UCRL-TR-235754

Finding Mixed-Memberships in Social Networks

Phaedon-Stelios Koutsourelakis*
Cornell University
369 Hollister Hall, Ithaca NY 14853
pk285@cornell.edu

Tina Eliassi-Rad
Lawrence Livermore National Laboratory
P.O. Box 808, L-560, Livermore, CA 94551
eliassi@llnl.gov

Abstract

This paper addresses the problem of unsupervised group discovery in social networks. We adopt a nonparametric Bayesian framework that extends previous models to networks where the interacting objects can simultaneously belong to several groups (i.e., *mixed membership*). For this purpose, a hierarchical nonparametric prior is utilized and inference is performed using Gibbs sampling. The resulting mixed-membership model combines the usual advantages of nonparametric models, such as inference of the total number of groups from the data, and provides a more flexible modeling environment by quantifying the degrees of membership to the various groups. Such models are useful for social information processing because they can capture a user's multiple interests and hobbies.

Introduction

Given a social network, a common task is to discover group structure within the network. Generally speaking, the approaches to group discovery (a.k.a., community finding) break down along three classes: graph-theoretic (Newman 2004; Palla, Barabasi, & Vicsek 2007), compression-based (Chakrabarti & Faloutsos 2006), and probabilistic (Kemp *et al.* 2006; Airoldi *et al.* 2006). In this paper, we present a probabilistic approach that (1) is nonparametric and (2) accounts for different mixtures of group memberships for each object, over all possible groups that are present in the whole network. These two characteristics make our model an excellent fit for social information processing.

A fundamental issue in all group discovery problems is that the actual number of groups is unknown *a priori*. In most cases, this is addressed by running the same model several times with a different cardinality each time and selecting the one that provides the best fit to the data (e.g., based on a Bayes factor criterion in the case of Bayesian techniques). Obviously, it would be preferable to develop techniques that are able to infer the number of groups from the data and simultaneously examine hypotheses of different cardinalities. So, we adopt a nonparametric Bayesian framework, which provides a very flexible modeling environment, where the size of the model (i.e., the number of groups) can adapt to the available data and readily accommodate outliers.

In this respect, our formulation is similar to the Infinite Relational Model (IRM) (Kemp *et al.* 2006). However, our *infinite mixed membership model* (IMMM) is able to capture the possibility that objects belong to several groups and quantify their relative degrees of membership. For this 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. We attempt to discover those components and estimate their proportions. Moreover, 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.

Section 2 reviews IRM, which serves as the basis for further developments. Sections 3 and 4 describe IMMM and several experimental studies, respectively. We conclude the paper in Section 5. It should also be noted that even though the majority of this paper's presentation is restricted to objects of a single type or domain (e.g., *people*) and pairwise binary links/connections of a single type (e.g., *is friend of*), the proposed framework can be readily extended to links of various types between several domains. Furthermore, our mixed-membership model can also be used to predict unobserved/missing links among objects. For brevity, we have omitted a discussion of this application.

Preliminaries

Consider a social network which contains observations about links between objects of various types (e.g., people, organizations, movies, etc). These links can be of various types and take binary, categorical, or real values. Without loss of generality, we will restrict the presentation to pairwise binary links $R_{i,j}$ in a single domain (i.e. *person i is friend of person j*) and follow the formalism introduced in IRM (Kemp *et al.* 2006). We present examples with two domains in subsequent sections. The basic goal is to group the objects based on the observables, i.e., the links. A generative model can be defined which postulates that the likelihood of any link between a pair of objects i and j depends exclusively on their respective identities I_i and I_j. In this respect, IRM is identical to the stochastic block-model (Nowicki & Snijders 2001) which

*Work performed while at Lawrence Livermore National Laboratory.

Copyright © 2007, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

is based, however, on a fixed, *a priori* determined number of groups and shares a lot of common characteristics with other latent variable models (Handcock, AE, & Tantrum 2007; Hoff, Raftery, & Handcock 2002). Formally, this leads to the following decomposition of the likelihood function:

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

The product above is over all pairs of objects 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 (Kemp *et al.* 2006). In fact, the η's can be readily integrated out which leads to a simpler expression for the likelihood $p(\boldsymbol{R} \mid identities\ \boldsymbol{I})$ that depends only on the counts $m_0(I, J)$, $m_1(I, J)$ of 0 and 1 links between each pair of identities $I,\ J$, respectively:

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

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:

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

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 non-parametric prior for I_i's is adopted. This is achieved by a distribution over the space of partitions induced by a Chinese Restaurant Process (CRP) (Antoniak 1974; Ferguson 1973). 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 $K \to \infty$ (Neal 1991). A fundamental characteristic 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 objects N and $1, 2, \ldots, N-1$, respectively and n_t is the number of objects already assigned to group t. Hence the Nth 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 (West 1992). Posterior inference with respect to the latent variables I_i can also be performed using Gibbs sampling (Escobar & West 1995; Neal 1998). This simply makes use of the prior conditionals (see Equation (4)) and the likelihood function (see Equation (2)).

The IRM is a flexible and lightweight model for group discovery. However, its most significant deficiency is that each object can belong to only a single identity (or group) and all the links that 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 friendship and the objects are people, then a person might be friends with other people because they belong to the same social group, or work for the same company, etc. This deficiency is particularly noticeable if several link types are simultaneously considered such as is-friend-of, is-colleague-of, and is-acquaintance-of, where depending on the type, each person exhibits different identities with varying degrees. The next section addresses this issue in detail.

Infinite 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 (Airoldi *et al.* 2005; 2006). Posed differently, an object can establish links as a member of multiple groups. This aspect is particularly important in social information processing where social networks can be used for detection of anomalous behaviors/identities. 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 the mixed-membership 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 $R^m_{i,j}$ 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.) Similar to the IRM (see Equation (1)), we assume that the likelihood can be decomposed as:

$$p(\boldsymbol{R} \mid\ \boldsymbol{I}) = \prod_m p(R^m_{i,j} \mid I_{i,m}, I_{j,m}) \tag{5}$$

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)

(Teh *et al.* 2006). 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. First, 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:

$$p(t_{i,m} = t \mid \boldsymbol{t_{i,-m}}) = \begin{cases} \frac{n_{i,t}}{m_i - 1 + a_i} & \text{if } n_{i,t} > 0 \\ \frac{a_i}{m_i - 1 + a_i} & \text{if } n_{i,t} = 0 \end{cases} \quad (6)$$

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):

$$p(d_{i,t} = d \mid \boldsymbol{d_{-(i,t)}}) = \begin{cases} \frac{s_d}{M - 1 + a_0} & \text{if } s_d > 0 \\ \frac{a_0}{M - 1 + a_0} & \text{if } s_d = 0 \end{cases} \quad (7)$$

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_0 accounts for the groups/identities associated with all the objects. The model is summarized below:

$$\begin{aligned} CRP_0 \mid a_0 &\sim CRP(a_0) \quad (8) \\ I_{i,m} \mid a_i &\sim CRP(a_i, CRP_0) \\ \eta(I_1, I_2) \mid \beta_1, \beta_2 &\sim Beta(\beta_1, \beta_2) \\ R^m_{i,j} \mid I_{i,m}, I_{j,m}, \boldsymbol{\eta} &\sim Bernoulli\,(\eta(I_{i,m}, I_{j,m})) \end{aligned}$$

Equations (6) and (7) readily imply how Gibbs sampling can be performed for posterior inference with respect to the latent variables $I_{i,m}$. The latter is not directly sampled but instead we sample the auxiliary variables $t_{i,m}$ and $d_{i,t}$. Further details are contained in (Teh *et al.* 2006). 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 for each object and discover 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.

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:

$$p_a = \sum_{t=1}^{T} p(t_{i,m_{i+1}} = t)\, p(d_{i,t} = k) \quad (9)$$

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 (6).

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:

$$p_b = p(t_{i,m_{i+1}} = T_i + 1)\, p(d_{i,T_i+1} = k) \quad (10)$$

The first term above is given by Equation (6) and the second from Equation (7) as it depends on the number of tables already assigned to dish k.

Object Set	Identity 1	Identity 2	Identity 3
Set_1 (Objects 1-10)	1.0	0.0	0.0
Set_2 (Objects 11-20)	0.2	0.8	0.0
Set_3 (Objects 21-30)	0.1	0.1	0.8
Set_4 (Objects 31-40)	0.1	0.4	0.5

Table 1: Degree of membership matrix for synthetic data

Experimental Evaluation

This section describes our experiments on a variety of synthetic and real data sets. Degrees of group membership were calculated from the maximum likelihood configuration, which was selected from 10 independent runs with $20K$ MCMC iterations each. A standard version of simulated annealing was used in order to avoid local modes with an initial temperature of 100 and a reduction factor of 0.99. In all experiments, the following hyper-priors were used: a) for β_1, β_2 : independent $Poisson(0.1)$, b) for a_0 and a_i's: independent $Gamma(1.0, 1.0)$.

Synthetic Data Set

An artificial data set consisting of 40 objects and 3 identities (or groups) was constructed. The objects were divided into four sets (Set_1 through Set_4) each consisting of 10 objects. The degree of membership of each set to the 3 groups can be seen in Table 1.

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.

The degrees of membership of the four sets of objects for the maximum likelihood configuration are reported in Table 2. IMMM is able to quantify the mixed membership

Object Set	Identity 1	Identity 2	Identity 3
Set_1 (Objects 1-10)	0.90	0.06	0.04
Set_2 (Objects 11-20)	0.14	0.82	0.05
Set_3 (Objects 21-30)	0.10	0.12	0.78
Set_4 (Objects 31-40)	0.16	0.30	0.53

Table 2: Degree of membership matrix for IMMM's maximum likelihood configuration

effect with good accuracy. It should also be noted that the maximum likelihood configuration from the IRM consists of 7 groups (and not 3 groups), to which the respective objects are assumed to belong exclusively (i.e. 100% membership).

In order to more accurately compare the two models, we calculated the similarity matrix $\boldsymbol{P_0}$ that expresses the probability that any pair of objects belong to the same group. For example, the probability that an object from Set_2 is in the same group with an object from Set_3 is $(0.2 \times 0.1) + (0.8 \times 0.1) + (0.0 \times 0.8) = 0.1$. This was compared with the similarity matrices calculated by IRM ($\boldsymbol{P_{IRM}}$) and IMMM ($\boldsymbol{P_{IMMM}}$) by averaging over the the posterior samples. Figure 1 depicts the absolute values of the deviations, i.e., $\mid P_{IRM}^{i,j} - P_0^{i,j} \mid$ and $\mid P_{mm}^{i,j} - P_0^{i,j} \mid, \forall i, j$. The errors are much smaller for IMMM, particularly within the Set_3 and Set_4 which exhibit the most significant mixed-membership characteristics. In fact, the ratio of the *Frobenius error norm* is $\frac{||\boldsymbol{P_{IMMM}} - \boldsymbol{P_0}||}{||\boldsymbol{P_{IRM}} - \boldsymbol{P_0}||} \approx 0.19$.

Another comparison metric is the *full-sample log-score* (LS_{FS}), which is defined as the average over all observable links $\boldsymbol{R_{i,j}}$ (Chen, Shao, & Ibrahim 2000; Krnjajic, Kottas, & Draper 2006). This metric is especially useful in real-world data sets where the ground-truth is unknown. It is formally defined as:

$$LS_{FS} = \frac{1}{M} \sum_{R_{i,j}} \log p(R_{i,j} \mid \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{\theta}) \qquad (11)$$

where $\boldsymbol{I}$ represents the latent variables indicating the identities and $\boldsymbol{\theta}$ the hyper-parameters (e.g., for the IRM, $\boldsymbol{\theta} = (\beta_1, \beta_2, a)$). $p(. \mid .)$ expresses the posterior predictive distribution, which can be readily calculated by averaging over the posterior samples. It should be emphasized that the observables $\boldsymbol{R}$ are not used twice but rather we choose to evaluate the posterior predictive distribution at those points. In contrast to the cross-validation log-score LS_{CV} (Chen, Shao, & Ibrahim 2000) (sometimes referred to as *perplexity*) which requires $O(M)$ independent MCMC runs (where M = the total number of observable links), LS_{FS} can be calculated using one MCMC run. Besides for large M, leaving a single observable out does not in general have a significant effect. Furthermore, as it is discussed in detail in (Krnjajic, Kottas, & Draper 2006), LS_{FS} has favorable properties over Bayes factors and is applicable to nonparametric models (in contrast to other metrics such as the *Deviance Information Criterion*). For our synthetic data, the LS_{FS} values for the IRM and the IMMM were -0.53 and -0.51, respectively. A higher LS_{FS} score denotes a more accurate model.

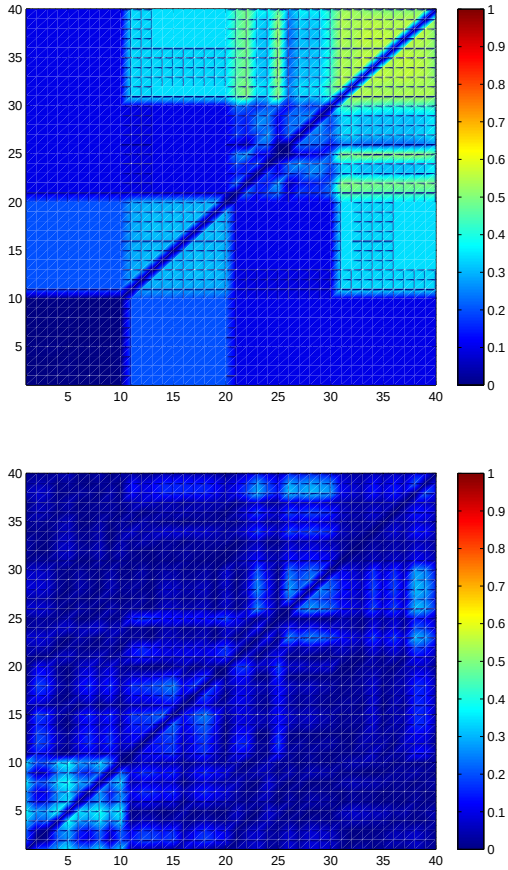

Figure 1: $\mid P_{IRM}^{i,j} - P_0^{i,j} \mid$ (top) and $\mid P_{IMMM}^{i,j} - P_0^{i,j} \mid$ (bottom) for all pairs of objects i, j. $P_M^{i,j}$ is the average over posterior samples that model M gives to i, j belonging to the same group. $M = 0$ represents the true model. The deviation matrix with entries closer to zero (i.e., the bottom one) is modeling the data better.

Real Data Sets

We tested our model on three real social networks (namely, Zachary's Karate Club, U.S. Political Books, and MIT Reality Mining)and one bipartite graph (namely, the animal-feature graph). The latter shows how our model can be extended to data sets with more than one domain. A description of each data set follows next.

The Animal-Feature data set consists 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 (source: http://www.ifs.tuwien.ac.at/∼andi/somlib/). Binary links indicate whether the animal has the particular feature. The animal domain consists of two basic groups: birds and 4-legged mammals. Some of the features are shared between the two basic groups. For example, even though the eagle is a bird, it has certain features with the mammals such as its medium size, being a hunter, and not being able to swim. Similarly, even though the cat belongs

in the mammals, it shares the some features of small-sized birds that cannot run.

Zachary's karate club (Zachary 1977) is a well-studied social network. 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 into two as illustrated in Figure 2 (Girvan & Newman 2002). 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 2 give us an idea about the topology of the network as described in (Girvan & Newman 2002). For example even though individuals 3 and 13 belong to the same group, the former seems closer to group 2 than the latter. For this data, we used a binary version of the friendship links as observables for our mixed-membership model.

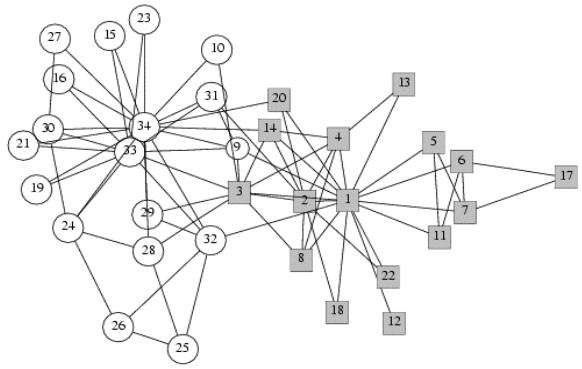

Figure 2: Friendship network for Zacharys karate club (Girvan & Newman 2002)

The U.S. Political data set consists of 105 political books sold by an online bookseller in 2004. The data was assembled by V. Krebs (source: http://www.orgnet.com/). Links were created based on frequent co-purchasing of books by the same buyers. The books have been assigned three labels – namely, liberal, conservative and neutral – by M.E.J. Newman (http://www-personal.umich.edu/∼mejn/) based on reviews and descriptions of the books.

The MIT Reality Mining data set utilizes proximity data for 97 individuals collected over a single week during the academic year 2004-2005 (source: http://reality.media.mit.edu/). Each person was given a cell phone with a blue-tooth device that registered other blue-tooth devices that were in close physical proximity. The individuals participating in this experiment were Sloan Business School students, faculty, staff and students of the MIT Media Lab as well as others. In the latter category we have added a single object to represent all outsiders.

Table 3 depicts the LS_{FS} scores for IRM and IMMM on four real data sets. IMMM's LS_{FS} scores are greater than IRM's in all four cases. (Again higher LS_{FS} scores indicate a better model.)

Discussion For the Animal-Feature data, Figure 3 depicts the comparison of the two rows of the similarity matrix as

Data set	IRM	IMMM
Animal-Feature	-0.30	-0.29
Zachary's Karate Club	-0.15	-0.13
U.S. Political Books	-0.15	-0.09
Reality Mining	-0.27	-0.23

Table 3: Full-sample log-score LS_{FS} on real data sets

calculated by IRM and IMMM. In particular, we calculated the posterior probabilities that the eagle and the cat belong to the same group with any other animal in the domain. Since this is an unsupervised learning task, the values of those probabilities are important in a relative sense rather than absolute. As it can be seen, the IRM assigns high values that the eagle belongs to the same group with the other birds but practically zero for the mammals. This is a direct consequence of the inability of the IRM to capture the possibility that the eagle might simultaneously belong to another group. The results are significantly better with IMMM which predicts higher probabilities that the eagle is in the same group with the rest of the birds, but also has considerable similarities with the mammals reflecting the fact that some of the features are common. For example, eagles are of medium-size like fox, dog, and wolf; they also hunt like some of the cat-family members. Similar results are observed in the case of the cat where the IMMM is shown to have superior performance.

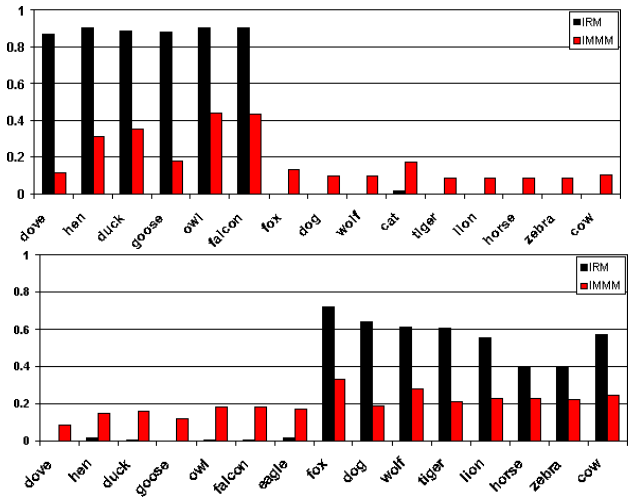

Figure 3: Posterior probabilities that the eagle (top) and the cat (bottom) belong to the same group with any of the other animals in the data set.

For Zachary's Karate Club data, IMMM's maximum likelihood configuration identified 4 groups/identities. IRM's maximum likelihood configuration identified 7 groups. The degrees of membership for each object are depicted in Figure 4. One can observe that identity 1 corresponds to the hard core of group 1, which consists primarily of the instructor (person 1) and to a lesser extent by 2, 4, 5, 6, 11.

Identity 2 represents the "outer layer" (or soft core) of group 1 (i.e. the remaining members of group 1). Similarly, identity 3 denotes the outer layer of group 2. Finally, identity 4 consists of the administrator (34) and a few other closely related individuals such as 33, 24, 25, 26, etc. It is important to note that the model is able to quantify mixed-membership effects. For example, person 3 belongs to group 1 but also has a considerable number of links with the hard (person 33) and soft (persons 10 and 29) cores of group 2; so it exhibits all 4 identities (though primarily those of group 1). Similarly persons 9, 10, 31, 32 (which belong to group 2) exhibit in their identities a component of group 1 due to their relations with other people from that group.

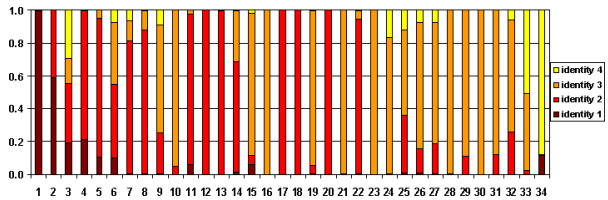

Figure 4: Friendship network for Zacharys karate club. Identities associated with the maximum likelihood configuration from IMMM.

Conclusions

Bayesian latent variable models provide a valuable tool for social network analysis in such tasks as group discovery and link prediction. Their descriptive ability is significantly increased by using nonparametric priors, which allow for the number of groups to be learned automatically from the data. The IRM is hampered by the assumption that each object is assumed to belong to a single group. In this paper, we introduced a mixed-membership model that allows each object to belong simultaneously to several groups. Our model is based on a hierarchical version of the CRP (namely, the Chinese Restaurant Franchise) and inference can be readily performed using Gibbs sampling. The proposed model combines all the advantages of the IRM and superior performance in several synthetic and real-world applications. In particular, our model is expressive enough that it can handle an important problem in social information processing, where one account (say on Amazon.com) is used by several people (like a family) with different interests and hobbies. Future work include extensions for hierarchical group structures and inclusion of time.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

Airoldi, E.; Blei, D.; Xing, E.; and Fienberg, S. 2005. A latent mixed membership model for relational data. In *LinkKDD'05: Proc. of the 3rd Int'l Workshop on Link Discovery*, 82–89.

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

Antoniak, C. 1974. Mixtures of Dirichlet processes with applications to nonparametric Bayesian problems. *Annals of Statistics* 2:1152–1174.

Chakrabarti, D., and Faloutsos, C. 2006. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys* 38(1).

Chen, M.-H.; Shao, Q.-M.; and Ibrahim, J. 2000. *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag.

Escobar, M., and West, M. 1995. Bayesian density estimation and inference using mixtures. *JASA* 90:577–588.

Ferguson, T. 1973. A Bayesian analysis of some nonparametric models. *Annals of Statistics* 1:209–230.

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

Handcock, M.; AE, R.; and Tantrum, J. 2007. Model-based clustering for social networks. *J. of the Royal Statistical Society A* 170(2):1–22.

Hoff, P.; Raftery, A.; and Handcock, M. 2002. Latent space approaches to social network analysis. *JASA* 97(460):1090.

Kemp, C.; Tenenbaum, J.; Griffiths, T.; Yamada, T.; and Ueda, N. 2006. Learning systems of concepts with an infinite relational model. In *Proc. of the 21st AAAI*.

Krnjajic, M.; Kottas, A.; and Draper, A. 2006. Parametric and nonparametric Bayesian model specification: A case study involving models for count data. Technical Report AMS-2006-17, Dept. of Applied Mathematics and Statistics, University of California, Santa Cruz.

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

Neal, R. 1998. Markov chain sampling methods for Dirichlet process mixture models. Technical Report 9815, Dept. of Statistics, University of Toronto.

Newman, M. E. J. 2004. Detecting community structure in networks. *European Physical Journal B* 38:321–330.

Nowicki, K., and Snijders, T. 2001. Estimation and prediction for stochastic blockstructures. *JASA* 96:1077–1087.

Palla, G.; Barabasi, A.-L.; and Vicsek, T. 2007. Quantifying social group evolution. *Nature* 446.

Teh, Y.; Jordan, M.; Beal, M.; and Blei, D. 2006. Hierarchical Dirichlet processes. *JASA* 101(476):1566–1581.

West, M. 1992. Hyperparameter estimation in Dirichlet process mixture models. Technical Report 92-A03, ISDS, Duke University.

Zachary, W. 1977. An information flow model for conflict and fission in small groups. *J. of Anthropological Research* 452–473.