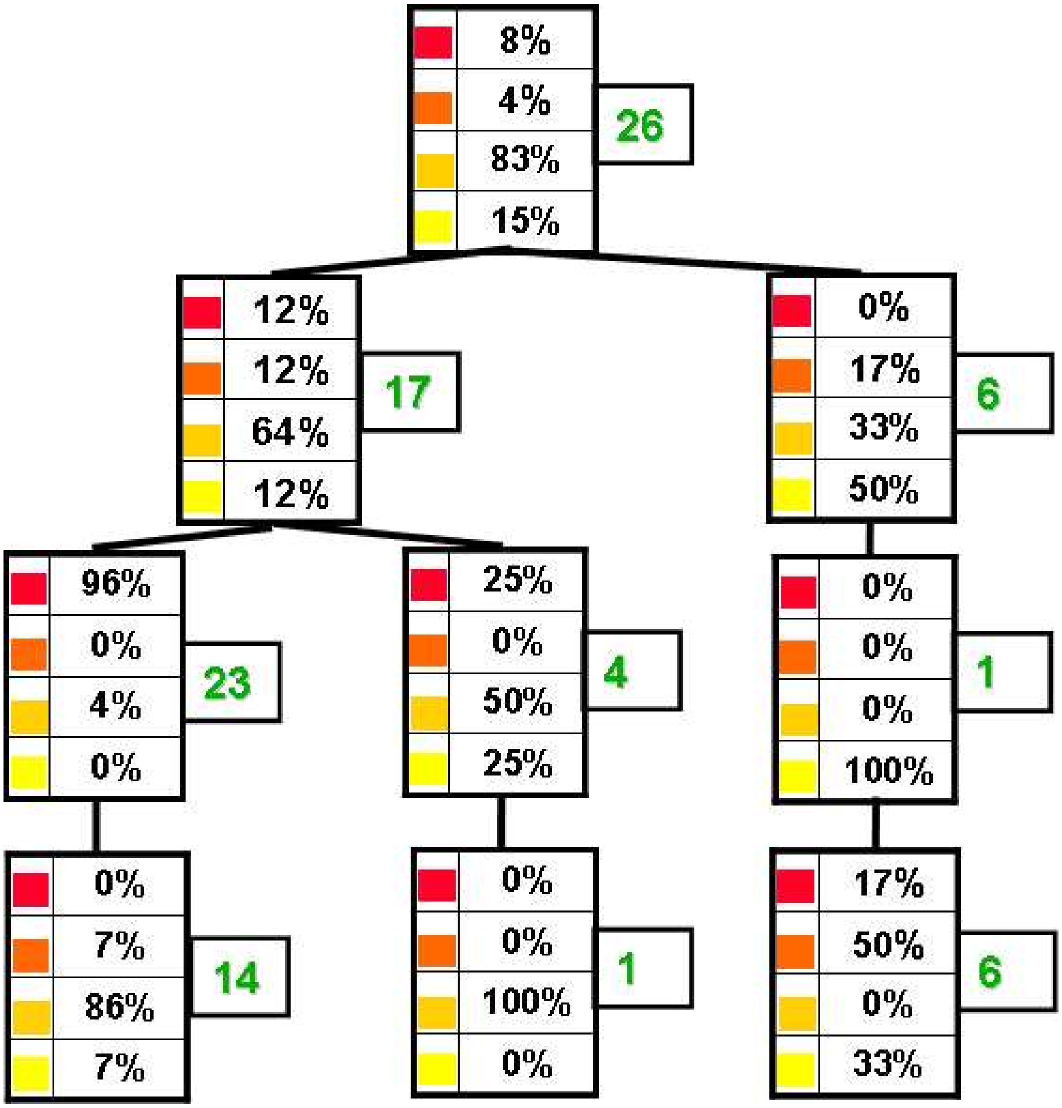


Figure 17: Maximum likelihood configuration for Reality Mining dataset. Each individual has been assigned to the group for which it had the highest degree of membership. Each box describes the percentage of Sloan (red), faculty & staff (orange), students (ocher) and others (yellow) individuals that it consists of. The total number of persons in each group is indicated with green

5 Dynamic Relational Datasets

In this section we discuss the case in which links are observed at various time instants and we wish to identify groups and how their members change in time as well as predict missing links at various time instants. One way to deal with the problem of time within the context of the aforementioned methods is by introducing an additional domain of objects that consists of the time instants and accordingly extend the link data in order to include the time instant that they were recorded. This would produce groups of time instants as well as clusters for the other domains of objects but these would nevertheless be static i.e. represent an averaged picture of the group evolution in time.

Apart from that, one can distinguish two basic modeling techniques that explicitly account for the effect of time. In the first category, models such as the one discussed in previous sections are used but the parameters involved are assumed to evolve based, in general, on some first-order Markov process similar to Hidden Markov Models. Such approaches have been followed in several cases ([28, 25]) with considerable success. The drawback is that the first-order Markov assumption can at times be too restrictive and the results obtained highly dependent on the selected time step. Their advantage is that they can be readily used to make predictions at future time instants.

The second set of techniques, treat the time stamps for each link as observables and introduce additional latent variables that can explain this data ([24]). Their drawback is that they cannot be used for predictions beyond the time range for which data is available. In the following we present one model for each of the aforementioned categories.

Dynamic non-parametric model I

For illustration purposes we consider that each object assumes a single identity at each time instant i.e. $I_{i,t}$ is the identity of object i at time t. This assumption can be readily extended as it was done in the mixed-membership model for the static case in order to account for the possibility that an object can assume different identities from one link to another even at the same time instant. The goal is to learn these identities from the links that have been observed $\boldsymbol{R} = \{R_{i,j}^t\}$ between various objects i and j and at several time instants t. Furthermore we are interested in predicting how these identities will evolve in the future and how they will affect the probabilities of future links. As mention in Section 2, the framework presented can be readily extended to take into account time-varying attributes e.g. $\boldsymbol{x}_{i,t}$ pertinent to the objects i.

Since the number of identities is unknown a priori, we need a prior that can grow with the data.Even though the CRP process gives rise to such a prior, the distribution defined is exchangeable i.e. does not depend on the order that the customers arrive in the restaurant. In time dependent problems though, this order is critical and contains significant information about patterns of evolution in time. For that purpose we propose a dynamic CRP (dCRP) which is defined using conditional probabilities as follows:

$$p(I_{i,t} = z \mid \boldsymbol{I_{-i,t}}) = \begin{cases} \frac{n_z^t}{N^t - 1 + a} & \text{if } n_z^t > 0 \\ \frac{a}{N^t - 1 + a} & \text{if } n_z^t = 0 \end{cases} \tag{46}$$

where $\boldsymbol{I_{-i,t}}$ represents the sitting assignments of all other customers up to time t. The difference with Equation (4) lies in the definition of counts. In particular n_z^t represents the weighted number of customers already seated at table z which is defined as:

$$n_z^t = \sum_l \sum_{k=1}^t w_{t-k} 1(I_{l,j} = z) \tag{47}$$

where w_k represents a sequence of non-increasing, non-negative factors such that $w_0 = 1$ that essentially encapsulate the effect of the order that the customers arrive in time. Apart from these requirements, there is tremendous flexibility in selecting the form of w_t. A good example is $w_t = e^{-\frac{t}{t_0}}$ where the parameter t_0 can be learned from the data. The summation with respect to l in Equation (47) should not include $I_{i,t}$. It should also be noted that the aforementioned prior is exchangeable with respect to the customers that arrive at the same time instant. Furthermore it allows for new states/identities to be created with a probability proportional to a. Finally N^t is used for normalization purposes and is defined as $N^t = \sum_z n_z^t = \sum_{k=1}^t w_{t-k}$.

In order to define the likelihood of each link $R_{i,j}^t$ we postulate that this depends exclusively on the identities of the participating objects at time t i.e. $I_{i,t}$ and $I_{j,t}$. Hence:

$$p(\boldsymbol{R} \mid \boldsymbol{I}) = \prod_t \prod_{i,j} p(R_{i,j}^t \mid I_{i,t}, I_{j,t}) \tag{48}$$

As in the IRM, for binary links $R_{i,j}^t$ the individual likelihoods can be modeled with a Bernoulli distribution with hyper-parameter $\eta I_{i,t}, I_{j,t}, t$ which expresses the probability of a link between a pair of groups/identities at a particular time t. As before, a beta prior $Beta(\beta_{1,t}, \beta_{2,t})$ can be used for η which similarly to Equation (2) would lead to a likelihood:

$$p(\boldsymbol{R} \mid \boldsymbol{I}) = \prod_t \prod_{I,J} \frac{beta(m_{0,t}(I,J) + \beta_{1,t}, m_{1,t}(I,J) + \beta_{2,t})}{beta(\beta_{1,t}, \beta_{2,t})} \tag{49}$$

where $m_{0,t}(I,J)$, $m_{1,t}(I,J)$ are the counts of 0 and 1 links respectively between each pair of identities $I,\ J$ at time t. A first-order Markov process call also be defined for the evolution of the β parameters in time such as $\log \beta_{j,t} \mid \log \beta_{j,t-1} \sim N(\log \beta_{j,t-1}, \sigma), \quad j = 1, 2$.

Dynamic non-parametric model II

Following the previous formulations we introduce a second set of latent variables $\tau_{i,m}$ which represent the activation time i.e. the time instant that object i was activated in some sense in order to participate in the link m. If $\{T_{i,j}^m\}$ denote the time instants which complement the observable links $\{R_{i,j}^m\}$, then the likelihood function can be expressed as follows:

$$p(\{R_{i,j}^m\}, \{T_{i,j}^m\} \mid \boldsymbol{t}, \boldsymbol{I}, \boldsymbol{\eta}) = \prod_m p(R_{i,j}^m \mid \eta(I_{i,m}, I_{j,m}, \tau_{i,m}, \tau_{j,m}) p(\{T_{i,j}^m \mid \tau_{i,m}, \tau_{j,m}) \tag{50}$$

where $I_{i,m}, I_{j,m}$ are the identities of the participating objects for this particular link, $\eta(I_1, I_2, \tau_1, \tau_2)$ expresses the probability of a link between groups $I_1,\ I_2$ when they are activated at times τ_1 , τ_2 respectively. As for the likelihood for $T^m_{i,j}$ we can adopt a form that depends on both activation times such as:

$$p\{T^m_{i,j} = T \mid \tau_{i,m}, \tau_{j,m}) \propto \exp\left\{-\lambda\frac{T - (\tau_{i,m} + \tau_{j,m})/2}{2}\right\} \tag{51}$$

or one of the two depending on the nature of the link.

In order to adopt a nonparametric prior for $\boldsymbol{I}$ we make use of the Chinese Restaurant Franchise model but now assume that we have a different restaurant for each object and for each activation time (the latter can be considered discrete without loss of generality). Essentially this requires defining the prior $p(\boldsymbol{I}, \boldsymbol{\tau})$ in two steps as $p(\boldsymbol{I}, \boldsymbol{\tau}) = p(\boldsymbol{I} \mid \boldsymbol{\tau})p(\boldsymbol{\tau})$ which implies first sampling the activation times $\boldsymbol{\tau}$ (from a finite Dirichlet for example) and given those, use the CRF to sample the group assignment variables $\boldsymbol{I}$.

6 Conclusions

Bayesian, latent variable models provide a valuable tool for unsupervised learning of relational datasets in such tasks as group discovery and link prediction. Their descriptive ability is significantly increased by using nonparametric priors which allows for the number of groups to be learned automatically from the data. The IRM, which serves as the basis of this exposition, is hampered by the assumption that each object is assumed to belong to a single group. For that purpose we introduced two mixed-membership models which can account for the fact that each object can belong simultaneously to several groups. These models are based on a hierarchical version of the CRP (the Chinese Restaurant Franchise) and a novel nonparametric prior on trees. Inference in the context of MCMC schemes remains a challenge particularly when a large number of objects is present. For that purpose we explore the possibility of employing approximate, variational inference techniques and present a novel implementation for CRFs. Finally we touch upon the subject of dynamic datasets and discuss two promising implementations that are able to identify the evolution of groups in time.

References

[1] E Airoldi, D Blei, S. Fienberg, and EP Xing. Latent mixed-membership allocation models of relational and multivariate attribute data. In *Valencia & ISBA Joint World Meeting on Bayesian Statistics*, 2006.

[2] E Airoldi, D.E Blei, EP Xing, and S. Fienberg. A latent mixed membership model for relational data. In *Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005*, 2005.

[3] CE Antoniak. Mixtures of dirichlet processes with applications to nonparametric bayesian problems. *Annals of Statistics*, 2:1152–1174, 1974.

[4] D Blei, T Griffiths, M Jordan, and J Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS 2003*, 2003.

[5] D.M. Blei and M.I. Jordan. Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144, 2005.

[6] C.A. Bush and SN MacEachern. A semiparametric bayesian model for randomised block designs. *Biometrika*, 83(2):75–85, 1996.

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

[8] TS Ferguson. A bayesian analysis of some nonparametric models. *Annals of Statistics*, 1:209–230, 1973.

[9] Z. Ghahramani and M.J. Beal. Variational inference for bayesian mixtures of factor analysers. In *NIPS*, volume 12, 2000.

[10] M Girvan and MEJ Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99:8271–8276, 2002.

[11] MS Handcock, Raftery AE, and JM Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society A*, 170(2):1–22, 2007.

[12] P Hoff. Multiplicative latent factor models for description and prediction in social networks. January 17, 2006.

[13] P. Hoff, AE Raftery, and MS Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090, 2002.

[14] S. Jain and RM Neal. A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

[15] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI 2006*, 2006.

[16] K Kurihara, M. Welling, and YW Teh. Collapsed variational dirichlet process mixture models. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007)*, 2007.

[17] R Neal. Bayesian mixture modeling by monte carlo simulation. Technical report, Technical Report CRG-TR-91-2, Dept. of Computer Science, University of Toronto, 1991.

[18] RM Neal. Markov chain sampling methods for Dirichlet process mixture models. Technical Report No. 9815, Dept. of Statistics, University of Toronto, 1998.

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

[20] J Pitman. Exchangeable and partially exchangeable random parti- tions. *Probab. Th. Rel. Fields*, 102:145–158, 1995.

[21] S. Sampson. *Crisis in a cloister*. PhD thesis, Cornell University, 1969.

[22] Y Teh, M Jordan, M Beal, and D Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.

[23] MJ Wainwright and MI Jordan. Graphical models, exponential families, and variational inference. Technical report, Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.

[24] X Wang and A McCallum. Topics over time: A non-markov continuous-time model for topical trends. In *ACM SIGKDD-2005*, 2005.

[25] X Wei, J Sun, and X Wang. Dynamic mixture model for multiple time series. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[26] M. West. Hyperparameter estimation in dirichlet process mixture models. Technical report, technical report 92-A03, Duke University, ISDS, 1992.

[27] M West, P. Müller, and M.D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In AMF Smith and PR Freeman, editors, *Aspects of Uncertainty: a tribute to D.V. Lindley*, pages 363–386. London: John Wiley & Sons, 1994.

[28] EP Xing. Dynamic nonparametric bayesian models and the birth-death process. Technical report, CMU-CALD Technical Report 05-114, 2005.

[29] WW Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, pages 452–473, 1977.