HSN-PAM: Finding Hierarchical Probabilistic Groups from Large-Scale Networks

Haizheng Zhang
College of Information Science and Technology
Pennsylvania State University
hzhang@ist.psu.edu

Wei Li, Xuerui Wang
Department of Computer Science
University of Massachusetts, Amherst
{weili,xuerui}@cs.umass.edu

C. Lee Giles, Henry C. Foley, John Yen
College of Information Science and Technology
Pennsylvania State University
{giles,hfoley,jyen}@ist.psu.edu

Abstract

Real-world social networks are often hierarchical, reflecting the fact that some communities are composed of a few smaller, sub-communities. This paper describes a hierarchical Bayesian model based scheme, namely HSN-PAM *(Hierarchical Social Network-Pachinko Allocation Model), for discovering probabilistic, hierarchical communities in social networks. This scheme is powered by a previously developed hierarchical Bayesian model. In this scheme, communities are classified into two categories:* super-communities *and* regular-communities*. Two different network encoding approaches are explored to evaluate this scheme on research collaborative networks, including* CiteSeer *and* NanoSCI*. The experimental results demonstrate that* HSN-PAM *is effective for discovering hierarchical community structures in large-scale social networks.*

1 Introduction

Social networks have been studied for decades. In recent years, this line of research has gained even more momentum with the prevalence of online social networking systems, such as *MySpace, LiveJournal, Friendster* and instant messaging systems. Despite the vast number of nodes, the heterogeneity of the user bases, and the variety of interactions among the members, most of these networks exhibit some common properties, such as the small-world property, power-law degree distribution, and community structures. An important task in these emerging networks is community discovery, which is to identify subsets of networks such that connections within each subset are dense and connections among different subsets are relatively sparse. Since large-scale complex networks based applications exist in many disciplines, community discovery is appealing to researchers from a variety of areas such as computer science, biology, social science and so on.

Although a wide array of approaches have been developed over years for finding communities, the current dominant community discovery algorithms tend to define various distance-based measures and cluster networks accordingly. However, such strategies fail to capture the overlap among communities, identify the multiple membership phenomenon, and discover inherent hierarchical communities. In order to address the aforementioned problems, we develop an *HSN-PAM*(Hierarchical Social Network-Pachinko Allocation Model) scheme by applying the Pachinko Allocation Model(*PAM*) [3], a DAG-structured mixture models, to identify and discover probabilistic hierarchical communities in complex, large-scale social networks. This technique is aligned with two previously developed graphical model approaches, namely *SSN-LDA* (Simple Social Network-Latent Dirichlet Allocation) [9] and *GWN-LDA*(Generic Weighted Network-Latent Dirichlet Allocation) [8], which discover hidden correlations among social actors using hierarchical Bayesian network models. However, the *HSN-PAM* model is able to discover not only correlations among social actors in networks but also correlations among hidden groups, thus making it possible to uncover complicated, hierarchical community structures.

In the rest of this paper, Section 2 introduces related studies; Section 3 introduces related terminology and notations for the *HSN-PAM* model and its corresponding learning procedures; Section 4 describes experimental results; Section 5 concludes the paper.

2 Related Work

Probabilistic graphical models such as Bayesian networks have been widely used as an important machine learning technique to represent dependency relations between visible and hidden random variables. As a well-received probabilistic graphical model, LDA(Latent Dirichlet Allocation) model was first introduced by Blei for modeling the generative process of a document corpus [1]. Its ability of modeling topics using latent variables has attracted significant interests and it has been applied to many domains such as document modeling [1], text classification [1], collaborative filtering [1], topic models detection [7, 6], and community discovery [9, 8]. The two topological community discovery approaches, *SSN-LDA* [9] and *GWN-LDA* [8], attempt to discover flat communities from social networks by utilizing only topological information in social networks. These two models encode the structural information of networks into profiles and discover community structures purely from these social connections among the nodes. With the only input information being the topological structure of a social network, these models can be easily extended to complex networks where no semantic information is available. PAM is DAG-structured mixture model that was proposed to capture the correlations among topics by introducing a DAG-structured mixture models [3]. This paper describes a community discovery approach, *HSN-PAM*, based on this hierarchical graphical model.

3 HSN-PAM model

In the hierarchical community structure that will be described in this section, namely *HSN-PAM*, the concept of communities is extended to include two different types of communities, namely *regular communities* and *super communities*. The two types of communities are denoted as ι^s (*super communities*), and ι^r (*regular communities*). A *regular community* is defined as a distribution on the social actor space while a *super community* is considered as a distribution on the *regular communities* or *super communities*. There can be arbitrary number of *super community* levels in *HSN-PAM*. In this section, we introduce related terminology and network encoding schemes for social networks in Sections 3.1 and 3.2 respectively and then describe a simplified *HSN-PAM* model, namely *TLC-HSN-PAM*, with a two-level community structure. Finally, the Gibbs sampler for solving *TLC-HSN-PAM* model is presented in Section 3.4.

3.1 Terminology and definitions

A typical social network G, as shown in Figure 1, is composed of a pair of sets, including the social actor set $V = \{v_1, v_2, ..., v_M\}$ and social interaction set $E(e_1, e_2, ..., e_N)$, together with a **Social Interaction Weight** function: $SIW : (V \times V) \rightarrow \mathbf{I}$, where I represents the integer set. The elements of social actor set V are the vertices of G and the elements of social interaction set E are the edges of G, representing the occurrence of social interactions between the corresponding social actors. The number of the social actors in the network is denoted as M. Each social interaction e_i in set E is considered as a binary relation between two social actors, i.e. $e_i(v_{i_1}, v_{i_2})$ and SIW function describes the strength of such interaction. Note that social interaction weight is specified as integer in order to be processed by the *HSN-PAM* model. Throughout this paper, terms *node*, *vertex*, and *social actor* are used interchangeably, and so are ; *edge* and *social interaction*.

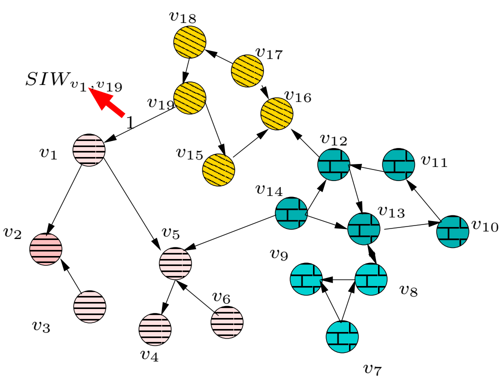

Figure 1. A typical social network

A node v_i's neighboring agents are encoded by vector $\vec{\omega_i}$ and each element $\omega_{ij} \in V$ in the vector represents node v_i's j^{th} neighbor. The connectivity of v_i in the network is characterized by its *social interaction profile* (SIP), which is defined as a sequence of all v_i's neighbors(ω_{ij}). In this sequence, the frequency of a neighbor ω_{ij} is set as the corresponding social interaction weight information ($SIW(v_i, \omega_{ij})$). Formally, v_i's social interaction profile is:

$$\vec{s_i} = (\omega_{i1}, \cdots, \omega_{i1}, \omega_{i2}, \cdots, \omega_{i2}, \cdots, \omega_{iN_i}, \cdots, \omega_{iN_i})$$

where N_i is the number of v_i's neighboring nodes and the count of a particular neighboring node ω_{ij} in $\vec{s_i}$ is $SIW(v_i, \omega_{ij})$. Throughout this paper, the variables in sequence $\vec{s_i}$ is specified as s_{ij}, where $s_{ij} \in \vec{\omega_i} \subseteq V$. Note that we assume the social interaction elements in this profile are exchangeable and therefore their order will not be concerned. This exchangeability allows these graphical models be used in this application domain [1].

3.2 Network encoding scheme

The set of social interaction profiles collectively determines the topological structure of a social network. The

Table 1. Notation for quantities in *HSN-PAM*

ι	hidden community variable
ι^s	super community variable
ι^r	regular community variable
γ	the root node
$\iota_{i,j}$	community for the jth social interaction in $\vec{s}_i$
$\vec{\theta}$	$p(\iota\|\vec{s}_j)$community mixture proportion for $\vec{s}_j$
$\vec{\phi}_k$	$p(s_{ki}\|\iota_k)$ the mixture component of community k

HSN-PAM model depends on the profile information to learn the graphical model and identify hidden communities in the pertaining social networks. In this paper we explore a straightforward encoding scheme, namely *DNES* , to generate social interaction profiles. In the *DNES* scheme, a social actor v_i's social interaction profile contains all directly connected neighbors and the count of each neighbor in the profile is 1. Hence, the social interaction profiles of all the social actors constitute the adjacent matrix of the social network. More formally, the SIW function is defined as:

$$SIW_D(v_{i_1}, v_{i_2}) = \begin{cases} 1 & \text{if } e(v_{i_1}, v_{i_2}) \in E; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

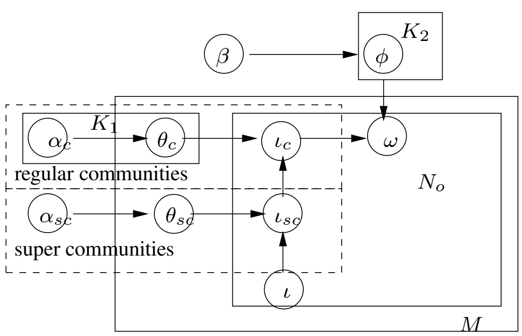

Figure 2. Graphical Model for *TLC-HSN-PAM*

3.3 TLC-HSN-PAM Model

This paper focuses on a simplified, two-level community structure, i.e *TLC-HSN-PAM* model, which is shown in Figure 3. The two level community structure consists of two types of communities: super communities $\vec{\iota^s} = \{\iota^s_1, \iota^s_2, ..., \iota^s_{k_1}\}$ and regular communities $\vec{\iota^r} = \{\iota^r_1, \iota^r_2, ..., \iota^r_{k_2}\}$. Figure 3 demonstrates that the root community γ is connected to all super communities and all super communities are fully connected to regular communities. Finally, regular communities are fully connected to all the social actors in the social network. associated with communities are Dirichlet component multinomials(DCM), Dir_{ι_i} [5]. A DCM distribution is defined as a distribution hierarchy, including a multinomial distribution and a Dirichlet prior. Dirichlet is often used as the prior distributions for multinomial distributions in Bayesian statistics in order to obtain close-form solutions. In the context of *HSN-PAM*, This means that the social interaction profile is generated by a multinomial distribution whose parameters are generated by its Dirichlet prior distribution.

Two different types of distributions are used in this two-level community structure. We specify that the distributions of root and super communities are Dirichlet component multinomial (DCM) distributions while the distributions of regular communities are modeled with fixed multinomial distributions $\phi_{\iota^r_i}$, sampled once for the whole social network from a single Dirichlet distribution $Dir(\beta)$. A DCM distribution is defined as a distribution hierarchy, including a multinomial distribution and a Dirichlet prior [5]. Dirichlet is often used as the prior distributions for multinomial distributions in Bayesian statistics in order to obtain close-form solutions. The corresponding graphical model is shown in Figure 2; The multinomials for the root and super communities are sampled individually for each social interaction profile. Each community ι_i is associated with a Dirichlet distribution.

Based on the graphical model in Figure 2, the generative process for a social actor's social interaction profile $\vec{s}_j$ is a two-step process:

1. Sample $\vec{\theta^j_r}$ from the root $Dir_t(\alpha_t)$, where $\vec{\theta^j_t}$ is a multinomial distribution over super-communities.
2. For each super-community ι^s_i, sample $\vec{\theta}_{\iota^s_i}$ from $Dir_i(\alpha_i)$, where $\vec{\theta}_{\iota^s_i}$ is a multinomial distribution over regular communities.
3. For each social actor in the social interaction profile,
   (a) Sample a super-community ι_ω from $\vec{\theta_r}$;
   (b) Sample a regular community ι^r_j from $\vec{\theta}_{\iota_\omega}$;
   (c) Sample word ω from $\vec{\phi}_{\iota^r_j}$.

The model structure and the generative process for this special setting are similar to *SSN-LDA* approach. The major difference is that it has one additional layer of super-topics modeled with Dirichlet multinomials, which is the key component capturing correlations among communities here. Another way to interpret this is that given the regular communities, each super-community is essentially an individual *SSN-LDA* structure. Therefore, this can be viewed as a mixture over a set of *SSN-LDA* models. Following this process, the joint probability of generating a social interaction profile, the community assignment $\vec{\iota}$, and the multinomial distribution $\vec{\theta}$ is:

$$P(\vec{s}_i, \vec{\iota}, \vec{\theta}|\alpha, \phi) = P(\vec{\theta}_t|\alpha_t)$$

$$\prod_{i=1}^{s} P(\vec{\theta}_{\iota_i}|\alpha_i) \times \prod_{\omega}(P(\iota_\omega|\vec{\theta}_r)P(\iota_j^r|\vec{\theta}_{\iota_\omega})P(\omega|\phi_{\iota_j^r})) \quad (2)$$

Integrating out $\vec{\theta}$ and summing over $\vec{\iota}$, we calculate the marginal probability of a social interaction profile as:

$$P(\vec{s}_i|\alpha, \Phi) = \int P(\vec{\theta}_t|\alpha_t) \prod_{i=1}^{s} P(\vec{\theta_{\iota_i}}|\alpha_i)$$

$$\times \prod_{\omega} \sum_{\iota_\omega} (P(\iota_\omega|\vec{\theta}_t)P(\iota_j^r|\vec{\theta}_{\iota_\omega^s})P(\omega|\Phi_{\iota_j^r})d\vec{\theta} \quad (3)$$

The probability of generating the entire social network $\vec{S}$ is the product of the probability for every social interaction profile $\vec{s}_i$, integrating out the multinomial distributions for regular communities Φ:

$$P(\vec{S}|\alpha, \beta) = \int \prod_j P(\phi_{\iota_j^r}|\beta) \prod_{\vec{s}_i} P(\vec{s}_i|\alpha, \Phi) d\phi$$

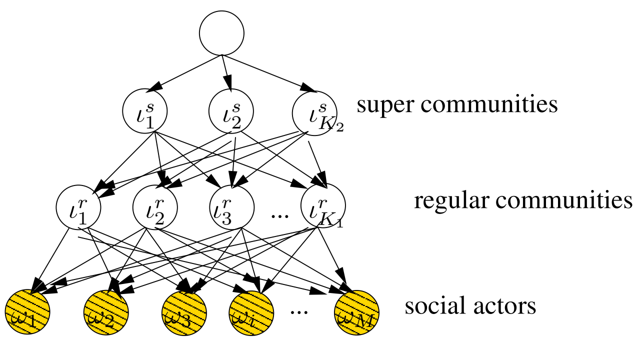

Figure 3. Tree structure of a two-level community structure *TLC-HSN-PAM* model, including K_2 super communities, K_1 regular communities, and M social actors.

3.4 Gibbs Samplers for HSN-PAM

Exact inference is generally intractable for even the two-level community *HSN-PAM* model. We employ Gibbs sampling to learn *HSN-PAM* models because it often yields relatively simple algorithms for approximate inference in high-dimensional models. Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) simulation [4] where the dimension K of the distribution are sampled alternately one at a time, conditioned on the values of all other dimensions [2].

For an arbitrary DAG, we need to sample a community path for each social actor given other variable assignments enumerating all possible paths and calculating their conditional probabilities. In the two-level community structure *HSN-PAM* model, each path contains the root, a super-community, and a regular community. Since the root is fixed, we only need to jointly sample the super-community and regular community assignments for each social actor, based on their conditional probability given observations and other assignments, integrating out the multinomial distributions. Θ; (thus the time for each sample is in the number of possible paths). The following equation shows the conditional probability given the assignment of other regular and super communities. For social actor ω_j in social interaction profile $\vec{s}_i$, we have:

$$p(\iota_{w2} = k_2, \iota_{w3} = k_3|D, \iota_{-w}, \alpha, \beta) \propto$$
$$\frac{n_{1k}^{(d)} + \alpha_{ak}}{n_1^{(d)} + \Sigma_{k'}\alpha_{1k'}} * \frac{n^{(d)} + \alpha_{kp}}{n_k^{(d)} + \Sigma_{p'}\alpha_{kp'}} * \frac{n_{pw} + \beta_w}{n_p + \Sigma_m \beta_m}.$$

Here we assume that the root community is k_1, ι_{w2} and ι_{w3} correspond to super community and regular community assignments respectively. ι_{-w} is the community assignments for all other social actors. Excluding the current social actor, $n_x^{(d)}$ is the number of occurrences of community k_x in social interaction profile sip; $n_{xy}^{(d)}$ is the number of times community k_y is sampled from its parent k_x in social interaction profile; n_x is the number of occurrences of regular-community k_x in the whole network and n_{xw} is the number of occurrences of social actor ω in regular-community k_x. Furthermore, α_{xy} is the yth component in α_x and β_w is the component for social actor ω in β.

Note that in the Gibbs sampling equation, we assume that the Dirichlet parameters are given. While *SSN-LDA* can produce reasonable results with a simple uniform Dirichlet, we have to learn these parameters for the super-communities in *TLD-HSN-PAM* since they capture different correlations among regular-communities. As for the root, we assume a fixed Dirichlet parameter. To learn α, we could use maximum likelihood or maximum a posterior estimation. However, since there are no closed-form solutions for these methods and we wish to avoid iterative methods for the sake of simplicity and speed, we approximate it by moment matching. In each iteration of Gibbs sampling, we update

$$\mu_{xy} = \frac{1}{N} * \Sigma_d \frac{n_{xy}^{(d)}}{n_x^{(d)}};$$

$$\sigma_{xy} = \frac{1}{N} * \Sigma_d (\frac{n_{xy}^d}{n_x^{(d)}} - \mu_{xy})^2;$$

$$m_{xy} = \frac{\mu_{xy} * (1 - \mu_{xy})}{\sigma_{xy}} - 1;$$
$$\alpha_{xy} \propto \mu_{xy};$$
$$\Sigma_y(\alpha_{xy}) = \frac{1}{5} * exp(\frac{\Sigma_y log(m_{xy})}{s_2 - 1}).$$

For each super-community k_x and regular-community k_y, we first calculate the sample mean μ_{xy} and sample variance σ_{xy}. $n_{xy}^{(d)}$ and $n_x^{(d)}$ are the same as defined above. Then we estimate α_{xy}, the yth component in α_x from sample mean and variance. N is the number of social actors and s_2 is the number of regular communities.

Smoothing is important when we estimate the Dirichlet parameters with moment matching. From the equations above, we can see that when one regular-community y does not get sampled from super-community x in one iteration, α_{xy} will become 0. Furthermore, from the Gibbs sampling equation, we know that this regular-community will never have the chance to be sampled again by this super-community. We introduce a prior in the calculation of sample means so that μ_{xy} will not be 0 even if $n_{xy}^{(d)}$ is 0 for every social interaction profile sip.

Table 2. An illustration of 4 regular communities that belong to the 48th super community (ι_{48}^s)

.

Community 63	Community 19
Signal Processing	*Learning,Robot*
Marc Moonen	Manuela Veloso
Robert W Dutton	Peter Stone
Brian L Evans	Anthony Skjellum
Thomas H Lee	Boi Faltings
Jung suk Goo	Edmund Burke
Community 140	Community 185
Medical,Image	*Multimedia, learning*
Ron Kikinis	Thomas S Huang
Ferenc A. Jolesz	Shih fu Chang
Simon K. Warfield,	Anoop Gupta
Mark A. Musen	Gonzalo Navarro
Martha Shenton	Kathleen R Mckeown

4 Experiments and Evaluation

We evaluate two-level community structure *HSN-PAM* on *CiteSeer*. CiteSeer is a free public resource created by Kurt Bollacker, Lee Giles, and Steve Lawrence in 1997-98 at NEC Research Institute (now NEC Labs), Princeton, NJ. It contains rich information on the citation, co-authorship, semantic information for computer science literature. In this paper we only consider the co-authorship information which constitutes a large-scale social network regarding academic collaboration with diversities spanning in time, research fields, and countries. *CiteSeer* contain unconnected subnetworks and the size of the largest connected subnetwork of *CiteSeer* is 249866. In this paper, we are only interested in discovering communities in the two largest subnetworks. Therefore, unless specially specify, we always mean the two subnetworks when referring *CiteSeer*.

Throughout the experiments, we assume a fixed Dirichlet distribution with parameter 0.01 for the root node. We can change this parameter to adjust the variance in the sampled multinomial distributions. We choose a small value so that the variance is high and each document contains only a small number of super communities, which tends to make the super communities more interpretable. We treat the regular communities in the same way as *SSN-LDA* and assume that they are sampled once for the whole corpus from a given Dirichlet with parameter 0.01. So the only parameters we need to learn are the Dirichlet parameters fro the super communities and multinomial parameters for the regular communities. For cross-validation purposes, 10% of the original datasets is held out as test set and we run the Gibbs sampling process on the training set for i iteration. In particular, in generating the exemplary communities, we set the number of the communities as 50, the iteration times i as 1000. α is set as $\frac{1}{K}$ and β is set as 0.01, where K is the number of the communities.

Tables 2, 3 demonstrate some exemplary communities that are discovered by *TLC-HSN-PAM* algorithm for the *CiteSeer* dataset with social interaction profiles being created using *DNES* enconding scheme. Each community is shown with the top 5 researchers that have the highest probability conditioned on the community. Note that *CiteSeer* dataset was crawled from Web and some authors were not recovered correctly, we keep the results in an “as is” fashion. In this dataset, the number of super communities is set as 50 while the number of regular communities is set as 200. These results illustrate that researchers from the regular communities that belong to the same a super community are often interested in related subjects. For instance, the four top regular communities in ι_{48}^s, as shown in Figure 2, include researchers that are working on “Signal processing” (ι_{63}^r), “Robot and learning” (ι_{19}^r), “Medical and image processing” (ι_{140}^r), and “Multimedia and learning” (ι_{185}^r) topics. Similarly, Figure 3 lists four regular communities that belong to super community ι_{36}^s, including four relevant areas such as “Agent and AI” (ι_{179}^r), “Algorithm theory” (ι_{33}^r), “Multi-Agent and distributed systems” (ι_{165}^r), and “Multimedia and learning” (ι_{185}^r). Note that a regular community can belong to many related super communities. For instance, regular community ι_{185}^r belongs to both super

Table 3. An illustration of 4 regular communities that belong to the 36th super community

Community 179	Community 33
Agent AI	*Algorithm Theory*
Nicholas R Jennings	Micha Sharir
Simon Parsons	Pankaj K Agarwal
Michael Wooldridge	John H Reif
Peter Mcburney	Boris Aronov
Timothy J. Norman	Leonidas J Guibas
Community 165	Community 185
Multi-Agent, distributed	*Multimedia, Learning*
Victor Lesser	Thomas S Huang
Thomas Wagner	Shih fu Chang
David Kotz	Anoop Gupta
Michael Gerndt	Gonzalo Navarro
Heinz Stockinger	Kathleen R Mckeown

community ι^s_{48} and ι_{36}.

In addition to empirical analysis on discovered communities, we also provide quantitative measurements to compare *HSN-PAM* with *SSN-LDA* approach. In Figure 4, *SSNLDA*, *S-4-HSNPAM*, and *S-10-HSNPAM* illustrate the likelihood for *SSN-LDA* and *HSN-PAM* models when the number of super communities is set as 4 and 10 respectively. Likelihood values indicate the uncertainty in predicting the occurrence of a particular social interaction given the parameter settings, and hence they reflect the ability of a model to generalize unseen data. The x axis represents the number of regular communities. This figure demonstrates that in general *HSN-PAM* is able to produce better higher likelihood value. These curves can be used to detect the approximate optimal regular communities given the number of super communities.

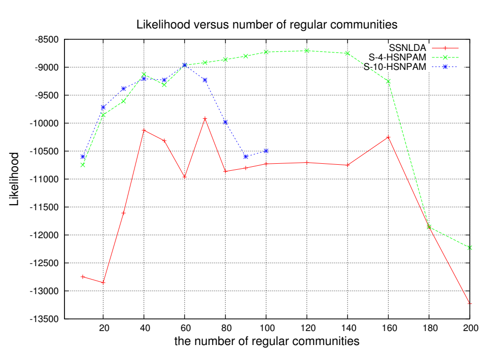

Figure 4. Likelihood versus the number of communities.

5 Conclusions and Future Work

Real-world social networks are often hierarchical, reflecting the fact that some communities are composed of a few smaller, sub-communities. This paper describes a hierarchical Bayesian model based scheme, namely *HSN-PAM* (Hierarchical Social Network-Pachinko Allocation Model), for discovering probabilistic, hierarchical communities in social networks. In this scheme, communities are classified into two categories: *super-communities* and *regular-communities*. Two different network encoding approaches are explored to evaluate this scheme on research collaborative networks, including *CiteSeer* and *NanoSCI*. The experimental results demonstrate that *HSN-PAM* is effective for discovering hierarchical community structures in large-scale social networks.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] G. Heinrich. Parameter estimation for text analysis. *Technical Report*, 2004.

[3] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, pages 577–584, 2006.

[4] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

[5] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *UAI '02: Proceedings of the 18th conference on Uncertainty in artificial intelligence*, 2002.

[6] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, pages 424–433, 2006.

[7] X. Wang, N. Mohanty, and A. McCallum. Group and topic discovery from relations and their attributes. In *NIPS*, 2005.

[8] H. Zhang, C. L. Giles, H. C. Foley, and J. Yen. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *AAAI*, pages 663–668, 2007.

[9] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. An LDA-based community structure discovery approach for large-scale social networks. In *IEEE International Conference on Intelligence and Security Informatics*, pages 200–207, 2007.