An LDA-based Community Structure Discovery Approach for Large-Scale Social Networks

Haizheng Zhang*, Baojun Qiu†, C. Lee Giles*, Henry C. Foley* and John Yen*
*College of Information Science and Technology
Pennsylvania State University, University Park, PA 16802
Email: {hzhang,giles,hfoley,jyen}@ist.psu.edu
†Department of Computer Science and Engineering
Pennsylvania State University, University Park, PA 16802
Email: {bqiu}@cse.psu.edu

***Abstract*— Community discovery has drawn significant research interests among researchers from many disciplines for its increasing application in multiple, disparate areas, including computer science, biology, social science and so on. This paper describes an LDA(latent Dirichlet Allocation)-based hierarchical Bayesian algorithm, namely *SSN-LDA*(Simple Social Network LDA). In *SSN-LDA*, communities are modeled as latent variables in the graphical model and defined as distributions over the social actor space. The advantage of *SSN-LDA* is that it only requires topological information as input. 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.**[1]

I. Introduction

Social networks have been studied for decades. In recent years, this line of research has drawn even more attentions with the prevalence of social network websites, such as *MySpace, LiveJournal, Friendster*. These social networks are being used by millions and have gained increasing popularity among very diverse user groups. 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, 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. 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 network based applications exist in many disciplines, community discovery study is appealing to not only computer scientists, but also researchers from disparate areas such as biology, social science and so on. A wide array of approaches have been developed over years for finding communities and will be introduced in Section II.

Unlike those previous community discovery studies, we design a hierarchical Bayesian network based approach, namely *SSN-LDA*(Simple Social Network-LDA) to discover probabilistic communities from social networks. This model is inspired by the success of the application of LDA(Latent Dirichlet Allocation) models in the information retrieval and image analysis domains. In this model, communities are modeled as latent variables and are considered as distributions on the entire social actor space. This way, each social actor contributes a part, big or small, to every community in the society. We also propose three different approaches to create social interaction profiles based on the social interaction information in the network. The latent probabilistic model and three pertaining representation 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. While this approach is proposed in the social network domain and evaluated in the context of co-authorship networks, it can shed light on a broad set of complex network-based applications, including protein interaction, gene co-occurrence graph[1], and Web etc.

In conclusion, the contributions of this paper include: (1) an LDA-based probabilistic community discovery model in large-scale networks which only only requires the topological structure of networks; (2) the exploration of the impact of three different social interaction profiles representation approaches on the community discovery.

The rest of this paper is organized as follows: Section II introduces related studies; Section III presents *SSN-LDA* and its corresponding Gibbs sampler. Section IV describes the two co-authorship networks and three different representation approaches. Experimental results are demonstrated and analyzed in Section V. Section VI discusses some issues related to this model and some potential applications. Section VII concludes the paper and discusses some possible directions for future work.

II. Background and Related Works

This section introduces the background of this study and describes a series of related work, ranging from graph partition, community discovery, clustering algorithms, and several variants of LDA models.

[1]To appear in IEEE Intelligence and Security Informatics 2007

A. Community discovery

Community structures exist in different types of networks including Web communities[2], [3], social networks[4], [5], [6], [7], [8], [9], [10], co-authorship networks[11], [12], [13], and biological networks[5], [8], [1]. The most representative approaches among these related studies include:

(1) *Centrality indices* or *betweenness* based approaches. The *betweenness* concept was introduced by Freeman[14] as a centrality measure. It is defined on a vertex v_i as the number of shortest paths between pairs of other vertices that contain vertex v_i. This measure has been used in many previous studies on co-authorship network[5], [1], [13]. Girvan et al extended this measure to edges and designed a clustering algorithm which gradually remove the edges with highest betweenness value[5]. A similar approach was taken to find community structures in gene networks by Wilkinson et al [1], where gene networks were created by collecting gene co-occurrence information from the literature and partitioning it into communities of related genes. However, a major problem with this approach is that the complexity of this approach is $O(m^2n)$, where m is the number of edges in the graph and n is the number of vertices in the network.

(2)*Minimum cut approaches*. The community discovery problem can also be viewed as a graph partition problem which has broad application in circuit design, web community discovery, and among others. The graph partition problem can be formulated as the balanced minimum cut problem where the goal is to find an optimal graph partition so that the edge weight between the partitions is minimized while maintaining partitions of a minimal size. The NP-complete complexity of this approach[15] requires approximate solutions. Flake et al developed approximate algorithms to partition the network by solving *s-t* maximum flow techniques[2], [3]. The main idea behind maximum flow is to create clusters that have small inter-cluster cuts and relatively large intra-cluster cuts. This idea was first used to explore the Web structure in order to provide guidance for crawlers to identify the authoritative nodes (sinks) and hubs etc[2].

The major difference between *SSN-LDA* approach and the aforementioned approaches is that *SSN-LDA* is a mixture-model based probabilistic approach. Each community weighs in the contributions from every social actors and this property can be exploited in many potential applications that will be introduced in Section VI. With appropriate statistical models (such as Gibbs sampling process), the computation complexity for *SSN-LDA* is advantageous to the previous introduced models[16]. Specifically, the complexity of each iteration of the Gibbs sampling process is $O(KM)$, where K is the number of the communities, M is the number of the social interactions(edges) in the network.

B. Topic-based Community Discovery and related LDA Models

LDA model was first introduced by Blei for modeling the generative process of a document corpus[17]. Its ability of modeling topics using latent variables has attracted significant interests and it has been applied to many domains such as document modeling [17], text classification [17], collaborative filtering [17], image processing[18], information retrieval [16], topic models detection[19], [20], [21], and semantic based community discovery[10]. For more information about LDA model, readers can refer to a technical report[22] where the models is described in great details with elaboration on the corresponding Gibbs samplers.

Among these variants of LDA models, the approaches proposed in [20], [10] are both concerned about the authors of the documents in the corpus. In particular, Zhou et al introduced a community latent variable in their graphical model and applied it to discover community information embedded in document corpus. This approach can discover the underlying social network based on social interactions and topical similarity. In their follow-up work[23], Zhou et al. investigated how research topics evolve over time and attempted to discover the most influential researchers involved in such transitions. However, the most significant difference between our approach and these approaches lies in the fact that the only input information in this paper is the topological structure of a social network instead of semantic information. *SSN-LDA* encodes the structural information of networks into profiles and discovers community structures purely from these social connections among the nodes. Therefore we claim that it is more generic and can be applied to any complex network based applications.

III. LDA based Mixture Model for Social Networks

This section describes the *SSN-LDA* model. Before diving into the details, we first introduce related terminology and notations in Section III-A. Thereafter, Section III-B describes the *SSN-LDA* model. Finally, the Gibbs sampler for solving *SSN-LDA* model is presented in Section III-C.

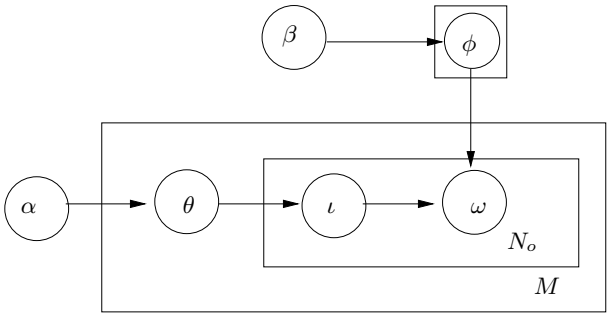

Fig. 1. Graphical Model for *SSN-LDA*

A. Terminology

A typical social network G 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: $SIW : (V \times V) \rightarrow \mathbf{I}$. 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

TABLE I
NOTATION FOR QUANTITIES IN *SSN-LDA*

M	number of social actors(social interaction profiles) in the social network
K	number of communities / mixture components
N_i	number of social interactions in a social interaction profile SIP_i
$\vec{\alpha}$	Dirichlet prior hyperparameter(known) on the mixing proportion
$\vec{\beta}$	Dirichlet prior hyperparameter(known) on the mixture component distributions for *SSN-LDA*
ι	hidden community variable, $\iota_{i,j}$ community for the jth social interaction in sip_i
$\vec{\theta}$	$p(\iota\|sip_j)$ the community mixture proportion for SIP_j
$\mathbf{\Theta}$	$\{\vec{\theta_m}\}_{m=1}^{M}$
$\vec{\phi_k}$	$p(\omega\|\iota_k)$ the mixture component of community k in *SSN-LDA*
$\mathbf{\Phi}$	$\{\vec{\phi_k}_{k=1}^{K}\}$ estimated parameter set for community mixture in *SSN-LDA*
ω	social interaction variable, $\omega_{i,j}$ means the j_{th} social interaction in SIP_i

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. In reality, these social interactions can be co-authorship, adviser/advisee, attendants of conferences, friendship and so on. In this paper, terms *vertex* and *social actor*, *edge* and *social interaction* are used interchangeably.

In this paper, a node v_i's neighboring agents are encoded by the variable $\vec{\omega_i}$ and ω_{ij} 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,

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

where m_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[17].

Subsequently, we specify that a social network contains a set of communities $\iota(\iota_1, \iota_2, ..., \iota_k)$ and each *community* in ι is defined as a distribution on the social actor space. In *SSN-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 following sections describe *SSN-LDA* model in more details.

B. Simple SN-LDA model(SSN-LDA)

The *SSN-LDA* model for social network analysis is illustrated in Fig. 1. Note that *SSN-LDA* resembles topic-based LDA model[17], with the social network being analogous to the corpus, the social interaction profiles being analogous to documents; and the occurrence of social interactions being analogous to words. The notations for all the variables in Fig. 1 is listed in table I. In particular, N_o is the number of social interactions in the pertaining social interaction profile. The distribution of topics in documents and the terms over topics are two multinomial distributions with two Dirichlet priors, whose hyperparameters are $\vec{\alpha}$ and $\vec{\beta}$ respectively. The dimensionality K of the Dirichlet distribution, which is also the number of community component distributions, is assumed to be known and fixed.

This generative process for an agent(ω_i)'s social interaction profile sip_i in a social network is:

1) Sample mixture components $\vec{\phi_k} \sim Dir(\vec{\beta})$ for $k \in [1, K]$
2) Choose $\vec{\theta_i} \sim Dir(\vec{\alpha})$
3) Choose $N_i \sim Poisson(\xi)$ (note that Poisson assumption is not critical to this model)
4) For each of the N_i social interactions ω_{ij}:
   (a) Choose a community $\iota_{ij} \sim Multinomial(\vec{\theta_i})$;
   (b) Choose a social interaction $\omega_{i,j} \sim Multinomial(\vec{\phi_{\iota_{i,j}}})$

According to the model, the probability that the j_{th} social interaction element $\omega_{i,j}$ in the social actor ω_i's social interaction profile sip_i instantiates a particular neighboring agent ω_m is:

$$p(\omega_{i,j} = \omega_m|\vec{\theta_i}, \underline{\mathbf{\Phi}}) = \sum_{k=1}^{K} p(\omega_{i.j} = v_m|\vec{\phi_k})p(\iota_{i,j} = k|\vec{\theta_i})$$

where $\vec{\theta_i}$ is the mixing proportion variable for sip_i and $\vec{\phi_k}$ is the parameter set for the k_{th} community component distribution.

Given the hyperparameters $\vec{\alpha}$ and $\vec{\beta}$, the joint distribution of all known and hidden variables is:

$$p(\vec{\omega_i}, \vec{\iota_i}, \vec{\theta_i}, \underline{\mathbf{\Phi}}|\vec{\alpha}, \vec{\beta}) = \prod_{j=1}^{N_i} p(\omega_{ij}|\vec{\phi_{\iota_{i,j}}})p(\iota_{i,j}|\vec{\theta_i})p(\vec{\theta_i}|\vec{\alpha})p(\underline{\mathbf{\Phi}}|\vec{\beta})$$

Exact inference is generally intractable for LDA model. There have been three major approaches for solving this model approximately, including variational expectation maximization [17], expectation propagation [24], and Gibbs sampling[25], [26], [22]. Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) simulation[27] where the dimension K of the distribution are sampled alternately one at a time, conditioned on the values of all other dimensions[22]. We select

this approach to solve *SSN-LDA* models because it often yields relatively simple algorithms for approximate inference in high-dimensional models. Section III-C gives further description on the Gibbs sampler that is used in this paper.

C. Gibbs Samplers for SSN-LDA

In *SSN-LDA*, the desired distribution is the posterior given evidence $p(\iota|\mathbf{w})$.

$$p(\iota|\omega) = \frac{p(\omega,\iota)}{\sum_{\iota} p(\omega,\iota)} \tag{1}$$

However, the computation complexity of the the denominator part is prohibitively high. In this section, we apply the Gibbs sampling algorithm that has been introduced in [26], [22] to solve the *SSN-LDA* model and reduce the computation requirement. The algorithm for *SSN-LDA* is listed in Algorithm 1.

Specifically, the joint distribution of *SSN-LDA* can be factored as:

$$p(\vec{\omega},\vec{\iota}|\vec{\alpha},\vec{\beta}) = p(\vec{\omega}|\vec{\iota},\vec{\beta})p(\vec{\iota}|\vec{\alpha}) \tag{2}$$

$$= \prod_{\iota=1}^{K} \frac{\Delta(\vec{n_{\iota}}+\vec{\beta})}{\Delta(\vec{\beta})} \prod_{m=1}^{M} \frac{\Delta(\vec{n_m}+\vec{\alpha})}{\Delta(\vec{\alpha})} \tag{3}$$

Subsequently, the update equation for the hidden variable can be derived [22]:

$$P(\iota_i = j|\vec{\iota}_{\neg i},\vec{w}) \propto \tag{4}$$

$$\frac{n^{w_i}_{\neg i,j}+\beta}{n^{(.)}_{\neg i,j}+W\beta} * \frac{n^{(sip_i)}_{\neg i,j}+\alpha}{n^{(sip_i)}_{\neg i}+T\alpha} \tag{5}$$

where $n^{(.)}_{\neg i}$ is the count that does not include the current assignment of ι_i and recall that sip is the variable for social interaction profiles. For the sake of simplicity, we assume that the Dirichlet distribution is symmetric in deriving the above formula.

Finally, the update formula for $\phi_{k,\omega}$ and $\theta_{m,k}$ are as follows:

$$\phi_{k,\omega} = \frac{n^{(\omega)}_k+\beta}{\sum_{v=1}^{V} n^{(v)}_k + W\beta} \tag{6}$$

$$\theta_{m,k} = \frac{n^{(k)}_m+\alpha}{\sum_{\iota=1}^{K} n^{(\iota)} + T\alpha} \tag{7}$$

The detailed algorithm is listed in Algorithm 1.

```
/* Initialization                                          */
foreach Social Interaction Profile sip_i ∈ [1, M] do
    foreach Social Interaction ω_{i,j} ∈ [1, N_i] do
        sample topic index ι_{i,j} ~ Mult(1/K);
        update counters: n_i^{(ι_{i,j})} + 1  n_i + 1, n_{ω_{i,j}}^{(ι_{i,j})} + 1,
        n_{ι_{i,j}} + 1;
    end
end
/* Gibbs sampling over burn-in period
   and sampling period                                     */
while not finished do
    foreach SIP sip_i do
        foreach ω_{i,j} ∈ [1, N_i] do
            decrement counts and sums: n_i^{(ι_{i,j})} − 1,
            n_i − 1, n_{ι_{i,j}}^{(ω_{i,j})} − 1, n_{ω_{i.j}} − 1;
            resample ω_{i,j} according to equation 4;
            update the counters accordingly;
        end
    end
    /* Check convergence and read out
       parameters                                          */
    if converged and L iterations then
        updated parameters φ and θ and readout
        parameters;
    end
end
```

Algorithm 1: Gibbs sampling algorithm for *SSN-LDA*

IV. Co-authorship networks and pertaining representation approaches

We evaluate the *SSN-LDA* approach in the context of research collaboration networks. This section describes the two co-authorship networks used in this paper as well as three different approaches of creating the corresponding social interaction profiles.

A. 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 *SSN-LDA* model on co-authorship networks collected from two distinct areas: computer science(*CiteSeer*) and nanotechnology(*NanoSCI*). Note that no name disambiguation has been done on either dataset.

1) CiteSeer *Dataset:* 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.

2) 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 [28]. The essential idea of this approach is to augment the keyword set until the returned results converges.

Table. II 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 subnetwork 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 subnetworks. Therefore, unless specially specify, we always mean the two subnetworks when referring *CiteSeer* and *NanoSCI*.

B. Social Interaction Profile Representations

The social interaction profiles of the social actors collectively determines the structure and dynamics of a social network. In this paper, we explore three different types of social interaction profile representations for social networks, namely *01-SIP*, *012-SIP*, and *k-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 schemes.

1) 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[14], [1]. More formally, the SIW function is defined as:

$$SIW_{01-SIP}(v_{i_1}, v_{i_2}) = \begin{cases} 1 & \text{if } e(v_{i_1}, v_{i_2}) \in E; \\ 0 & \text{else.} \end{cases} \quad (8)$$

2) 012-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. In order to mitigate this problem, we propose a *012-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 sparser which can improve the performance of the LDA model[29]. 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:

$$SIW_{012-SIP}((v_{i_1}, v_{i_2}) = \begin{cases} 1 & \text{if } (e(v_{i_1}, v_{i_n}) \in E) \\ & \text{AND } (e(v_{i_n}, v_{i_2}) \in E) \\ & \text{AND } (\ e(v_{i_1}, v_{i_2}) \notin E); \\ 2 & \text{if } e(v_{i_1}, v_{i_2}) \in E; \\ 0 & \text{else.} \end{cases} \quad (9)$$

3) k-SIP: The two approaches of defining social interaction profiles fall short of considering the frequency of collaboration. This section describes a *K-SIP* model where the weight information for an edge is defined as the times of the collaboration between the two authors. That is, $SIW_{k-SIP}(v_{i_1}, v_{1_2}) = k$ iff researcher v_{i_1} and researcher v_{i_2} has coauthored for k times in the past. This way, the SIW function reflects the strength of the interactions.

V. Experimental Settings and Evaluation

In evaluating the model and different SIP construction approaches, we first build up SIP in the three 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. In both case, α is set as $\frac{1}{K}$ and β is set as 0.01, where K is the number of the communities.

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

A. Examples of Communities

Table III shows 6 exemplary communities from a 50-community solution for the *CiteSeer* dataset with social interaction profiles being created using *012-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, 6 out of 10 researchers (Don Towsley, James F. Kurose, Victor Lesser, Prashant Shenoy, Jim Kurose, Paul R Cohen) in Community 15 listed in Fig III are from University of Massachusetts, Amherst, although they work in disparate areas spanning from networking, knowledge management, operating systems and multi-agent research; Similarly, community 29 is clearly a Berkeley community and most researchers in community 43

TABLE II

STATISTICS FOR DATASETS *CiteSeer* AND *NanoSCI*

Dataset	Size	Paper number	Edge number	average author number per paper	size of largest component
CiteSeer	398831	716793	1181133	1.648	249866
NanoSCI	225313	195997	877609	4.48	203762

TABLE III

AN ILLUSTRATION OF 6 COMMUNITIES FROM A 50-COMMUNITY SOLUTION FOR THE *CiteSeer* DATASET AFTER 1000 ITERATIONS BASED ON $012 - SIP$ APPROACH. EACH COMMUNITY IS SHOWN WITH THE 10 RESEARCHERS THAT HAVE THE HIGHEST PROBABILITY CONDITIONED ON THAT TOPIC

Community 15	Community 26	Community 12
John A Atankovic	Manuela Veloso	Jiawei Han
Don Towsley	Peter Stone	Dragomir R. Radev
Krithi Ramamritham	Milind Tambe	Senior Member
James F. Kurose	Andrew Barto	Kathleen R. Mckeown
Victor Lesser	Minoru Asada	Shih Fu Chang
Prashant Shenoy	Xuemei Wang	Terrence J. Sejnowski
Jean Yves le Boudec	Hiroaki Kitano	Ke Wang
Jim Kurose	Thomas G. Dietterich	Hongjun Lu
Subhabrata Sen	Craig A. Knoblock	Beng Chin Ooi
Paul R Cohen	Itsuki Noda	Thomas S. Huang
Community 29	**Community 43**	**Community 47**
David E.Culler	Alex Waibel	Geoffrey Fox
Eric A Brewer	Alon Lavie	Ken Kennedy
Y.H Katz	Jaime Carbonell	Alok Choudhary
Ion Stoica	Masaru Tomita	Cisco Systems
Hari Balakrishnan	Stanley Osher	Deborah Estrin
Steven D. Gribble	M. J. Irwin	Andrew Chien
David A. Patternson	Lori Levin	Sanjay Ranka
Srinivasan Seshan	Robert Frederking	Scott Shenker
Randy H. Katz	Jie Yang	Charles Koelbel
Scott Shenker	R. G. Mamahon	Ian Foster

are from CMU. The second type of community is"topic-based", as illustrated by Community 16, where most researchers in this community fall into AI and machine learning research area; and most members in Community 12 are working in information retrieval and data mining areas. 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 interests tend to collaborate together more and build closer social ties.

B. 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.

TABLE IV

PERPLEXITY RESULTS ON *CiteSeer* AFTER 300 ITERATIONS WITH DIFFERENT *SIP* APPROACHES

SIP	T=20	T=30	T=50
0-1	17853.24	14582.90	8620.29
0-1-2	9435.13	7382.17	5696.51
0-1-k	16873.29	12648.33	7967.10

Perplexity PP is defined as

$$PP(\tilde{W}) = \prod_{m=1}^{M} p(\vec{\omega_m})^{-\frac{1}{N_m}} \quad (10)$$

$$= exp^{-\frac{\Sigma_{m=1}^{M} logp(\vec{\omega_m})}{\Sigma_{m=1}^{M} N_m}} \quad (11)$$

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

$$p(\vec{\omega_m}) = \prod_{n=1}^{N_d} \Sigma_{k=1}^{K} p(\omega_n = t|\iota_n = k)p(\iota_n = k|d = m)$$

$$= \prod_{v=1}^{V} \left(\Sigma_{k=1}^{K} \phi_{k,t} * \theta_{m,k}\right)^{n_m^{(v)}}$$

where $n_M^{(v)}$ is the number of times term t has been observed in document m. Note that the Φ can be determined by the training set, but hyperparameter Θ for the unseen documents in the test sets has to be estimated. The estimation can be achieved by conducting another Gibbs sampling process[22]:

$$p(\tilde{\iota}_i = k|\vec{\tilde{\iota}_{\neg i}}, \tilde{\vec{\omega}}, \vec{\iota_{\neg i}}, \vec{\omega}) = \frac{n_k^{(t)} + \tilde{n}_{k,\neg i}^{(t)} + \beta}{n_k^{(.)} + \tilde{n}_{,}^{(.)} W\beta} \frac{n_{\tilde{m}+\alpha}^{k,\neg i}}{n_{\tilde{m}}^{(.)} T\alpha}$$

N_m is the size of the social interaction profile in the test set.

And then we have

$$\theta_{\tilde{m},k} = \frac{n_{\tilde{m}}^{(k)} + \alpha}{n_{\tilde{m}}^{(.)} + T\alpha} \quad (12)$$

Table IV lists the perplexity results for a selected set of topic numbers for the three 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 *012-SIP* approach has lower perplexity value than the other two approaches.

C. Clustering Analysis

In this section, we evaluate the quality of the communities discovered by *SSN-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 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 preprosessing has minimal impact on concerned researchers.

Figures 2 and 3 demonstrate the compactness and well-separateness measures for *01-SIP*, *012-SIP*, and *k-SIP* approaches for datasets *CiteSeer* and *NanoSCI* respectively. In particular, the two *x* axes in Figures 2 and 3 show the shortest distance and the two *y* axes show the numbers of top-ranked author pairs with the corresponding shortest distance. Note that the two authors in the author pair are within the same communities. In Figure 2, The mean for *01-SIP* approach is 5.62, with standard deviation as 1.58; the mean for *012-SIP* approach is 4.63, with standard deviation being 1.49. The mean for *k-SIP* approach is 5.10, with the standard deviation being 1.36. In Figure 3, the mean for *01-SIP* is 4.097 with standard deviation being 0.999; the mean for *012-SIP* is 2.34, and the corresponding standard deviation is 0.73; the mean for *k-SIP* approach is 3.62, with the standard deviation being 1.196. The *t*-test results show that the *012-SIP* approach is significantly better the other two approaches for both datasets.

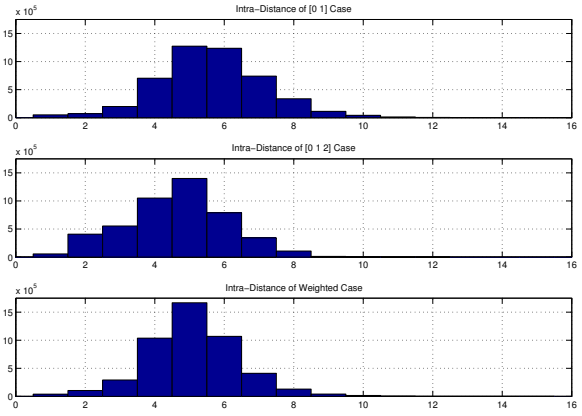

Fig. 2. The shortest distance (*x* axis) and the number of top-ranked researcher pairs from different communities with the corresponding distance(for dataset *CiteSeer*)

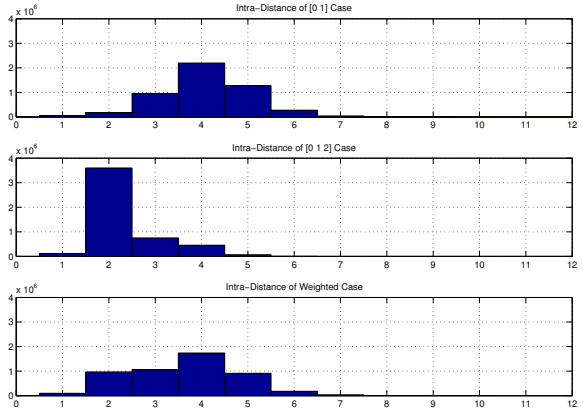

Fig. 3. The shortest distance (*x* axis) and the number of top-ranked researcher pairs from different communities with the corresponding distance(for dataset *NanoSCI*)

VI. Discussions

While the community discovery approach introduced in this paper is evaluated in the context of research collaborative networks, it has broad implications on social network research. This section lists three possible applications for the *SSN-LDA* model.

(1)*Detect the importance and roles of community members*

The probabilities that can be derived from *SSN-LDA* model can be helpful in determining the importance and roles of community members. For instance, the importance of community members conditioned on the community variable ι_j can be measured through the probability $p(\omega_i|\iota_j)$, which can be easily derived from this model based on the learned $\vec{\phi}$. In addition, $p(\iota_j|\omega_i)$ can reveal how strong a social actor is associated with a particular community. This is related to the work of locating "sinks" or "Hubs" or locating leaders in local communities.[30].

(2)*Measure the similarity between communities*

SSN-LDA model provides an elegant way to measure the similarity of two communities by calculating the corresponding KL(Kullback-Leibler) distance and convert it to similarity measure. KL divergence is a distance measure for two distributions and the corresponding formula for calculating the distance between two communities ι_i and ι_j is:

$$D_{KL}(\iota_i, \iota_j) = \Sigma_k p(\omega_k|\iota_i) log \frac{p(\omega_k|\iota_i)}{p(\omega_k|\iota_j)}$$

And then the similarity $Sim(\iota_i, \iota_j)$ between communities ι_i and ι_j can be derived by:

$$Sim(\iota_i, \iota_j) = 10^{-\zeta * D_{KL}(\iota_i, \iota_j)}$$

(3)*Identity recognition and name disambiguation*

Name disambiguation is very important to social network studies because (1) most of the current social network information is extracted from online and errors are inevitable. (2)many social actors may possess the same names although

they may share very different research interests and belong to different social communities. Conversely, a same person may be identified as multiple ones due to the confusion on middle name or maiden name. We believe that *SSN-LDA* model is able to provide some insights on whether two distinct individuals are actually the same person, or whether multiple members share the same name. For instance, in Community 15 in Fig. III, we have reasonable doubt that members *James F. Kurose* and *Jim Kurose* may be the same person. On the other hand, if a member belongs to very different communities, he may be a candidate deserving more attention for name disambiguation purpose.

VII. Conclusions and Future Work

Community discovery has drawn significant research interests among researchers from many disciplines for its increasing application in multiple, disparate areas, including computer science, biology, social science and so on. This paper describes an LDA(latent Dirichlet Allocation)-based hierarchical Bayesian algorithm, namely *SSN-LDA*(Simple Social Network LDA). In *SSN-LDA*, communities are modeled as latent variables in the graphical models and defined as distributions over social actor space. The advantage of *SSN-LDA* is that it only requires topological information as input. 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 fairly generic and can be naturally extended to other complex network research area including protein interaction recognition and can have broad implication on homeland security studies.

Acknowledgment

The authors would like to thank Wei Li, Xuerui Wang, and Ding Zhou for the helpful discussions on this paper. The authors would also thank NSF and Dr. Padma Raghavan for kindly providing the computational resources for this project.

References

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

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

[3] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis, "Graph clustering and minimum cut trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, 2004.

[4] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, p. 066111, 2004. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0408187

[5] M. Girvan and M. E. Newman, "Community structure in social and biological networks." *Proc Natl Acad Sci U S A*, vol. 99, no. 12, pp. 7821–7826, June 2002.

[6] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Tracking evolving communities in large linked networks." *Proc Natl Acad Sci U S A*, vol. 101 Suppl 1, pp. 5249–5253, April 2004.

[7] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, p. 066133, 2004. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0309508

[8] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, p. 814, 2005. [Online]. Available: doi:10.1038/nature03607

[9] J. P. Scott, *Social Network Analysis: A Handbook*. SAGE Publications, January 2000. [Online]. Available: http://www.amazon.co.uk/exec/obidos/ASIN/0761963391/citeulike-21

[10] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities." in *WWW*, 2006, pp. 173–182.

[11] K. Börner, J. T. Maru, and R. L. Goldstone, "The simultaneous evolution of author and paper networks," pp. 5266–5273, Apr 2004.

[12] M. E. Newman, "Coauthorship networks and patterns of scientific collaboration." *Proc Natl Acad Sci U S A*, vol. 101 Suppl 1, pp. 5200–5205, April 2004.

[13] T. Krichel and N. Bakkalbasi, "A social network analysis of research collaboration in the economics community," *Journal of Information Management and Scientometrics*, To Appear.

[14] L. Freeman, "A set of measures of centrality based upon betweeness," in *Sociometry*, 1977, pp. 35–41.

[15] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1979.

[16] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval." in *SIGIR*, 2006, pp. 178–185.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

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

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

[20] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*. Arlington, Virginia, United States: AUAI Press, 2004, pp. 487–494.

[21] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations." in *ICML*, 2006, pp. 577–584.

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

[23] D. Zhou, X. Ji, H. Zha, and C. L. Giles, "Topic evolution and social interactions: how authors effect research." in *CIKM*, 2006, pp. 248–257.

[24] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," 2002.

[25] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning." *Machine Learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[26] T. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences*, 2004.

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

[28] R. N. Kostoff, J. A. Stump1, D. Johnson1, J. S. Murday, C. G. Lau, and W. M. Tolles, "The structure and infrastructure of the global nanotechnology literature," pp. 301–321, 2006.

[29] L. Si and R. Jin, "Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis." in *PAKDD*, 2005, pp. 622–631.

[30] L. C. Freeman, T. J. Fararo, and W. B. J. M. H. Sunshine, "Locating leaders in local communities," in *American Sociological Review*, 1963, pp. 791–798.