Probabilistic Community Discovery Using Hierarchical Latent Gaussian Mixture Model

Haizheng Zhang and **C. Lee Giles** and **Henry C. Foley** and **John Yen**
College of Information Science and Technology
Pennsylvania State University
University Park, PA 16803
{hzhang,giles,hfoley,jyen}@ist.psu.edu

Abstract

Complex networks exist in a wide array of diverse domains, ranging from biology, sociology, and computer science. These real-world networks, while disparate in nature, often comprise of a set of loose clusters(a.k.a communities), whose members are better connected to each other than to the rest of the network. Discovering such inherent community structures can lead to deeper understanding about the networks and therefore has raised increasing interests among researchers from various disciplines. This paper describes GWN-LDA(Generic weighted network-Latent Dirichlet Allocation) model, a hierarchical Bayesian model derived from the widely-received LDA model, for discovering probabilistic community profiles in social networks. In this model, communities are modeled as latent variables and defined as distributions over the social actor space. In addition, each social actor belongs to every community with different probability. This paper also proposes two different network encoding approaches and explores the impact of these two approaches to the community discovery performance. This model is evaluated on two research collaborative networks:*CiteSeer* and *NanoSCI*. The experimental results demonstrate that this approach is promising for discovering community structures in large-scale networks.

Introduction

Complex networks exist in a wide range of real world systems, such as communication networks, ecological webs, protein interaction networks, and social networks. Despite their disparate nature, these networks often exhibit common topological properties, including the small-world property, and power-law degree distribution. In addition, some members in the networks form loose clusters, making them better connected to each other than to the rest of the network. Discovering and identifying these clusters is referred as community discovery problem which has raised significant interest among researchers from a variety of disciplines.

While the concept of community is self-explanatory, there is no quantitative, rigorous definition that is commonly accepted. This is partly due to the fact that members in the network can potentially belong to more than one communities and the boundaries between communities are often blurry and difficult to draw. 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 and multiple membership phenomenon. Therefore, the outcome of such algorithms is often artificial and difficult to explain. In order to address this problem, we design a hierarchical Bayesian network based approach, namely *GWN-LDA*(Generic-Weighted Network-LDA), to discover probabilistic communities from complex networks. This model is inspired by the success of the application of LDA(Latent Dirichlet Allocation) models in information retrieval and image analysis domains. In the *GWN-LDA* model, communities are modeled as latent variables and are considered as distributions on the entire social actor space. Therefore, each social actor contributes a part, big or small, to every community in the society. In addition, this paper proposes and compares two different approaches to encode the networks in order to be processed by the *GWN-LDA* model. Finally, this latent probabilistic model and the two pertaining network encoding approaches are evaluated on two co-authorship networks from two distinct academic communities, i.e *NanoSCI* from the nanotechnology domain and *CiteSeer* from the computer science domain. Note that while this approach is evaluated in the social network domain with co-authorship networks, it can be easily extended to other complex network-based applications.

In conclusion, the contributions of this paper include: (1) an LDA-based probabilistic community discovery model in large-scale generic-weighted networks which only requires the topological structure of networks; (2) the exploration of the impact of different network encoding strategies on the community discovery.

The rest of this paper is organized as follows. We first introduces related studies and then present the technical details of *GWN-LDA* model and the corresponding Gibbs sampler. In the experiment and result section, we describe the two co-authorship networks and two encoding approaches, followed by the experimental settings and results analysis. Finally we conclude the paper with a brief summary of this model.

Background and Related Works

Community discovery problems have been studied in a variety of networks, including World Wide Web(Flake, Lawrence, & Giles 2000), distributed information retrieval(Zhang *et al.* 2004), social networks(Clauset, Newman, & Moore 2004; Girvan & Newman 2002; Newman 2004b; Palla *et al.* 2005; Scott 2000; Zhou *et al.* 2006b; Newman 2004a), and biological networks(Girvan & Newman 2002; Palla *et al.* 2005; Wilkinson & Huberman 2004). Most of these approaches are characterized by the use of distance-based measures, including *Centrality indices*(a.k.a betweeness)(Freeman 1977; Girvan & Newman 2002; Wilkinson & Huberman 2004; Ruan & Zhang 2006) or *Minimum cut approaches*(Flake, Lawrence, & Giles 2000). The common ground for these studies is the definition of community distance measures and (iterative) clustering process for minimizing such measures. This paper describes a probabilistic community discovery model, *GWN-LDA*, based on mixture-model based approach. LDA model was first introduced by Blei for modeling the generative process of a document corpus(Blei, Ng, & Jordan 2003). Its ability of modeling topics using latent variables has attracted significant interest and it has been applied to many domains including document modeling (Blei, Ng, & Jordan 2003; Li & McCallum 2006), text classification (Blei, Ng, & Jordan 2003), image processing(Sudderth *et al.* 2005), contextual community discovery(Zhou *et al.* 2006b; 2006a).

GWN-LDA, similar to a previously developed model (*SSN-LDA*), encodes the structural information of networks into profiles and discovers community structures purely based on the social connections among the social actors. These two models do not depend on semantic information as does in (Zhou *et al.* 2006b; 2006a). However, the major drawback of *SSN-LDA* approach is its inability of modeling the weight of social interactions since the community component distribution is specified as multinomial distribution with a Dirilet prior and unable to handle the situation when weight is real number. *GWN-LDA* model uses a Gaussian distribution with inverse-Wishart prior to model the arbitrary weights that are associated with the social interaction occurrences.

LDA Model for Generic-Weighted Networks

This section describes the *GWN-LDA* model. Before diving into the details, we introduce the technical details of *GWN-LDA* model and the corresponding Gibbs sampler.

GWN-LDA Model Description

A typical social network is composed of a pair of sets, including the social actor set $V = \{v_0, v_1, ...v_M\}$ and social interaction set $E = \{e_0, e_1, ..., e_N\}$, together with a **Social Interaction Weight** function: $SIW : (V \times V) \rightarrow \mathbf{R}$. The number of the social actors is denoted as M. The elements of social actor set V are the vertices of the network and the elements of social interaction set E are the edges of G, representing the occurrence of social interactions between the corresponding social actors. 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. Different from *SSN-LDA*, SIW function in *GWN-LDA* function can take on arbitrary real number. In this paper, *vertex* and *social actor*, *edge* and *social interaction* are used interchangeably.

A node v_i's neighboring agents are encoded by the variable $\vec{\omega_i}$ and $\omega_{ij} \in V$ means node v_i's j_{th} neighbor. Each actor is characterized by its *social interaction profile* (SIP), which is defined as a set of neighbor(ω_{ij}) and the corresponding weight($SIW(v_i, \omega_{ij})$) pair. Formally,

$$SIP(v_i) = \{(\omega_{i1}, SIW(v_i, \omega_{i1})), \cdots, (\omega_{iN_i}, SIW(v_i, \omega_{iN_i}))\}$$

where N_i is the size of v_i's social interaction profile. Note that we consider the social interaction elements in this profile are exchangeable and therefore their order will not be concerned. It is this exchangeability that permits the application of LDA model(Blei, Ng, & Jordan 2003).

Subsequently, we specify that a social network contains K communities $\iota(\iota_1, \iota_2, ..., \iota_k)$ and each *community* in ι is defined as a distribution on the social actor space. In *GWN-LDA*, community assignments are modeled as a latent variable(ι) in the graphical model. The community proportion variable (θ) is regulated by a Dirichlet distribution with a known parameter α. Meanwhile, each social actor belongs to every community with different probabilities and therefore its social interaction profiles can be represented as random mixtures over latent communities variables. The graphical model is illustrated in Figure 1, where each social interaction profile is considered as a multinomial distribution of community variables and each community is modeled as a multivariate Gaussian distribution of social interaction weight variables. The parameter sets for the two distributions are specified as $\mathbf{\Theta} = \{\vec{\theta_m}\}_{m=1}^{M}$ and $\underline{\mathbf{\Omega}} = \{\vec{\mu_k}, \vec{\Sigma_k}\}$ respectively.

Given $\vec{\theta_i}$ and $\underline{\mathbf{\Omega}}$, the probability of $\omega_{i,j} = r$ is:

$$p(\omega_{i,j} = r|\vec{\theta_i}, \underline{\mathbf{\Omega}}) = \sum_{j=1}^{K} p(\omega_{i.j} = r|\vec{\mu_k}, \vec{\Sigma_k})p(\iota_{i,j} = k|\vec{\theta_i}) \tag{1}$$

$$p(\vec{\omega_m}|\vec{\mu_m}, \vec{\Sigma_m}) \propto |\Sigma_m|^{-} \; exp(-\frac{1}{2}tr(\Sigma_m^{-1} S_0)) \tag{2}$$

where $S_0 = \sum_{i=1}^{n}(\omega_{mi} - \mu_m)^T(\omega_{mi} - \mu_m)$

In order to obtain a close-form distribution, the prior distribution for the social interaction profiles is set as a Dirichlet prior with hyperparameter α and the prior for community component distribution is set as Multivariate Gaussian/inverse-Wishart distribution. For the sake of simplicity, we denote the hyperparameter for the prior distribution is $\underline{\mathbf{\Psi}} = \{\vec{\mu_0}, \vec{\kappa_0}\}$ and $\underline{\mathbf{\Upsilon}} = \{\vec{v}, \lambda\}$ respectively. According to this prior distribution definition, the joint prior density is (Gelman *et al.* 2004):

$$p*(\underline{\mathbf{\Omega}}) \propto |\Sigma|^{-(\frac{(v_0+d)}{2}+1)}$$
$$exp\left(\frac{-1}{2}tr(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu - \mu_0)^T\Sigma^{-1}(\mu - \mu_0)\right)$$

The parameters υ_0 and Λ_0 describe the degree of freedom and the scale matrix for the inverse-Wishart distribution on Σ. The remaining parameters are the prior mean, μ_0, and the number of prior measurements, κ_0 on the Σ scale.

Specifically,

$$\Sigma \sim W^{-1}(\upsilon_0, \Lambda_0^{-1}) \tag{3}$$

$$\mu|\Sigma \sim N(\mu_0, \frac{\Sigma}{\kappa_0}), \tag{4}$$

where W^{-1} represents inverse-Wishart distribution, which is a multivariate generalization of the scaled inverse-χ^2 distribution. The *pdf* function and more information about the Wishart distribution can be referred to (Gelman *et al.* 2004).

From the topology of the Bayesian network, we can further specify that the complete-data likelihood of a social interaction profile (the joint distribution of all known and hidden variables given the hyperparameters:

$$p(\vec{\omega_i}, \vec{\iota_i}, \vec{\theta_i}, \underline{\mathbf{\Omega}}|\vec{\alpha}, \vec{\upsilon}, \vec{\Lambda}, \vec{\kappa}, \vec{\mu_0}) = \prod_{j=1}^{N_i} p(\omega_{i,j}|\vec{\mu_k}, \Sigma_k)p(\iota_{i,j} = k|\vec{\theta_i})p(\vec{\theta_i}|\vec{\alpha})p(\vec{\mu}, \vec{\Sigma}|\vec{\upsilon}, \vec{\Lambda}, \vec{\kappa}, \vec{\mu_0})$$

Finally the likelihood of the complete network $W = \{\vec{\omega_m}\}_{m=1}^{M}$ is determined by the product of the likelihoods of the independent nodes:

$$p(W|\vec{\alpha}, \Psi, \Upsilon) = \prod_{m=1}^{M} p(\omega_m|\vec{\alpha}, \Psi, \Upsilon)$$

The generative process can be explained as that the Bayesian network of *GWN-LDA* model generates a stream of observable edges $\omega_{m,n}$, partitioned into node profile $\vec{\omega_m}$. For each of these node profiles, a profile proportion θ_i is drawn, and from this, profile-specific edges are emitted. That is, for each connection, a community indicator $\iota_{m,n}$ is sampled according to the node-specific mixture proportion, and then the corresponding community-specific node distribution ω_m used to draw a connection and its corresponding weight. The communities ι_i are sampled once for the entire network.

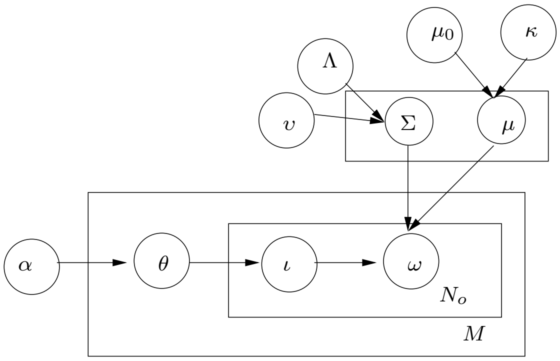

Figure 1: Graphical Model for *GWN-LDA*

Gibbs Samplers for GWN-LDA

The desired distribution for solving the *GWN-LDA* model is the posterior probability given evidence $p(\iota|\omega)$.

$$p(\iota|\omega) = \frac{p(\omega, \iota)}{\sum_\iota p((\omega), \iota)} \tag{5}$$

However, the computation complexity of the the denominator part is prohibitively high. Analogous to the derivation in *SSN-LDA* model, in order to derive the posterior distribution over community assignment in *GWN-LDA* model, $p(\iota_j|\vec{\iota_{\neg j}}, \vec{\omega})$, we start from the joint distribution:

$$p(\vec{\omega}, \vec{\iota}|\vec{\alpha}, \underline{\mathbf{\Upsilon}}, \underline{\mathbf{\Psi}}) = p(\vec{\omega}|\vec{\iota}, \underline{\mathbf{\Upsilon}}, \underline{\mathbf{\Psi}})p(\vec{\iota}|\vec{\alpha}) \tag{6}$$

And

$$p(\vec{\omega}|\vec{\iota}, \underline{\mathbf{\Upsilon}}, \underline{\mathbf{\Psi}}) = \int p(\vec{\omega}|\vec{\iota}, \underline{\mathbf{\Omega}})p(\underline{\mathbf{\Omega}}|\underline{\mathbf{\Upsilon}}, \underline{\mathbf{\Psi}})d\underline{\mathbf{\Omega}} \tag{7}$$

$p(\vec{\omega}|\vec{\iota}, \underline{\mathbf{\Omega}})$ is determined by Formula 2 and $p(\underline{\mathbf{\Omega}}|\underline{\mathbf{\Upsilon}}, \underline{\mathbf{\Psi}})$ is determined by Formula 1. Integrating over the parameters of the normal-inverse-Wishart posterior distribution in equation 7, the predictive likelihood of a new observation is a multivariate Student-t with $(\upsilon - d + 1)$ degree of freedom. Such a distribution can be approximated by a moment-matched Gaussian(Sudderth *et al.* 2005) distribution:

$$p(\omega_{i,j}|\iota_{ij} = k, \iota_{\vec{\neg i.j}}, \vec{\omega}, \underline{\mathbf{\Psi}}, \underline{\mathbf{\Upsilon}}) \approx N(x_{i,j}; \mu_k, \Sigma_k) \tag{8}$$

where

$$\mu_k = \frac{1}{n_k^{(\cdot)}} \sum_{i=1}^{M} \sum_{l|z_{i,l}=k} \omega_{i,l} \tag{9}$$

$$\delta_j = \frac{n_k + 1}{n_k(n_k + \upsilon_p - 3)} \tag{10}$$

$$\Sigma_k = \delta_k \left(\Delta_p + \sum_{i=1}^{M} \sum_{l|\iota_{m,l}=k} (\omega_{il} - \mu_k)(\omega_{il} - \mu_k)^T \right) \tag{11}$$

Finally,

$$p(\iota_{ij} = k|\omega_{i,j}, \iota_{\vec{\neg i.j}}, \omega_{\vec{\neg i,j}}, \vec{\alpha}, \underline{\mathbf{\Psi}}, \underline{\mathbf{\Upsilon}}) \propto p(\omega_{i,j}|\iota_{ij} = k, \iota_{\vec{\neg i.j}}, \vec{\omega}, \underline{\mathbf{\Psi}}, \underline{\mathbf{\Upsilon}}) * p(\iota_{i,j} = k|\vec{\alpha}) \tag{12}$$

where

$$p(\iota_{i,j} = k|\vec{\alpha}) = \int p(\vec{\iota}|\Theta)p(\Theta|\vec{(\alpha)})d\Theta \tag{13}$$

$$= \prod_{m=1}^{M} \frac{\Delta(\vec{n_m} + \vec{\alpha})}{\Delta(\vec{\alpha})} \tag{14}$$

The Gibbs sampling process is analogous to the *SSN-LDA* model and the detailed algorithm is elaborated in (Zhang *et al.* 2007).

Table 1: Statistics for datasets *CiteSeer* and *NanoSCI*,*PN* denotes the number of papers; *EN* denotes the number of edges; *AAP* denotes the average author number per paper, and *SLC* denotes the size of the largest component

Dataset	Size	*PN*	*EN*	*AAP*	*SLC*
CiteSeer	398831	716793	1181133	1.65	249866
NanoSCI	225313	195997	877609	4.48	203762

Experiments and Evaluation

Two Co-Authorship Networks

In co-authorship networks, the vertices represent researchers and the edges in the network represent the collaboration relation between researchers. In this section we evaluate *GWN-LDA* model on two co-authorship networks collected from computer science(*CiteSeer*) and nanotechnology(*NanoSCI*). Note that no name disambiguation has been done on either dataset.

***CiteSeer* Dataset** CiteSeer is a free public resource created by Kurt Bollacker, C. 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.

NanoSCI *NanoSCI* is a collection of nanotechnology related articles published and indexed by *SCI*(Science Citation Index) in 2000-2006 period. The records are acquired by inquiring *Thomson Scientific* website (http://scientific.thomson.com/products/sci/) directly. The query used in this paper is generated using an iterative relevance feedback technique (Kostoff *et al.* 2006). The essential idea of this approach is to augment the keyword set until the returned results converges.

Table. 1 lists the statistics for the two data collections. Both *CiteSeer* and *NanoSCI* contain unconnected subnetworks. In particular, *CiteSeer* contains 31998 subgraphs and *NanoSCI* contains 5241 unconnected subnetworks. The size of the largest connected sub-network of *CiteSeer* is 249866 while the size of the largest connected subnetwork in *NanoSCI* is 203762. In this paper, we are only interested in discovering community structures in the two largest sub-networks. Therefore, unless specially specify, we always mean the two sub-networks when referring *CiteSeer* and *NanoSCI*.

Encoding Social Interaction Profiles

The social interaction profiles of the social actors collectively determines the structure and dynamics of a social network. In this paper, we explore two different types of social interaction profile representations for social networks, namely *01-SIP*, and *Real-SIP*. It is worth to mention that such exploration is by no means comprehensive. Nevertheless it provides valuable insights for designing more sophisticated social interaction profile encoding schemes.

01-SIP In the *01-SIP* approach, an edge is drawn between a pair of scientists if they coauthored one or more articles. Collaborating multiple times does not make a difference in this model. Therefore, the social interaction profiles of the social actors constitute the adjacent matrix of the social network. Many previous studies on social networks use this type of representation(Freeman 1977; Wilkinson & Huberman 2004). More formally, the SIW function is defined as:

$$SIW_{01}(v_{i_1}, v_{i_2}) = \begin{cases} 1 & \text{if } e(v_{i_1}, v_{i_2}) \in E; \\ 0 & \text{else.} \end{cases} \tag{15}$$

Real-SIP However, one of the disadvantage of *01-SIP* is that the social interaction profiles give no consideration to the nodes other than their direct neighbors and fall short of considering the frequency of collaboration. In order to mitigate this problem, we propose a *Real-SIP* model which takes a node's neighbors' neighbors into consideration. This way, the social interaction profiles reflect the proximity of the nodes in the network more accurately. Furthermore, the final matrix defined by the social interaction profiles are less sparse which can improve the performance of the LDA model(Si & Jin 2005). In this model, we distinguish a node's direct neighbors from its neighbors' neighbors by giving different weights to them. The SIW function for a node is defined as follows:

$$SIW_R(v_{i_1}, v_{i_2}) = \begin{cases} p_{i_n,i_2} * c_{i_1,i_n} & \text{if } (e(v_{i_1}, v_{i_n}) \in E) \\ & \&\&(e(v_{i_n}, v_{i_2}) \in E) \\ & \&\&(e(v_{i_1}, v_{i_2}) \notin E); \\ c_{i_1,i_2} & \text{if } e(v_{i_1}, v_{i_2}) \in E; \\ 0 & \text{else.} \end{cases} \tag{16}$$

where c_{i_1,i_2} is the frequency of the collaboration between researchers v_{i_1} and v_{i_2}; and p_{i_n,i_2}, a normalized discount coefficient, is defined as

$$p_{i_n,i_2} = \frac{c_{i_n,i_2}}{\sum_j ci_n, i_j}$$

Therefore, *Real-SIP* encoding approach takes into account not only the strength of the collaboration, but also the second-order neighbors.

Experimental Settings and Evaluation

In evaluating the model and different SIP construction approaches, we first build up SIP in the two different ways for the researchers in the two networks. And then, 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. In perplexity computation, i is set as 300 in order to shorten the computation time.

We evaluate this model in both descriptive and quantitative ways: first, we demonstrate the exemplary communities discovered by the algorithms and briefly discuss the

Table 2: An illustration of 4 communities from a 50-community solution for the *CiteSeer* dataset after 1000 iterations based on $Real - SIP$ approach. Each community is shown with the 10 researchers and the corresponding $\mu_{m,n}$(magnitude is 10^{-2}) that have the highest probability conditioned on that topic

Community 8	Community 19
H. V Jagadish(2.2)	Sebastian Thrun(2.3)
Rakesh Agrawal(2.0)	Tom Mitchell(1.7)
Christos Faloutsos(1.9)	Manuela Veloso(1.7)
David Lomet(1.8)	Reid Simmons(1.6)
Divesh Srivastava(1.7)	Takeo Kanade(1.6)
Krithi Ramamritham(1.7)	Dieter Fox(1.5)
Sharad Mehrotra(1.6)	Milind Tambe(1.5)
Kenneth A Ross(1.6)	Frank Dellaert(1.4)
Serge Abiteboul(1.5)	Peter Stone(1.3)
Jiawei Han(1.5)	Andrew Moore(1.2)
Community 22	Community 36
David Culler(2.1)	Alex Waibel(2.6)
Eric Brewer(2.0)	Nelson Morgan(1.9)
Ion Stoica(1.9)	Lynette Hirschman(1.8)
Garth Gibson(1.8)	John Lafferty(1.8)
Y.H Katz(1.8)	Stanley Osher(1.6)
Steven Gribble(1.8)	M. J. Irwin(1.6)
David Patternson(1.7)	Lori Levin(1.5)
Scott Shenker(1.7)	Robert Frederking(1.4)
Srinivasan Seshan(1.7)	Jie Yang(1.4)
Amin Vahdat(1.7)	R. G. Mamahon(1.2)

results. Therefore, we compare the perplexity values for a set of community numbers for three different SIP encoding approaches. Furthermore, we investigate the quality of the discovered communities from a clustering perspective.

Examples of Communities

Table 2 shows 4 exemplary communities from a 50-community solution for the *CiteSeer* dataset with social interaction profiles being created using *Real-SIP* representation. Each community is shown with the top 10 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.

These exemplary communities give us some flavor on the communities that can be discovered by this approach. Specifically, we observe that some communities are “institution-based”, some others are “topic-based’. For instance, community 22 is clearly a Berkeley community although their research interest span a variety of areas. and most researchers in community 36 are from CMU. The second type of community is“topic-based”, as illustrated by Community 8 and 19, where most researchers in these two communities fall into database and AI domains respectively; Note that these two types of communities are not exclusive, meaning that many communities are actually “hybrid”, with some members being from the same institutions and others work on the same area. This observation reveals the fact that researchers from same institution or with similar research interest tend to collaborate together more and build closer social ties.

Table 3: Perplexity Results on *CiteSeer* after 300 iterations with the two *SIP* approaches

SIP	K=20	K=30	K=50
0-1	19473.3	16812.9	9752.9
Real	8763.1	7989.7	6289.3

Perplexity Analysis

Perplexity is is a common criterion for measuring the performance of statistical models in information theory. It indicates the uncertainty in predicting the occurrence of a particular social interaction given the parameter settings, and hence it reflects the ability of a model to generalize unseen data.

Perplexity PP is defined as

$$PP(\tilde{W}) = \prod_{m=1}^{M} p(\vec{\omega_m})^{-\frac{1}{N_m}} \quad (17)$$

$$= exp^{-\frac{\Sigma_{m=1}^{M} logp(\vec{\omega_m})}{\Sigma_{m=1}^{M} N_m}} \quad (18)$$

where $\vec{\omega_i}$ is the social interaction profiles in the test set and

$$p(\vec{\omega_i}) = \prod_{j=1}^{N_i} \Sigma_{k=1}^{K} p(\omega_{i,j}|\iota_n = k) p(\iota_n = k|d = m) \quad (19)$$

Note that the $p(\omega_{i,j}|\iota_n = k)$ can be determined by the training set, but hyperparameter $p(\iota_n = k|d = m)$ for the unseen documents in the test sets has to be estimated. The estimation can be achieved by conducting another Gibbs sampling process.

Table 3 lists the perplexity results for a selected set of topic numbers for the two different representation approaches. It shows that the perplexity value is high initially and decreases when the number of communities increases. In addition, the results show that the *Real-SIP* approach has lower perplexity value than the *0-1*, which implies that *Real-SIP* leads to more accurate prediction.

Clustering Analysis

In this section, we evaluate the quality of the communities discovered by *GWN-LDA* by comparing their compactness. Compactness of a community is measured through the average shortest distance among the top-ranked N_r researchers in this community. Short average distance indicates a compact community. In particular, N_r is set as 1000 in this paper. Both *CiteSeer* and *NanoSCI* have more than $200,000$ nodes in the network. In order to reduce the computational complexity and memory usage in calculating the shortest distances among the researchers, we pre-process the two networks by conducting a graph reduction algorithm to reduce the number of the nodes in the network. In

Table 4: Compactness Results for *CiteSeer* and *NanoSCI* with the two *SIP* approaches

SIP	CiteSeer(K=50)		NanoSCI(K=50)	
	mean	deviation	mean	deviation
0-1	5.82	1.52	4.30	1.04
Real	4.74	1.69	3.57	0.98

this graph-reduction algorithm, we iteratively eliminate the nodes whose degree is 1 (i.e, only one co-author). Subsequently, we run *Johnson*'s algorithm for calculating all-pair shortest paths for the processed networks. Since we focus on the top ranked researchers, this preprocessing has minimal impact on concerned researchers.

Table 4 demonstrates the compactness and well-separateness measures for *01-SIP* and *Real-SIP* approaches for datasets *CiteSeer* and *NanoSCI* respectively. In this experiment, the number of communities K is set as 50. The *t*-test results show that the *Real-SIP* approach is significantly better the other two approaches for both datasets.

Conclusions

Community discovery has drawn significant research interest among researchers for its increasing applications in a wide range of areas, including computer science, biology, social science and so on. This paper describes an LDA(latent Dirichlet Allocation)-based hierarchical Bayesian algorithm, namely *GWN-LDA*, to discover community structure from large-scale complex networks. In this model, communities are modeled as latent variables in the pertaining graphical model and defined as distributions over social actor space. This model is evaluated on two research collaborative networks:*CiteSeer* and *NanoSCI*. The experimental results demonstrate that this approach is promising for discovering community structures in large-scale networks. While this approach is developed and evaluated in social network domain, the model is generic and can be naturally extended to other complex network research areas.

Acknowledgments

The authors would like to thank Jian Huang and Dr. Lee Giles for providing *CiteSeer* dataset. The authors would also thank NSF and Dr. Padma Raghavan for kindly providing the computational resources for this project.

References

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

Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E* 70:066111.

Flake, G. W.; Lawrence, S.; and Giles, C. L. 2000. Efficient identification of web communities. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 150–160. New York, NY, USA: ACM Press.

Freeman, L. 1977. A set of measures of centrality based upon betweeness. In *Sociometry*, 35–41.

Gelman, A.; Carlin, J. B.; Rubin, D. B.; and Stern, H. S. 2004. *Bayesian Data Analysis, Secone Edition*. Chapman Hall Llc.

Girvan, M., and Newman, M. E. 2002. Community structure in social and biological networks. *Proc Natl Acad Sci U S A* 99(12):7821–7826.

Kostoff, R. N.; Stump1, J. A.; Johnson1, D.; Murday, J. S.; Lau, C. G.; and Tolles, W. M. 2006. The structure and infrastructure of the global nanotechnology literature.

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

Newman, M. E. 2004a. Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci U S A* 101 Suppl 1:5200–5205.

Newman, M. E. J. 2004b. Fast algorithm for detecting community structure in networks. *Physical Review E* 69:066133.

Palla, G.; Derenyi, I.; Farkas, I.; and Vicsek, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435:814.

Ruan, J., and Zhang, W. 2006. Identification and evaluation of weak community structures in networks. In *AAAI*.

Scott, J. P. 2000. *Social Network Analysis: A Handbook*. SAGE Publications.

Si, L., and Jin, R. 2005. Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis. In *PAKDD*, 622–631.

Sudderth, E. B.; Torralba, A.; Freeman, W. T.; and Willsky, A. S. 2005. Learning hierarchical models of scenes, objects, and parts. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, 1331–1338. Washington, DC, USA: IEEE Computer Society.

Wilkinson, D. M., and Huberman, B. A. 2004. A method for finding communities of related genes. *Proc Natl Acad Sci U S A* 101 Suppl 1:5241–5248.

Zhang, H.; Croft, W. B.; Levine, B.; and Lesser, V. 2004. A multi-agent approach for peer-to-peer based information retrieval systems. In *Proceedings of Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

Zhang, H.; Qiu, B.; Giles, C. L.; Foley, H. C.; and Yen, J. 2007. An lda-based community structure discovery approach for large-scale social networks. In *IEEE International Conference on Intelligence and Security Informatics*.

Zhou, D.; Ji, X.; Zha, H.; and Giles, C. L. 2006a. Topic evolution and social interactions: how authors effect research. In *CIKM*, 248–257.

Zhou, D.; Manavoglu, E.; Li, J.; Giles, C. L.; and Zha, H. 2006b. Probabilistic models for discovering e-communities. In *WWW*, 173–182.