Entropic Regularization of Mixed-membership Network Models using Pseudo-observations

Ramnath Balasubramanyan RBALASUB@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA USA

William W. Cohen WCOHEN@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA USA

Abstract

Mixed-membership network models permit a node in a graph to take on different latent roles in different interactions. However, while mixed-membership block models often do out-perform classical network models, the actual degree of mixed-membership in many graphs is small, with nodes usually taking on only a handful of many possible roles. We thus present a novel *slightly mixed membership* stochastic block model, in which the degree of mixed-membership can be controlled. This model is based on a novel regularization method, where the generative model is extended to include variables that measure aggregate statistics (e.g., the entropy of the distribution of latent roles assigned to nodes), as well as "noisy copies" of these aggregates. We then *pseudo-observe* a desired value for the noisy copies, which has the effect of penalizing models whose aggregates differ greatly from the desired value. Here we demonstrate two applications of this technique: one which encourages slightly-mixed membership, and one which encourages balanced clusters. Experiments with several networks from different domains show that the new models improve performance, as measured by link perplexity and cluster recovery.

1. Introduction

Modeling relations between pairs of objects by representing them as graphs with nodes corresponding to objects is a frequently encountered setting in machine learning and statistics. Common examples of such graphs are web graphs, where the relations indicate hyperlinks between webpages, and social networks, with nodes representing people and edges representing a social link between pairs of people. Models of relational data serve as a foundation for tasks like clustering—i.e., grouping nodes by similarities in interaction patterns, de-noising network representations, and visualizing large complex networks.

The task of studying network structure has been a fertile area of research. In this paper, we mainly focus on stochastic network models (Goldenberg et al., 2010) i.e. generative models that produce random graphs. Stochastic block models (Holland et al., 1983; Snijders & Nowicki, 1997) posit that nodes play a single latent role and the probability of an edge depends only on the latent roles of the nodes. While this approach is simple and elegant, nodes in complex graphs often exhibit multiple latent roles. For instance, in a social network, a person might assume a personal role while creating a link with a relative or a family member and don a more professional role while doing the same with a colleague. Airoldi et al. (2008) introduced the mixed membership stochastic block model (MMSB) that models this phenomenon. This idea of mixed membership has been successfully used previously in non-relational settings in models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) where a word is free to take on different latent roles when it appears multiple times in the corpus. An alternate network model (henceforth the PSK model) presented by (Parkkinen et al., 2009) models sparse networks more efficiently.

In this paper, we introduce a novel technique to regularize mixed membership stochastic blockmodels, demonstrated using the PSK model, which has been shown to outperform MMSB(Parkkinen et al., 2009; Balasubramanyan & Cohen, 2011) on sparse networks, to obtain *slightly mixed membership* stochastic blockmodels. In this approach, we extend the model to include a noisy copy of an aggregate function over latent variables (e.g., the entropy of latent role distributions). By pretending to see a desired value for the copy (e.g.,

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

low entropy of latent role memberships) the model is coaxed to push the variables that participate in the aggregate functions to values that make the pseudo-observed variables likely. This form of regularization therefore permits us to impose biases that cannot be obtained by simply modifying the parameters of prior distributions. By designing the right form for the aggregate functions, restrictions can be imposed on distributions that are not explicity sampled. Moreover independence assumptions between multiple draws from a prior can be potentially overcome by using aggregate functions that span the multiple draws.

Technically, this regularization method can be easily added to any generative model, as it requires only adding new variables and adding observed values of these variables. In our application, the generative model posits that the pseudo-observed variables are distributed as Gaussians, which are parameterized by the aggregate functions of latent variable (such as entropies of the latent role distributions of nodes, or the entropy of the cluster volume distribution). This approach has the advantage of keeping posterior inference simple; we need only extend the Gibbs sampler used for the standard network model by adding a few additional terms, which do not increase the computational order of complexity of inference. We believe that this approach is general and can be extended to model other forms of useful regularizations.

An alternate approach to control the latent role distributions of nodes is to impose a suitable prior on the distributions. However, this approach has the shortcoming of often requiring non-conjugate priors, which make MCMC sampling complicated, computationally expensive, and slower to converge. The pseudo-observation approach also allows the modeler to express easily preferences on distributions that are not explicity sampled—such as the latent role distributions of nodes.

We consider two applications of pseudo-observations. First, we constrain the latent role membership distributions of nodes by penalizing high entropy. By varying the noise model associated with the copy process for the aggregate variables, one can obtain any desired degree of mixed membership, ranging from a fully mixed membership block model (such as the PSK or the MMSB models) to a classical non-mixed network model.

The second application of pseudo-observed variables is motivated by spectral clustering (Shi & Malik, 2000; Luxburg, 2007), a widely used class of techniques for clustering nodes in a graph, and in particular by the Normalized Cut technique, which strives to produce clusters that are balanced in terms of the cluster *volumes*. (Here *volume* is the sum of degrees of nodes belonging to the cluster). In this spirit, we propose a second regularization term for mixed membership stochastic models that imposes a preference on balanced volumes.

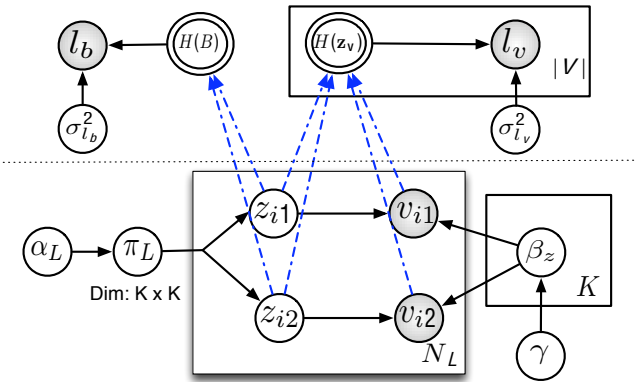


Figure 1. The sparse network model with role entropy and volume entropy regularization.

Results of experiments show that adding either of these penalty terms, or their combination, is beneficial, as measured by the average accuracy in recovering cluster labels with known cluster labels.

The rest of the paper is organized as follows. The mixed membership network model used in the paper is presented in Section 2. Details of the two types of regularization proposed are presented in Sections 3 and 4. Section 5 shows experimental results, followed by a discussion of related work in Section 6 which is followed by the conclusion.

2. Sparse Network Model

The sparse network model(the PSK model) introduced in (Parkkinen et al., 2009) allows nodes to take on different latent roles in different interactions like the MMSB model. As in LDA, clusters are modeled as multinomial distributions over nodes. Recent literature (Parkkinen et al., 2009; Balasubramanyan & Cohen, 2011) suggests that the model outperforms MMSB when modeling sparse networks.

Figure 1 shows the plate diagram for the regularized version of the PSK model that generates a graph representing links between nodes with an underlying block structure. The top part of the figure above the dotted line shows variables related to the regularization introduced in later sections and can be ignored for now. Clusters in this model are represented as distributions over nodes. Nodes participating in an edge are generated from cluster specific node distributions conditioned on the cluster pairs sampled for the edge. Cluster pairs for edges(links) are drawn from a multi-

nomial defined over pairs of cluster labels. Each node in the set of nodes V in the graph therefore has mixed memberships in clusters. The generative process to obtain links in a graph with K clusters is as follows.

1. Generate cluster distributions:

For each cluster $k \in 1, \ldots, K$, sample $\beta_k \sim \text{Dirichlet}(\gamma)$, the cluster specific multinomial distribution over nodes.

2. Generate edges between nodes:

(a) Sample $\pi_L \sim \text{Dirichlet}(\alpha_L)$ where π_L denotes the multinomial distribution over cluster pair labels.

(b) For every link $v_{i1} \rightarrow v_{i2}$, $i \in \{1 \cdots N_L\}$, where $v_{i1}, v_{i2} \in V$:

(i) Sample a cluster pair $\langle z_{i1}, z_{i2} \rangle \sim \text{Multinomial}(\pi_L)$

(ii) Sample $v_{i1} \sim \text{Multinomial}(\beta_{z_{i1}})$

(iii) Sample $v_{i2} \sim \text{Multinomial}(\beta_{z_{i2}})$

In contrast to MMSB, this model only generates realized links that are observed, which provides better fits for sparse graphs.

Given the hyperparameters α_L and γ, the joint distribution over the links, the cluster pair distribution, the cluster node distributions and cluster assignments for edges is given by

$$\mathcal{L} = p(\pi_L, \boldsymbol{\beta}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma)$$
$$\propto \left[\prod_{z=1}^{K} \text{Dir}(\beta_z|\gamma)\right] \text{Dir}(\pi_L|\alpha_L) \prod_{i=1}^{N_L} \pi_L^{\langle z_{i1}, z_{i2} \rangle} \beta_{z_{i1}}^{v_{i1}} \beta_{z_{i2}}^{v_{i2}}$$

Since exact inference in the PSK model is intractable, we use a collapsed Gibbs sampler to perform approximate inference. A cluster pair for every link conditional on cluster pair assignments to all other links after collapsing π_L and $\boldsymbol{\beta}$, is sampled using the expression:

$$p(z_i = \langle k_1, k_2 \rangle | \langle v_{i1}, v_{i2} \rangle, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle^{\neg i}, \alpha_L, \gamma)$$
$$= \frac{p(z_i = \langle k_1, k_2 \rangle, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma)}{p(\langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma)}$$
$$\propto \left(n^{L\neg i}_{\langle k_1, k_2 \rangle} + \alpha_L\right) \frac{\left(n^{\neg i}_{k_1 v_{i1}} + \gamma\right)\left(n^{\neg i}_{k_2 v_{i2}} + \gamma\right)}{\left(\sum_v n^{\neg i}_{k_1 v} + |V|\gamma\right)\left(\sum_v n^{\neg i}_{k_2 v} + |V|\gamma\right)} \quad (1)$$

The n's are counts of observations in the training set, where n_{kv} is the number of times a node v is observed under cluster k and $n^L_{\langle k_1, k_2 \rangle}$ is the number of edges assigned to cluster pair $\langle k_1, k_2 \rangle$. Counts with superscript $\neg i$ indicate that edge i is removed from the counts.

The cluster multinomial parameters and the cluster pair distributions of links are recovered using their point estimates after inference using the counts of observations - $\beta^v_k = \frac{n_{kv}+\gamma}{\sum_{v'} n_{kv'}+|V|\gamma}$ and $\pi_L^{\langle k_1, k_2 \rangle} = \frac{n^L_{\langle k_1, k_2 \rangle}+\alpha_L}{\sum_{k'_1, k'_2} n^L_{\langle k'_1, k'_2 \rangle}+K^2\alpha_L}$. A de-noised form of the entity-entity link matrix can also be recovered from the estimated parameters of the model. Let B be a matrix of dimensions $K \times |V|$ where row $k = \beta_k$, $k \in \{1, \cdots, K\}$. Let Z be a matrix of dimensions $K \times K$ s.t $Z_{p,q} = \pi_L^{\langle p,q \rangle}$. The de-noised matrix M of the strength of association between the nodes in V is given by $M = B^T Z B$.

3. Role Entropy Regularization

Each node $v \in V$ in the PSK model has a set of associated latent roles (z's) it plays when participating in edges. For every node v, we define a distribution $\mathbf{z}_v$ of dimension K where

$$z_v^k = \sum_{v_{i1} \rightarrow v_{i2}} \frac{\mathbf{I}(v_{i1} = v)\mathbf{I}(z_{i1=k}) + \mathbf{I}(v_{i2} = v)\mathbf{I}(z_{i2} = k)}{\mathbf{I}(v_{i1} = v) + \mathbf{I}(v_{i2} = v)} \quad (2)$$

where $\mathbf{I}(\cdot)$ takes the value 0 or 1 depending on the condition being true. The expression effectively computes $p(z = k|v)$, the latent role distribution of the node v. Figure 1 shows blue dashed edges from the latent role and edge-node variables to variables that represents the entropy of $\mathbf{z}_v$. Note that there is no distinction made between the occurrences of the node as a source or destination node, i.e. directionality of the edges is ignored while determining the latent role distribution.

It should be noted that $\mathbf{z}_v$ is not explicitly sampled during the generative process and is an aggregate function of the z and v variables that are generated. Since the different z and v values are independent draws conditioned on π_L and $\boldsymbol{\beta}$, any preference on a function that aggregates over different z and v values cannot be imposed by simply adjusting the parameters of the Dirichlet prior α_L.

We now introduce the role entropy regularization term by adding pseudo-observed variables, l_v, one for each node in V, which are noisy copies of $H(\mathbf{z}_v)$, to the generative process as seen in Fig. 1. $H(\mathbf{z}_v)$ is defined as $-\sum_k z_v^k \log_2 z_v^k$ and represents the Shannon entropy of $\mathbf{z}_v$. These pseudo-observed variables are drawn from Gaussians with mean $H(\mathbf{z}_v)$ and variance $\sigma^2_{l_v}$ which is a hyperparameter to the model. The addition of the terms penalizes large entropies in the latent role distributions while retaining the generative nature of the model. The $\sigma^2_{l_v}$ parameter dictates the strictness of the penalty. The imposition of the penalty therefore allows us to overcome the independence assumption between the different z draws for a given node v.

The regularization term is expressed as

$\prod_v l_v, \quad v \in V, l_v \sim \mathcal{N}\left(H(\mathbf{z}_v), \sigma^2_{l_v}\right)$. Therefore

$$
\begin{aligned}
&p\left(\prod_v l_v | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma^2_{l_v}\right) \\
&= \prod_v p(l_v | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma^2_{l_v}) \\
&\propto \prod_v \exp \frac{-\left(l_v - H(\mathbf{z}_v)\right)^2}{2\sigma^2_{l_v}}
\end{aligned}
$$

Since l_v is observed, i.e., its value is known during inference, the inference procedure tends to push the mean of the Gaussians i.e. $H(\mathbf{z}_v)$ close to the l_v values. We therefore set (*pseudo-observe*) l_v to be 0 to coax the inference procedure to return low entropy latent role distributions for nodes. The variance parameter $\sigma^2_{l_v}$ can be used to adjust the tightness of the Gaussian to permit more or less entropy in the label distributions. In the limit, as $\sigma^2_{l_v}$ tends to 0, the model reduces to the stochastic block model since the regularization will require the entropies to be close to 0 implying that the distribution over latent roles has all its mass on one cluster. Similarly, as the variance tends to ∞, the model reduces to a fully unconstrained mixed membership model. [1]

The joint distribution of the model with regularization is given by

$$
\begin{aligned}
\mathcal{L}^m &= p(\pi_L, \boldsymbol{\beta}, \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma, \sigma^2_{l_v}) \\
&= \mathcal{L} \times \prod_v p(l_v | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma^2_{l_v}) \qquad (3)
\end{aligned}
$$

To obtain the conditional distribution required for Gibbs' sampling with the regularization term added, we see that (derivation in the Appendix)

$$
\begin{aligned}
&p(z_i = \langle k_1, k_2 \rangle | \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \alpha_L, \gamma, \sigma^2_{l_v}) \\
&\propto \left(n^{L^{\neg i}}_{\langle k_1, k_2 \rangle} + \alpha_L\right) \frac{\left(n^{\neg i}_{k_1 v_{i1}} + \gamma\right)\left(n^{\neg i}_{k_2 v_{i2}} + \gamma\right)}{\left(\sum_v n^{\neg i}_{k_1 v} + |V|\gamma\right)\left(\sum_v n^{\neg i}_{k_2 v} + |V|\gamma\right)} \\
&\times \exp \frac{-\left(l_{v_{i1}} - H(\mathbf{z}_{v_{i1}})\right)^2}{2\sigma^2_{l_v}} \ \exp \frac{-\left(l_{v_{i2}} - H(\mathbf{z}_{v_{i2}})\right)^2}{2\sigma^2_{l_v}} \qquad (4)
\end{aligned}
$$

where $\mathbf{z}_{v_{i1}}$ and $\mathbf{z}_{v_{i2}}$ use the assignment of $z_{i1} = k_1$ and $z_{i2} = k_2$.

It can be seen from the expression that adding the role entropy regularization is computationally inexpensive since the extra terms introduced in the Gibbs sampling expression only require the entropies of the current edge's source and destinate nodes' latent role distributions to be computed and does not require any computation over nodes and edges that do not participate in the edge being considered.

[1] The pseudo-observed variables l_v can be modeled using a variety of distributions parameterized on $H(\mathbf{z}_v)$. We use Gaussian distributions for this in this paper because of its property of controllable peakiness (by adjusting the variance) around a desired mean and due to its minimal impact on sampling complexity.

4. Cluster Volume Regularization

Cluster balance is an important aspect in clustering. Spectral clustering methods which are relaxed versions of Ratio Cut and Normalized Cut (Shi & Malik, 2000) use different ways to define notions of cluster balance. In the case of Normalized Cut, the clusters are coaxed to have balanced volumes (which is defined as the sum of the degrees of the nodes in the cluster). To impose a similar preference in the PSK model, we propose a regularization scheme that prefers a higher entropy in the volume distribution $\mathbf{B}$, which is defined as follows: $B_k, k \in 1, \ldots K = \sum_{i=1}^{N_L} \frac{\mathbf{I}(z_{i1}=k)+\mathbf{I}(z_{i2}=k)}{2*N_L}$.

The regularization term l_b (seen in Figure 1) is drawn from the Gaussian $\mathcal{N}(H(\mathbf{B}), \sigma^2_{l_b})^{-1}$. It should be noted that the regularization term is the multiplicative inverse of the density.

Since the Gaussian term in the sampling equation below (Equation 5) is raised to the power -1, setting l_b to 0 will cause the sampling procedure to diminish the probability $p(l_b | H(\mathbf{B}), \sigma^2_{l_b})$ by returning a mean for the Gaussian i.e. $H(\mathbf{B})$ that is far from 0, which means that $H(\mathbf{B})$ will tend to be high, implying that $\mathbf{B}$ will tend to be more balanced. $\sigma^2_{l_b}$ like $\sigma^2_{l_v}$ in the case of role entropy regularization, controls the strictness of this preference. This value can be set to a lower value in cases where the network is believed to have more balanced clusters and can be set higher when bigger variations in the volumes of clusters is expected.

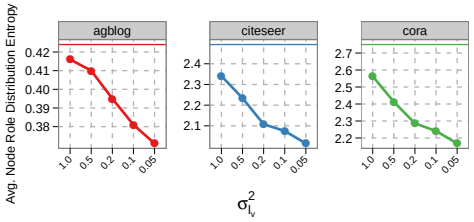
The joint distribution after adding volume and role entropy regularization terms to the PSK model is defined as $L^{bm} = p(\pi_L, \boldsymbol{\beta}, l_b, \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{e_1}, \mathbf{e_2} \rangle | \alpha_L, \gamma, \sigma^2_{l_v}, \sigma^2_{l_b})$. Therefore $\mathcal{L}^{bm} = \mathcal{L}^m \times p(l_b | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \sigma^2_{l_b})$.

Similar to Equation 4, the Gibbs sampling equation for the latent cluster pair of an edge, with $\mathbf{B}$ using the assignment $z_{i1} = k_1$ and $z_{i2} = k_2$ is now defined as

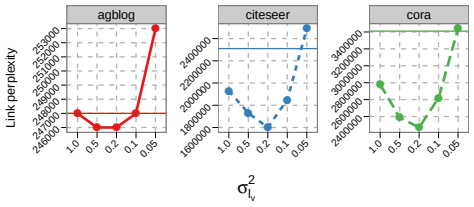
$$
\begin{aligned}
&p(z_i = \langle k_1, k_2 \rangle | l_b, \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \alpha_L, \gamma, \sigma^2_{l_v}) \\
&\propto \left(n^{L^{\neg i}}_{\langle k_1, k_2 \rangle} + \alpha_L\right) \frac{\left(n^{\neg i}_{k_1 v_{i1}} + \gamma\right)\left(n^{\neg i}_{k_2 v_{i2}} + \gamma\right)}{\left(\sum_v n^{\neg i}_{k_1 v} + |V|\gamma\right)\left(\sum_v n^{\neg i}_{k_2 v} + |V|\gamma\right)} \\
&\times \exp \frac{-\left(l_{v_{i1}} - H(\mathbf{z}_{v_{i1}})\right)^2}{2\sigma^2_{l_v}} \ \exp \frac{-\left(l_{v_{i2}} - H(\mathbf{z}_{v_{i2}})\right)^2}{2\sigma^2_{l_v}} \\
&\times \left(\exp \frac{-\left(l_b - H(\mathbf{B})\right)^2}{2\sigma^2_{l_b}}\right)^{-1} \qquad (5)
\end{aligned}
$$

Table 1. Evaluation of regularization in the smaller datasets.

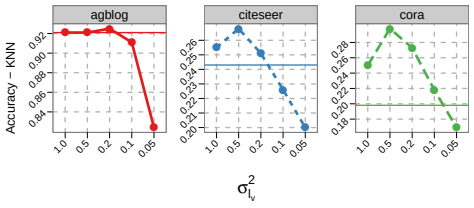
Dataset	Perplexity				Accuracy			
	Regularization				Regularization			
	None	Role	Volume	Both	None	Role	Volume	Both
agblog	2.47e+05	2.47e+0.5	2.33+05	**2.31+05**	0.921	0.925	0.922	**0.947**
citeseer	2.31e+06	1.73+06	1.60e+06	**1.51+06**	0.243	0.268	0.282	**0.291**
cora	3.41e+06	**2.27e+06**	2.52+06	2.62e+06	0.198	**0.268**	0.210	0.230
dolphin	2.10e+03	2.03e+03	2.03e+03	**2.00e+03**	0.871	**0.935**	0.897	0.881
football	1.31e+04	**0.79e+04**	0.96e+04	**0.79e+04**	0.161	**0.833**	0.515	0.560
karate	5.72e+02	**5.35e+02**	5.39e+02	5.54e+02	0.941	**1.00**	0.951	0.961
polbook	4.95e+03	4.91e+03	4.84e+03	**4.77e+03**	0.752	**0.778**	0.774	**0.778**
senatevote	3.50e+03	3.50e+03	**3.44e+03**	3.46e+03	0.969	**0.980**	**0.980**	0.971



(a) Effect of role distribution entropy



(b) Perplexity



(c) Cluster label prediction accuracy

Figure 2. Role entropy - varying the variance hyperparameter. (Horizontal line indicates no-regularization baseline)

5. Experimental Results

5.1. Datasets

We investigate the effects of regularization on a collection of datasets consisting of social networks, citation networks, yeast protein-protein interaction networks and other similar networks that have been studied in the sociology literature.

The first set of graphs (Balasubramanyan et al., 2010) are relatively small and have one known true cluster label for every node, which is used solely for evaluating the accuracy of node clustering. Statistics about the datasets are shown in the first 8 rows of Table 2.

Table 2. Dataset statistics.

Dataset	Nodes	Edges	Clusters	#Labels per-node
ag	1222	33428	2	1
cora	2485	10138	7	1
citeseer	2114	7396	6	1
dolphin	62	318	2	1
football	115	1226	10	1
karate	34	156	2	1
polbooks	105	882	3	1
senate	98	9506	2	1
yeast	844	14780	15	2.5
blogcatalog	10,312	333,983	39	1.4
youtube	1,138,499	2,990,443	47	1.6

Next, we study the Munich Institute for Protein Sequencing (MIPS) database which contains a collection of protein interactions covering protein complex associations in yeast. We use a subset of this collection containing 844 proteins, for which all interactions were hand-curated. The proteins in the dataset are also annotated with functional categories based on the functions that they play. There are 15 top-level functional categories which are treated as known cluster labels. On average, a protein is annotated with 2.5 functional categories.

In addition to the smaller networks described above, we also run experiments on two larger benchmark networks, namely the BlogCatalog and YouTube datasets (Zafarani & Liu, 2009). These larger networks also have known mixed-membership labels for nodes. On average across all nodes, nodes in the BlogCatalog dataset have 1.4 labels per node and nodes in the YouTube dataset have 1.6 labels per node. More statistics about these networks are in Table 2.

Table 3. Predicting cluster labels in mixed-membership datasets.

Model	BlogCatalog			YouTube			Yeast		
	Micro F-1	Macro F-1	Avg KL	Micro F-1	Macro F-1	Avg KL	Micro F-1	Macro F-1	Avg KL
Unregularized	0.131	0.076	2.271	0.154	0.084	0.117	0.435	0.284	1.91
Role entropy	0.153	0.077	2.141	0.165	0.086	0.114	0.485	**0.321**	**1.80**
Volume entropy	0.154	0.080	2.150	0.171	0.089	**0.112**	0.468	0.305	1.83
Both	**0.161**	**0.082**	**2.075**	**0.198**	**0.096**	0.115	**0.523**	0.310	1.83

5.2. Experimental Setup

We evaluate the regularized and unregularized versions of the PSK model using the following metrics. The first metric used is average node entropy defined as $\sum_v H(\mathbf{z}_v)/|V|$. This metric shows the extent to which each node participates in multiple latent roles. The second metric used to evaluate the model is link perplexity which is a function of the likelihood of the edges in the dataset and is defined as

$$2^{-\frac{\sum_{v_{i1}\to v_{i2}} log_2 p(v_{i1}\to v_{i2})}{N_L}}$$

A lower perplexity value indicates an higher likelihood of data and a better fit.

For datasets with only one node per label, we can evaluate the model by checking its accuracy in predicting node labels. Nodes in the PSK model are associated with a distribution over clusters which can be obtained by normalizing β_z^v. Predicted class labels can be assigned to nodes using the 1-NN algorithm using the Jensen-Shannon distance between these cluster distributions as the metric to measure the distance between two nodes.

Performance in the larger networks which have multiple labels per node is measured using micro and macro averaged F-1 measures of retrieving the known cluster labels. The prediction of multiple labels for a node is done in two stages. In the first stage, the Hungarian algorithm (Kuhn, 1955) is used to align true cluster labels to clusters in the model. Next, labels corresponding to elements from posterior role distributions of nodes that are above a threshold are treated as predicted labels.

For every dataset, we run experiments with 1) the baseline PSK model with no regularization, 2) PSK with role entropy regularization, 3) PSK with volume entropy and finally 4) PSK with both the regularization terms introduced in the paper. In all experiments, the number of clusters in the model is set to be the number of known clusters in the dataset. The Gibbs sampler is set to run for 100 iterations and the average of the last 10 samples is taken. Since Gibbs sampling results can vary depending on the random starting point, the accuracy and perplexity values reported are the means of 10 separate runs. The variance values $\sigma_{l_b}^2$ and $\sigma_{l_v}^2$ are set to 0.5 and we place priors which favor diagonal blocks over off-diagonal blocks by using a non-symmetric Dirichlet for α_L.

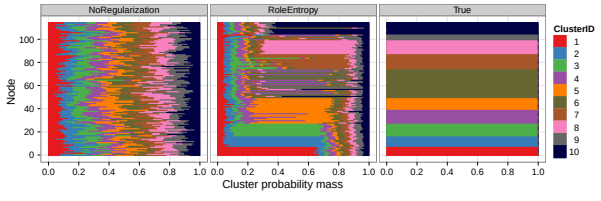


Figure 3. Role entropy demonstration: Inferred Latent Role distributions in the football network.

5.3. Results

First we study the direct impact of role entropy and volume entropy on the smaller networks measured by link perplexity and 1-NN clustering accuracy. It can be seen from Table 1 (bold values indicate the best performing model) that using role and volume entropy consistently decreases link perplexity and increases cluster prediction accuracy when compared to the baseline unregularized model. The improvements for both perplexity and accuracy for all variants of regularization is statistically significant at the 0.05 level using the Wilcoxon paired-sign test. The direct impact of role entropy regularization is illustrated further in Figure 2(a) which plots the average node entropy of 3 sample datasets obtained using different values of $\sigma_{l_v}^2$. It can be seen from the figure that the average node entropy in these datasets decreases as the variance parameter value is increased which shows that tightening the variance leads to models that tend closer to the stochastic block model where the entropy of the latent role distribution is 0. Figure 3 shows the reduction of entropy in latent role distribution more clearly. The figure shows a heatmap of the latent role distributions of each node in the network with nodes on the y-axis. The panel on the far right shows the known true label distributions with solid single colors in each row since nodes in the football dataset have one cluster label per node. The panel on the left shows the latent role distribution after inference using an unregularized

PSK model. The middle panel shows the latent role distribution after inference with a role entropy regularized model. It can be seen clearly that the regularized model returns more peaky distributions with large probability masses residing in certain roles as compared to the unregularized model where the distributions are more equally distributed. The rows where the dominant color in the left and middle panels do not match the color in the right panel indicate cluster assignment errors.

Figures 2(b) and 2(c) show how perplexity and clustering accuracy vary with different values for the variance term $\sigma^2_{l_v}$ on the same 3 sample networks used in Figure 2(a). The perplexity curves show a general U-shaped pattern that dips below the horizontal line representing the perplexity of the unregularized model, indicating a "sweet spot" for the variance value. Similarly in the 3 accuracy plots, the regularized model accuracies rise above the baseline value with increasing variance values and then fall when it is increased further. At very low variance values, the model effectively approximates a single latent role stochastic block model since the nodes are restricted to only one role with high probability. This behavior tends to offer insufficient flexibility in modeling networks that inherently possess some mixed membership characteristics leading to drastic fall-offs in accuracies. These results indicate the although the smaller networks have only one true label per node, the actual structure of the networks does exhibit mixed-membership characteristics.

Next, we evaluate the impact of regularization in the larger multi-labelled networks by checking the ability of the model to recover the known labels of nodes. Because nodes in these networks can have multiple labels, we use the micro and macro averaged F-1 measures to evaluate the clustering rather than accuracy. The models are also evaluated by computing the Kullback-Leiber(KL) divergence between the known true role distributions and the predicted role distributions averaged over all the nodes in the network. Table 3 shows the F-1 measures and KL divergences obtained from the clustering. It can be seen from the table that adding role and volume regularization improves performance in all 3 datasets. These networks have an average of 1.4 to 2.5 true labels per node and adding role entropy forces the model to restrict the number of roles a node can participate in which is a better fit to the true nature of the network. Volume entropy improves performance similarly by penalizing the formulation of trivially small clusters.

6. Related Work

Regularization by entropy has been used previously for semi-supervised learning in (Grandvalet & Bengio, 2005), (Jiao et al., 2006) and (Corduneanu & Jaakkola, 2005) where entropy based regularizers are used to constrain the unknown labels of unlabeled data points. (Celeux & Soromenho, 1996) also use criteria based on entropy to determine the optimal number of clusters in mixture models. The approach presented in this paper uses entropy for a different purpose i.e. to impose preferences on the mixed-membershipness of nodes and the volume balance in clusters. Regularization in models based on Latent Dirichlet Allocation have been previously proposed in works such as (Cai et al., 2008) and (Mei et al., 2008), which use a regularization term in the likelihood expression to remove the independence assumption between documents by placing them on a manifold. (Newman et al., 2011) present a technique that uses structured priors over words as a way to regularize topic models to be more sensitive to known cooccurence patterns. Airoldi et al.(Airoldi et al., 2008) also describe a method to regularize the MMSB model to permit better fits for sparse graphs. The regularization techniques described in this paper however are designed to specifically influence the mixed-membership and balance characteristics of the network which is different from the goal of the regularization in the work described above. It also differs from previous regularization approaches through its use of pseudo-observed variables which allows the model to retain a generative story. This method provides a general alternate way to impose preferences without the use of non-conjugate prior distributions by adding noisy copies of aggregate functions of latent variables to models which can be set to desired values to prefer desirable properties in the latent variable distributions.

7. Conclusion

We presented a general technique to impose preferences in latent variable models using pseudo-observed variables and and used it to regularize stochastic network models to control nodes' ability to take on different roles and to obtain balanced clusters. The regularization scheme permits the use of Gibbs sampling for inference with only an addition of a few terms to the sampling equations of the original PSK model. The technique of using pseudo-observed variables can also be used to impose other soft restrictions on networks such as controlling the incoming and outgoing latent roles separately and also to other stochastic network models such as MMSB. Experiments on real world network data both small and large, show that using slightly mixed-membership models using the regular-

ization introduced provides better fits and consistently improves link perplexity and cluster label prediction.

References

Airoldi, Edoardo M., Blei, David M., Fienberg, Stephen E., and Xing, Eric P. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9: 1981—-2014, September 2008.

Balasubramanyan, Ramnath and Cohen, William W. Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In *SDM*, pp. 450–461. SIAM / Omnipress, 2011. ISBN 978-0-898719-92-5.

Balasubramanyan, Ramnath, Lin, Frank, and Cohen, William W. Node clustering in graphs: An empirical study. In *NIPS Workshop on Networks Across Disciplines in Theory and Applications*, 2010.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Cai, Deng, Mei, Qiaozhu, Han, Jiawei, and Zhai, Chengxiang. Modeling hidden topics on document manifold. In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, pp. 911, New York, New York, USA, October 2008. ACM Press.

Celeux, Gilles and Soromenho, Gilda. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

Corduneanu, Adrian and Jaakkola, Tommi S. Distributed Information Regularization on Graphs. *Advances in Neural Information Processing Systems 17*, .17:297–304, 2005.

Goldenberg, Anna, Zheng, Alice X., Fienberg, Stephen E., and Airoldi, Edoardo M. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2:129–233, February 2010.

Grandvalet, Yves and Bengio, Yoshua. Semi-supervised Learning by Entropy Minimization. *Advances in neural information processing systems*, 17:529–536, 2005.

Holland, Paul W., Laskey, Kathryn B., and Leinhardt, Samuel. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Jiao, Feng, Wang, Shaojun, Lee, Chi-Hoon, Greiner, Russell, and Schuurmans, Dale. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pp. 209–216, Morristown, NJ, USA, July 2006. Association for Computational Linguistics.

Kuhn, Harold W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2 (1-2):83–97, 1955. doi: 10.1002/nav.3800020109.

Luxburg, Ulrike. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, August 2007.

Mei, Qiaozhu, Cai, Deng, Zhang, Duo, and Zhai, ChengXiang. Topic modeling with network regularization. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, pp. 101, New York, New York, USA, April 2008. ACM Press.

Newman, David, Bonilla, Edwin V., and Buntine, Wray L. Improving topic coherence with regularized topic models. In *NIPS*, pp. 496–504, 2011.

Parkkinen, Juuso, Sinkkonen, Janne, Gyenge, Adam, and Kaski, Samuel. A block model suitable for sparse graphs. *The 7th International Workshop on Mining and Learning with Graphs*, 2009.

Shi, Jianbo and Malik, Jitendra. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Snijders, Tom A.B. and Nowicki, Krzysztof. Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1):75–100, January 1997.

Zafarani, Reza and Liu, Huan. Social computing data repository at ASU, 2009.

Appendix

We derive the Gibbs sampling equation for the PSK model regularized with role entropy.

$$
\begin{aligned}
&p(z_i = \langle k_1, k_2\rangle | \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2}\rangle | \alpha_L, \gamma, \sigma^2_{l_v}) \\
&= \frac{p(z_i = \langle k_1, k_2\rangle, \langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2}\rangle | \alpha_L, \gamma)}{p(\langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2}\rangle | \alpha_L, \gamma)} \\
&\times \frac{p(\mathbf{l_v} | \langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, z_i = \langle k_1, k_2\rangle, \langle \mathbf{v_1}, \mathbf{v_2}\rangle, \sigma^2_{l_v})}{\sum_{k'_1}\sum_{k'_2} p(\mathbf{l_v} | \langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, z_i = \langle k'_1, k'_2\rangle, \langle \mathbf{v_1}, \mathbf{v_2}\rangle, \sigma^2_{l_v})}
\end{aligned}
\tag{6}
$$

The first term in the product is the same as the unregularized model and we can replace the term with the expression from Equation 1. In the second term of the product, the denominator is not dependent on $\langle k_1, k_2\rangle$ and can therefore be discarded as it only serves as a normalizing constant.

$$
p(\mathbf{l_v} | \langle \mathbf{z_1}, \mathbf{z_2}\rangle, \langle \mathbf{v_1}, \mathbf{v_2}\rangle, \sigma^2_{l_v}) = \prod_v p(l_v | \langle \mathbf{z_1}, \mathbf{z_2}\rangle, \langle \mathbf{v_1}, \mathbf{v_2}\rangle, \sigma^2_{l_v})
$$

Terms in the product that do not pertain to v_{i1}, v_{i2}, z_{i1} and z_{i2} can be discarded since they are constants over all cluster pair label assignments. Therefore the second term in Equation 6 is only dependent on $\mathbf{z}_{v_{i1}}$ and $\mathbf{z}_{v_{i2}}$. The Gibbs sampling equation can now be expressed as

$$
\begin{aligned}
&\left(n^{L\neg i}_{\langle k_1, k_2\rangle} + \alpha_L\right) \frac{\left(n^{\neg i}_{k_1 v_{i1}} + \gamma\right)\left(n^{\neg i}_{k_2 v_{i2}} + \gamma\right)}{\left(\sum_v n^{\neg i}_{k_1 v} + |V|\gamma\right)\left(\sum_v n^{\neg i}_{k_2 v} + |V|\gamma\right)} \\
&\times \exp\frac{-\left(l_{v_{i1}} - H(\mathbf{z}_{v_{i1}})\right)^2}{2\sigma^2_{l_v}} \ \exp\frac{-\left(l_{v_{i2}} - H(\mathbf{z}_{v_{i2}})\right)^2}{2\sigma^2_{l_v}}
\end{aligned}
\tag{7}
$$

where $\mathbf{z}_{v_{i1}}$ and $\mathbf{z}_{v_{i2}}$ use the assignment of $z_{i1} = k_1$ and $z_{i2} = k_2$.