A Topical Link Model for Community Discovery in Textual Interaction Graph

Guoqing Zheng[†], Jinwen Guo[†], Lichun Yang[†], Shengliang Xu[†],
Shenghua Bao[‡], Zhong Su[‡], Dingyi Han[†], Yong Yu[†]

[†]Shanghai Jiao Tong University
Shanghai, 200240, China
{gqzheng, guojw, lichunyang, slxu, handy, yyu}@apex.sjtu.edu.cn

[‡]IBM China Research Laboratory
Beijing, 100094, China
{baoshhua, suzhong}@cn.ibm.com

ABSTRACT

This paper is concerned with community discovery in textual interaction graph, where the links between entities are indicated by textual documents. Specifically, we propose a Topical Link Model(TLM), which leverages Hierarchical Dirichlet Process(HDP) to introduce hidden topical variable of the links. Other than the use of links, TLM can look into the documents on the links in detail to recover sound communities. Moreover, TLM is a nonparametric model, which is able to learn the number of communities from the data. Extensive experiments on two real world corpora show TLM outperforms two state-of-the-art baseline models, which verify the effectiveness of TLM in determining the proper number of communities and generating sound communities.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms, Experimentation

Keywords

Community discovery, Nonparametric statistical model, Topical link model

1. INTRODUCTION

Community discovery is one of the important research topics in multiple disciplines. Traditionally, it is performed on an entity link graph in which the vertices represent the entities and the edges indicate links between pairs of entities. Several methods have been proposed to discover communities in previous work. Most approaches, including graph cut based methods[6], modularity based methods[4], flow based methods[1] and spectral based methods[5], typically choose an objective function which captures the above intuition of a community and then try to optimize the objective function[2]. In [3], Mei et al. propose NetPLSA for discovering smoothing topics over network. These methods only model the links by assigning a certain weight to each link between pairs of entities, as the co-authorship network shown in Figure 1(a).

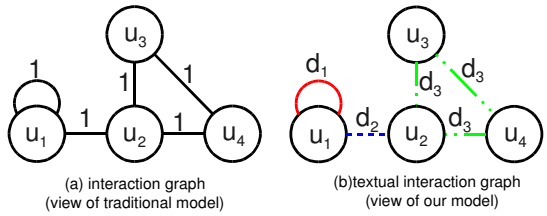

Figure 1: Difference between interaction graph and textual interaction graph

In particular, this paper explores community discovery in *textual interaction graph*. In our setting, the links between entities are indicated by text documents. We refer to such kind of data as *textual interaction graph*. Figure 1(b) gives a sample of a research proceeding corpus. In this paper, we propose a Topical Link Model(TLM). Given the textual interaction graph, TLM leverages Hierarchical Dirichlet Process(HDP) to introduce hidden topic variables of the links. Moreover, TLM can look into the documents on the links in detail to recover sound communities. We first generate the hidden topic variables of the links and then documents are generated by these according to the hidden variables. Besides, TLM is a nonparametric model. Experimental results on two real world corpora show the effectiveness of TLM in determining the proper number of communities and generating sound communities.

2. PROBLEM STATEMENT

Definition 1. **(Textual Interaction Graph)**: A textual interaction graph is viewed as $G_0 = (V_0, E_0)$ with associated documents $\mathcal{D}$, where V_0 represents the set of users (vertices), E_0 represents the set of interactions (edges) between users and $\mathcal{D}$ represents the set of documents that record those interactions. If vertex v_i and v_j have interacted once which is recorded by document $d_k \in \mathcal{D}$, there is an edge $e_{v_i,v_j} \in E_0$ to indicate this **interaction**, and the document representing this interaction is d_k. Figure 2(a) shows one example of interaction graph.

Definition 2. **(Topical Community)**: A topical community is a soft partition of the users with a multinomial

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

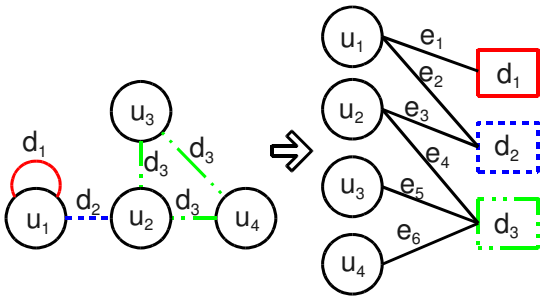

Figure 2: An interaction graph and its corresponding participation graph

word distribution over the vocabulary $\mathcal{V}$. We denote the topical community space as ϕ_∞.

Definition 3. **Task (Community Discovery)**: Given an participation graph G with interaction corpus $\mathcal{D}$, the task of *Community Discovery* is to find a set of topical communities $\{c_1, c_2, ..., c_K\}$, where the community number K is detected automatically and to calculate the community distribution of each user u, i.e. $\boldsymbol{\theta_u}$.

3. COMMUNITY DISCOVERY MODELING

3.1 Participation Graph

In this paper, in order to consider the content of the interactions, we transform G_0 to an equivalent form $G = (V, E)$ with corpus $\mathcal{D}$ where $V = V_0$, $\mathcal{D}$ is the set of interaction documents generated among the interactions of the users and each $e \in E$ represents a user's **participation** in a document. If vertex v_i and v_j have interacted once and the corresponding interaction document is d, there is an edge $e_{v_i,d}$ connecting v_i and d indicating that v_i participates in the interaction represented by d, and the same with v_j. Figure 2(b) shows the equivalent form of Figure 2(a).

3.2 Topical Link Model

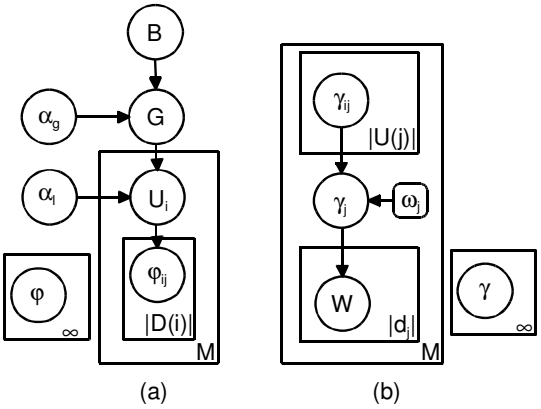

Figure 3: Separated graphical model of TLM. (a) models the participation graph, (b) models the documents.

Participation Graph Modeling. Figure 3(a) shows the graphical model of our Participation Graph Modeling, where M is the number of users and $|D(i)|$ is the number of documents linked to user i. The outside global Dirichlet Process(DP) provides a shared infinite number of variables

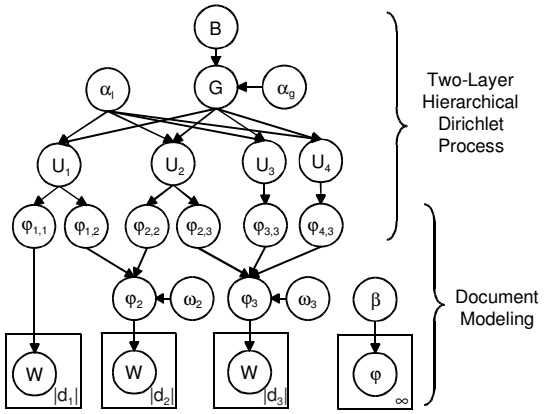

Figure 4: Combined graphical model of TLM for the sample graph in Figure 2 (b)

for all the users' evidence variables. Formally, the Dirichlet process is

$$G|\alpha_g, B \sim DP(\alpha_g, B) \tag{1}$$

The inside local DP models all participations of a single user in her documents. This models the clustering property of a user's participation in documents. Formally,

$$U|\alpha_l, G \sim DP(\alpha_l, G) \tag{2}$$

Then, every user's evidence variables are drawn from the infinite space ϕ_∞ as follows:

$$p(\phi_{i,j_{n+1}}|\phi_{i,j_1}, \cdots, \phi_{i,j_n}; \alpha_l) = \frac{\alpha_l G + \sum_{h=1}^{n} \delta(\phi_{i,j_h})}{n + \alpha_l} \tag{3}$$

where ϕ_{i,j_h} denotes the hth evidence variable of U_i, and δ is the atom function. This actually forms a two-layer HDP.

Document Modeling. Figure 3(b) shows the graphical model of our Document Modeling, where $|d_j|$ is the number of words in d_j and $|U(j)|$ denotes the number of topic variables connected to γ_j. On the top, γs are the topic variables of all the authors of the document, which are draws from the infinite space γ_∞. Then, ω is the weighting vector parameters for the topic selection process, satisfying $\sum_{h=1}^{|U(j)|} \omega_{j,h} = 1$. The relation of these variables is as

$$p(\gamma_j|\gamma_{(1),j}, \gamma_{(2),j}, \cdots, \gamma_{(|U_j|),j}) = \sum_{h=1}^{|U_j|} \omega_h \delta(\gamma_{(h),j}) \tag{4}$$

where $\gamma_{(h),j}$ denotes the hth evidence variable that connects γ_j. Then, we assume that all the words in that document are drawn from the topic model, as the lower part of the Figure 3(b). The joint probability of a document model γ_j and its generated words $\boldsymbol{W}_j$ is:

$$p(W_{j,1}, W_{j,2}, \cdots, W_{j,|\boldsymbol{W}_j|}, \gamma_j) = p(\gamma_j) \prod_{h=1}^{|\boldsymbol{W}_j|} p(W_{j,h}|\gamma_j) \tag{5}$$

where γ_j is a draw from an infinite semantic space γ_∞.

Combination of Participation Graph Modeling and Document Modeling. Considering the community space ϕ_∞ and the document topic space γ_∞ are both semantic spaces, we unite them into a single space, ϕ_∞. See Figure 4 for the complete model. Figure 4 just gives a sample according to the participation graph in Figure 2(b).

3.3 Model Inference

Table 1: symbols used in model inference

M	the number of users
N	the number of documents
$\boldsymbol{W}_j$	the jth observed document
$D_s(i)$	the set of single-user documents linked to user i
$D_m(i)$	the set of multi-user documents linked to user i
$U(j)$	the set of users linked to document j
$T(i)$	the set of tables in restaurant i
$t_{i,j}$	table index of customer (i,j), $i \in U(j)$, $j \in D(i)$
$k_{t_{i,j}}$	the dish serving on the table that $t_{i,j}$ refers
$k_{i,t}$	$t \in T(i)$, the dish on the tth table in restaurant i
d_j	the mixture component (i.e. dish) index of multi-user document j
$m_{i,k}$	number of tables in restaurant i serving dish k
$n_{i,t,k}$	the number of customers in restaurant i, sitting at table t, eating dish k

We employ Gibbs sampling for model inference.Firstly, we derive two likelihood expressions. The conditional density of $\boldsymbol{W}_j$ under mixture component k given all other observed documents is:

$$f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) \tag{6}$$
$$= \frac{\int Mul(\boldsymbol{W}_j|\phi_k)\prod_{j'\neq j,\begin{Bmatrix} d_{j'}=k, j'\in D_m(\cdot) \\ k_{t_{\cdot,j'}}=k, j'\in D_s(\cdot)\end{Bmatrix}} Mul(\boldsymbol{W}_{j'}|\phi_k)Dir(\phi_k)d\phi_k}{\int \prod_{j'\neq j,\begin{Bmatrix} d_{j'}=k, j'\in D_m(\cdot) \\ k_{t_{\cdot,j'}}=k, j'\in D_s(\cdot)\end{Bmatrix}} Mul(\boldsymbol{W}_{j'}|\phi_k)Dir(\phi_k)d\phi_k}$$

where $Mul()$ denotes multinomial distribution and the selection probability distribution is $s(d_j|k_{t_{\cdot,j}}) = \sum_{i'\in U(j)}\delta(k_{t_{i',j}})$ We have set all ω's to be uniform so that these parameters can be omitted. There are three sets of hidden variables to be sampled: table indices $\boldsymbol{t}$ of customers, mixture component indices $\boldsymbol{d}$ of documents, and dish indices $\boldsymbol{k}$ of tables.

Sampling $\boldsymbol{t}$. The variable set $\boldsymbol{t}$ should be split to two sets because they are different in sampling. Firstly, if the document j is a single user interaction document, $t_{i,j}$ is sampled by combing the likelihood of generating the observed documents.

$$p(t_{i,j}=t|\boldsymbol{t}^{-i,j},\boldsymbol{k}) \tag{7}$$
$$\propto \begin{cases} n_{it\cdot} f_{k_{i,t}}^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) & if\ t\ has\ been\ used \\ \alpha_l p(\boldsymbol{W}_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) & if\ t=t^{new}\end{cases}$$

where $p(\boldsymbol{W}_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k})$ is the liklihood for $t_{i,j}=t^{new}$:

$$p(\boldsymbol{W}_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) \tag{8}$$
$$= \sum_{k=1}^{K}\frac{m_{\cdot,k}}{m_{\cdot,\cdot}+\alpha_g} f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) + \frac{\alpha_g}{m_{\cdot,\cdot}+\alpha_g} f_{k^{new}}^{-\boldsymbol{W}_j}(\boldsymbol{W}_j)$$

If the sampled value of $t_{i,j}$ is t^{new}, we need to obtain a sample of $k_{i,t^{new}}$:

$$p(k_{i,t^{new}}|\boldsymbol{t},\boldsymbol{k}^{-i,t^{new}}) \propto \begin{cases} m_{\cdot,k} f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) & if\ k\ has\ been\ used \\ \alpha_g f_{k_{new}}^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) & if\ k=k^{new}\end{cases} \tag{9}$$

Secondly, for $\boldsymbol{t}$ of multi-user interaction documents, we have a similar sampling process but the likelihood function is replaced by the selection function.

$$p(t_{i,j}=t|\boldsymbol{t}^{-ij},\boldsymbol{k},\boldsymbol{d}) \tag{10}$$
$$\propto \begin{cases} n_{i,t,\cdot}^{-i,j} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k_{i,t}) & if\ t\ has\ been\ used \\ \alpha_l p(d_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) & if\ t=t^{new}\end{cases}$$

Table 2: Statistics of PAPER and NYT

	# of users	# of docs	# of links	# of links/user
PAPER	9415	5308	25034	2.7
NYT	1677	2461	83367	49.7

Table 3: Best CCD on NYT

	NCut	NetPLSA	TLM
NYT	173.9	180.35	141.82

where $p(d_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k})$ is:

$$p(d_j|\boldsymbol{t}^{-i,j}, t_{i,j}=t^{new},\boldsymbol{k}) = \tag{11}$$
$$\sum_{k=1}^{K}\frac{m_{\cdot,k}}{m_{\cdot,\cdot}+\alpha_g} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k) + \frac{\alpha_g}{m_{\cdot,\cdot}+\alpha_g} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k^{new})$$

And in the case of choosing t^{new}:

$$p(k_{i,t^{new}}|\boldsymbol{t},\boldsymbol{k}^{-i,t^{new}}) \tag{12}$$
$$\propto \begin{cases} m_{\cdot,k} s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k) & if\ k\ has\ been\ used \\ \alpha_g s(d_j|k_{t_{\cdot,j}}^{-i}, k_{t_{i,j}}=k^{new}) & if\ k=k^{new}\end{cases}$$

Sampling $\boldsymbol{d}$. These variables relate only to the selection processes and the multi-user document likelihood. They are thus sampled as

$$p(d_j=k|\boldsymbol{W}_\cdot,\boldsymbol{d}^{-j},\boldsymbol{t},\boldsymbol{k}) \propto s(d_j=k|k_{t_{\cdot,j}}) f_k^{-\boldsymbol{W}_j}(\boldsymbol{W}_j) \tag{13}$$

Sampling $\boldsymbol{k}$. Since changing $k_{i,t}$ will change the mixture components of all the $t_{i,\cdot}$, the sampling relates to both the selection likelihood and the document likelihood.

$$p(k_{i,t}=k|\boldsymbol{t},\boldsymbol{k}^{-it}) \tag{14}$$
$$\propto \begin{cases} m_{\cdot,k}^{-it} f_k^{-\boldsymbol{W}_{\{j,j\in D_s(i)\}}}(\boldsymbol{W}_{\{j,j\in D_s(i)\}}) & if\ k\ has \\ \prod_{j'\in D_m(i)} s(d_{j'}|k_{t_{\cdot,j'}}^{-i}, k_{t_{i,j'}}=k) & been\ used \\ \alpha_g f_{k^{new}}^{-\boldsymbol{W}_{j,j\in D_s(i)}}(\boldsymbol{W}_{j,j\in D_s(i)}) & if\ k=k^{new} \\ \prod_{j'\in D_m(i)} s(d_{j'}|k_{t_{\cdot,j'}}^{-i}, k_{t_{i,j'}}=k^{new}) & \end{cases}$$

After model training, the mixture components can be estimated as

$$\hat{\phi}_k = \int \phi_k p(\phi_k|\{\boldsymbol{W}_j|\begin{Bmatrix} d_j=k, j\in D_m(\cdot) \\ t_{\cdot,j}=k, j\in D_s(\cdot)\end{Bmatrix}\})d\phi_k \tag{15}$$

And the community distribution θ_i of a user i is estimated from the community assignment variables $d_j, j\in D_m(i)$ and $t_{i,j}, j\in D_s(i)$ as

$$\hat{\theta}_i = \frac{1}{|D_m(i)|}\sum_{j\in D_m(i)}\delta(\phi_{k_{d_j}}) + \frac{1}{|D_s(i)|}\sum_{j\in D_s(i)}\delta(\phi_{k_{t_{i,j}}}) \tag{16}$$

4. EXPERIMENTS

4.1 Data Collection

We mainly use the research proceeding corpus to evaluate the performance of TLM. The dataset contains the abstracts from 7 research conferences, i.e. ACL, ICML, SIGGRAPH, SIGIR, SIGKDD, SIGMOD, and WWW, from 2005 to 2009. We call this corpus PAPER. We also collect a set of companies[1] and their news articles from New York Times. The dataset consists of all the articles that mention about at least 3 companies. And hereafter we refer to it as NYT. Table 2 shows some statistics of PAPER and NYT.

4.2 Community Membership Evaluation

We use the Categorical Clustering Distance(CCD)[7] to compare the similarity between the computed community

[1] `http://topics.nytimes.com/topics/news/business/companies/index.html`

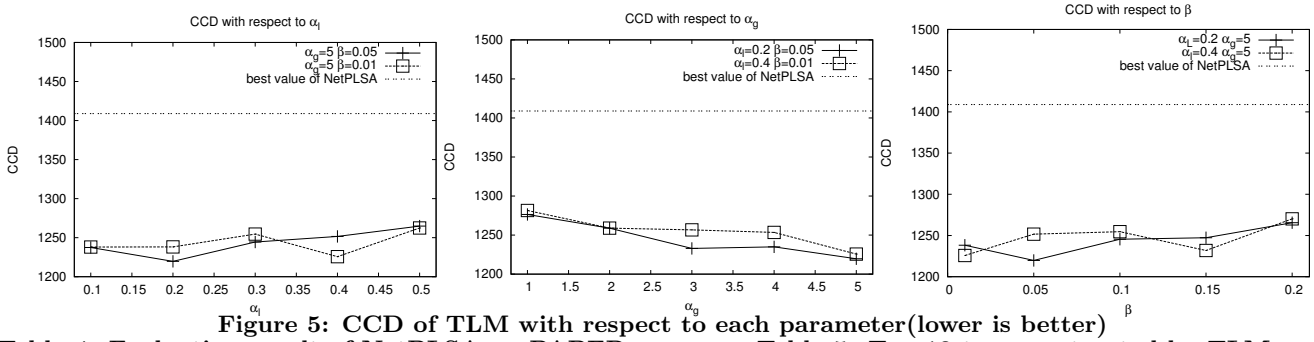

Figure 5: CCD of TLM with respect to each parameter(lower is better)

Table 4: Evaluation result of NetPLSA on PAPER

λ	0.1	0.2	0.3	0.4	0.5
$CCD(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$	1474.2	1463.8	1435.6	1509.9	1486.5
λ	0.6	0.7	0.8	0.9	
$CCD(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$	1408.4	1432.6	1435.7	**1408.9**	

distribution and the ideal community distribution. For PAPER, we treat each conference as a community and the proportion of the number of papers one author published in each conference as the ideal probability the author belongs to that community. On PAPER, The CCD of NCut is 1351.02, and we list the evaluation results of NetPLSA regarding to its parameter λ in Table 4.

From Table 4, the best CCD value of NetPLSA is 1408.9, which outperforms NCut, so from now on we choose the best evaluation result of NetPLSA as comparison baseline for TLM. In our experiments, we also try different sets of parameters of TLM with α_l varying from 0.1 to 0.5, α_g from 1.0 to 5.0 and β from 0.01 to 0.20. We show part of the detailed evaluation results of TLM with respect to α_l, α_g and β in Figure 5. With $\alpha_l = 0.2$, $\alpha_g = 5.0$ and $\beta = 0.05$, we get the minimum value of $CCD = 1219.65$ giving a maximum improvement over NetPLSA as $\frac{|1219.65-1408.9|}{1408.9} = 13.4\%$.

With $\alpha_l = 0.2$, $\alpha_g = 5.0$ and $\beta = 0.05$, TLM detects 8 communities and we make statistics of papers from each conference as shown in Figure 6. Note here that the community membership of a paper is set to be community assignment in the last iteration of Gibbs sampling.

4.3 Community Semantic Analysis

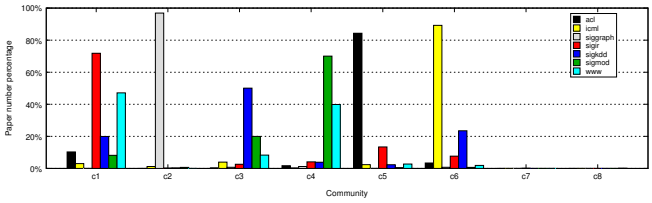

Figure 6: Paper number percentage distribution

From Figure 6, we see that our model discovered 6 major communities and 2 minor communities. We examine the 6 major communities first. Considering both Figure 6 and Table 5, it is easy to see that c_1 well corresponds to information retrieval, c_2 is closely related to computer graphics, c_3 is mainly about data mining, c_4 covers the database community, c_5 mainly concerns about computer linguistics, c_6 is closely related to machine learning. Note that papers from WWW scatter around several communities, which is quite reasonable because the WWW conference covers multiple topics. As to the 2 minor communities, c_7 and c_8 both contain only 1 paper. After an investigation of the data, we find that the authors of the above 2 papers have no co-authorship

Table 5: Top 10 terms extracted by TLM

c_1	c_2	c_3	c_4
search	imag	data	data
web	model	network	queri
queri	method	algorithm	web
user	motion	mine	system
inform	base	pattern	servic
model	surfac	graph	applic
retriev	present	model	databas
document	mesh	base	base
base	time	cluster	user
result	algorithm	propos	process

c_5	c_6	c_7	c_8
model	learn	portinari	economi
base	algorithm	work	engin
languag	model	paint	perspect
word	data	brazil	understand
method	method	social	long
approach	problem	project	chang
translat	cluster	present	arriv
system	propos	document	econom
paper	base	develop	largest
show	classif	import	compani

with the rest authors and that the contents of these papers are very dissimilar from others.

5. CONCLUSIONS AND FUTURE WORK

Although community discovery techniques have been developed for decades, there is no much work done in developing general algorithms for textual interaction graph. This paper proposes a principle solution. In the future, we will try some other document modeling e.g. LDA, and apply TLM to large-scale datasets.

6. ACKNOWLEDGEMENT

We would like to thank the three anonymous reviewers for their elaborate and helpful comments.

7. REFERENCES

[1] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proceedings of KDD 2000*, pages 150–160, 2000.

[2] J. Leskovec, K. J. Lang, and M. W. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of International Conference on WWW 2010*, pages 631–640, 2010.

[3] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *Proc. of WWW'08*, pages 101–110, 2008.

[4] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E.*, 69(026113), 2004.

[5] J. Ruan and W. Zhang. An efficient spectral algorithm for network community discovery and its applications to biological and social networks. In *Proc. of ICDM*, pages 643–648, 2007.

[6] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transcations of Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[7] D. Zhou, J. Li, and H. Zha. A new mallows distance based metric for comparing clusterings. In *Proc. of ICML'05*, pages 1028–1035, 2005.