Modeling Topic and Community Structure in Social Tagging: the TTR-LDA-Community Model[1]

Daifeng Li[1], Ying Ding[2], Cassidy Sugimoto[2], Bing He[2], Jie Tang[3], Erjia Yan[2], Nan Lin[4], Zheng Qin[1], Tianxi Dong[5]

[1]*School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China*
[2]*School of Library and Information Science, Indiana University, Bloomington, IN, USA*
[3]*Department of Computer Science and Technology, Tsinghua University,China*
[4]*School of International Business Administration, Shanghai University of Finance and Economics, Shanghai, China*
[5]*Rawls College of Business, Texas Tech University, TX, USA.*
ldf3824@yahoo.com.cn ,{binghe, dingying, sugimoto, eyan}@indiana.edu, jietang@tsinghua.edu.cn

Abstract
The presence of social networks in complex systems has made networks and community structure a focal point of study in many domains. Previous studies have focused on the structural emergence and growth of communities and on the topics displayed within the network. However, few scholars have closely examined the relationship between the thematic and structural properties of networks. Therefore, this paper proposes the TTR-LDA-Community model which combines the Latent Dirichlet Allocation (LDA) model with the Girvan-Newman community detection algorithm through an inference mechanism. Using social tagging data from Delicious, this paper is able to demonstrate the clustering of active taggers into communities; the topic distributions within communities; and the ranking of taggers, tags, and resources within these communities. The data analysis evaluates patterns in community structure and topical affiliations diachronically. The paper evaluates the effectiveness of community detection and the inference mechanism embedded in the model and finds that the TTR-LDA-Community model outperforms other traditional models in tag prediction. This has implication for scholars in domains interested in community detection, profiling, and recommender systems.

Key Words: Topic mining, Community detection, Social tagging system, TTR-LDA-Community

1. Introduction

Social networks have been extensively studied in many domains—using everything from sociological to mathematical perspectives. Researchers have found that most real word networks, differing from random networks, exhibit three common properties: a small world property, a power-law degree distribution and community structures with relatively high clustering coefficients (Erdos, et al., 1959; Milgram, et al., 1967; Newman, et al., 2001; Newman, et al., 2003). Exploring these communities is critical for the understanding of social networks. Communities in a social network might represent real social groupings by language, interests or demographic background; communities on the Web might represent pages on related topics; communities in a citation network might represent related papers on a specific project or a particular research question. The

[1] * Corresponding author at: School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China.
E-mail address: ldf3824@yahoo.com.cn (Daifeng Li).
This paper is a thorough extension of a short paper of CIKM conference (Li, et al., 2010).

heterogeneous nature of actors and the variety of interactions among the actors suggest dynamic forces underpinning the formation of communities. Being able to detect these communities and explore their features could help us to understand and utilize these networks more effectively. Many studies in various disciplines have been devoted to community detection (e.g., Girvan, et al., 2002). However, these studies focused on the structural property of communities and neglected other important aspects of communities, especially the thematic features of communities. Moreover, different aspects of communities may have interactive and iterative impacts on each other; for example, the formation of communities may be driven by common interests while common interests may emerge and be reinforced by community structure. Few previous studies have systematically and quantitatively delved into the interactive impact between structural and thematic properties of communities.

In recent years, the prevalence of online social bookmarking websites (e.g., Delicious) have created new opportunities for quantitative analyses of large-scale social networking datasets. The high participation in these social tools has drawn increasing research attention to understanding user communities and their behavior patterns. Exploring structural and thematic features of user communities in social tagging systems can improve community-supporting services--such as friend, resource, or tag recommendation. In parallel with studies on community detection in social tagging systems, graphical topic modeling has recently been proposed to mine the semantics of large corpora of tags, which uses specific features such as word occurrences within documents to reveal meaningful semantic content of documents (Blei, et al., 2003). Many studies have introduced LDA-based topic modeling into social tagging systems, showing the "hot" topics among users based on the tags they posted. However, few studies have integrated community detecting and topic modeling in studying user communities in social tagging systems.

This paper proposes a TTR-LDA-Community model, which is an inferential combination of an extended LDA model and a betweenness-based community detection algorithm. Moreover, two indicators--network community plot and modularity--are used to evaluate the quality of the detected community. The TTR-LDA-Community model can provide rich, systematic, and quantitative information about the profiles of detected communities.

The contributions of this model include: 1) more nuanced understanding of the underlying motifs associated with community structure than existing community detection algorithms; 2) integration of community structure and topic modeling, not explored with with previous LDA-based topic models; and 3) a diachronic analysis that, by integrating LDA-based topic modeling and a community detection algorithm, reveals how community structure evolves over time. In addition, the dataset employed here, comprehensive coverage (2005-2008) of the Delicious social tagging data, enables examination of the relationship between community and topics in a real-world example.

2. Related work

Two groups of research, community detecting and topic modeling, are closely related to our research. Various community detection algorithms and LDA-based topic models are summarized in this section.

2.1 Community Detection

Researchers have used a number of methods to detect communities within networks. Two widely used approaches are those based on 1) centrality and 2) graph partitioning. Girvan and Newman (2002) used betweeness-centrality to examine the community structure in large networks. Their algorithm functions by 1) calculating edge-betweeness for all edges in the network, 2) removing the edges with highest betweeness, 3) recalculating betweenenss for all edges affected by the removal, and 4) continuing this pattern until no edges remain. The original algorithm was improved upon by Clauset et al. (2004), who reduced the complexity from $O(m^2n)$ to $O(md\log n)$ (where d is the depth of the dendrogram of the community structure). This algorithm has been tested empirically and validated as an appropriate model for community detection (e.g., Radicchi et al., (2004)).

Two standard examples of the graph partitioning approach are the local spectral partitioning algorithm (Andersen, et al., 2008) and the flow-based Metis_MQI algorithm (Flake, et al., 2003). These approaches were compared to the Girvan-Newman algorithm by Leskovec et al. (2010). In Leskovec's research (2010), all the representative algorithms can detect similar compact clusters with high conductance inside and low conductance outside from a large-scale network. The Newman-Girvan algorithm is established as an efficient method to identify these clusters in large-scale networks; it is also one of the most commonly used topology-based community detection approaches. In this paper, our main purpose is to detect the topic distribution from each community's structure, which has not been taken into consideration in existing research. In addition, the Newman-Girvan algorithm provides three advantages over other algorithms: 1) compared with traditional k-means clustering method, there is no need to provide the number of clusters in advance; 2) there is also no need to find the cut point for the dendrogram if using hierarchical clustering; and 3) instead of trying to construct a measure that tells us which edges are the most central to communities, the communities are detected by progressively removing edges from the original graph, rather than by adding the strongest edges to an initially empty network (Girvan, 2002). Thus, the Newman-Girvan algorithm is used in this study.

2.2 Topic Modeling

Since the introduction of the LDA model (Blei, et al., 2003), various extended LDA models have been used for automatic topic extraction from large-scale corpora. In the LDA model, a document is viewed as a collection of "bags of words". It assumes that each article has the probability of several topics and that topics are associated with different conditional distributions over a fixed set of words. The words within the entire corpora are used to generate topics and each particular document is assigned a probability distribution

according to how likely it is to be about the given topic. Simply put, LDA is a "mixture of mixtures model": the mixture components are shared across all documents but each document exhibits different mixture proportions (Blei, et al., 2003). In the context of tagging systems, where multiple users are annotating resources, the resulting topics reflect a shared view of the document; and the tags belonging to the topics reflect a common vocabulary.

Many researchers have extended the basic LDA model. Rosen-Zvi et al. (2004) introduced the Author-Topic (AT) model, which extended LDA to include authorship as a latent variable. In the AT model, each author is associated with a multinomial distribution over topics. The primary benefit of the Author-Topic Model is that it provides a general framework for answering queries and making predictions at the level of authors as well as the level of documents. Based on the Author-Topic model, McCallum et al. (2004) presented an Author-Recipient-Topic (ART) model for social network analysis, which provided topic distributions based on the direction-sensitive messages sent between entities. They added the attribute that distribution over topics is conditioned distinctly on both the sender and recipient. Tang et al. (2008) further extended the LDA and Author-Topic model and proposed the Author-Conference-Topic (ACT) model, which is a unified topic model for simultaneously modeling different types of information in academic networks. Tang et al. (2008) found that the proposed method had a high performance in expertise search and association search. Si and Sun (2009) proposed a tag-LDA model, which extended the LDA model by adding a tag variable, and applied it to social tagging systems.

Similar to social tags, the link structure of networks has served as an additional area for network research. Cohn and Hoffman (2001) proposed Probabilistic HITS (PHITS), an extension to Probabilistic Latent Semantic Indexing (PLSI), which defines a generative process for hyperlinks and thereby models topic-specific influence of web documents. It assumes the generation of each hyperlink in a document is a multinomial sampling of the target document from the topic-specific distribution of the documents. Erosheva et al. (2004) proposed a similar model, in which PLSA was replaced by LDA as the fundamental generative building block, usually referred to as the Link-LDA model. Later, Dietz et al. (2007) presented a new LDA-based approach that utilizes the flow of topic information from the cited documents to the citing documents. Nallapati and Cohen (2008) proposed a Link-PLSA-LDA model as a scalable LDA-type model for topic modeling and link prediction. Chang and Blei (2009) introduced the relational topic model (RTM) to model the link between documents as a binary random variable conditioned on their contents. Although research has been done in both the areas of community detection and topic analysis, very few researchers have sought to combine the two. One notable exception is the work of Liu, Niculescu-Mizil, and Gryc (2009) who examined topic and author communities for a set of blog posts and citation data through jointly modeling underlying topics, author community, and link formation in one unified model. However, it was done synchronically, rather than diachronically: thereby, it did not provide an evaluation of how the model functions in examining change in topics over time.

As discussed above, studies on community detection haven't taken other aspects of community profile into consideration, while researches on topic modeling largely neglect potential relationships between topics and community structure. This paper proposes a different approach to address this question, combining topic modeling and community detection through an inference mechanism.

3. Methodology

In this section, three core modules of the TTR-LDA-Community model are described in detail, including topic model, community detection and an inference mechanism. Additionally, the sampling and processing of the experiment data is also presented. In order to make the content easy to understand, the notation is summarized in Table 1.

Table 1
Notation Table

Notations	Meaning
I	The whole data set (including all taggers, tags, resources and posts)
$N[tap]_{p->I}$	The tagger ta in post p is the $N[tap]$th tagger in I.
$N[tp]_{p->I}$	The tth tag in post p is the $N[tp]$th tag in I.
$N[rp]_{p->I}$	The resource r in post p is the $N[rp]$th resource in I.
α	hyperparameter for generating Θ from Dirichlet Distribution
β	hyperparameter for generating φ from Dirichlet Distribution
μ	hyperparameter for generating Ψ from Dirichlet Distribution
Θ	The multinomial distribution of taggers over topics
φ	The multinomial distribution of topics over tags
Ψ	The multinomial distribution of topics over resources
P	The number of posts
TA	The number of taggers
T	The number of tags
R	The number of resources
K	The number of topics
$nw[m][z]$	The number of times to assign mth tag in lexicon to topic z in I.
$nwksum[z]$	The number of times to assign all tags to topic z in I.
$na[a][z]$	The number of times to assign zth topic to ath tagger in I.
$naksum[a]$	The number of times to assign all topics to ath tagger in I.
$nc[r][z]$	The number of times to assign rth resource to topic z in I.
$ncksum[z]$	The number of times to assign all resources to topic z in I.
U	The set of all taggers in the social tagging networks
V	The set of all edges among taggers in the social tagging network
$gp_e(u,v)$	The number of all paths that pass through node u, edge e and node v
$gp(u,v)$	The number of all paths that pass through node u and v.

3.1 Topic Model

This section introduces the TTR-LDA model, which extends the ACT model to the context of social tagging (Tang, et al., 2008). The objective of this paper is to detect thematic features of each post in a social tagging system, which is important for analyzing topic distribution in communities. The TTR-LDA is proposed

as a solution for this objective. The TTR-LDA model is an extension of the ATC1 model (Tang, 2008), which was shown to outperform other LDA-based models (e.g., Language model, LDA, Author-Topic, etc.). The ACT1 model is used to simulate an author writing and submitting an article. Similar with the ACT1 model, the TTR-LDA model is used to simulate a tagger bookmarking tags for a certain resource, thus these simulations are used to make further analysis on the structure of communities in order to improve results. TTR-LDA is three-layer Bayesian model with taggers *tap* in each post *p* as the first layer, tags *t* and resource *r* as the third layer and all the topics denoted as latent variable *z* as the middle layer. In this model, the notation for total number of posts is *P*, distinct taggers is *TA*, total resources is *R*, and distinct tags is *T*. *K* is the total number of topics (to be determined by perplexity analysis, discussed later in this section). Using this notation, the process of TTR-LDA can be described as follow:

1. Choose $\theta \sim Dir(\alpha)$, $\phi \sim Dir(\beta)$, $\psi \sim Dir(\mu)$;
2. For each post p (P posts):
3. For each tags t_{pi} in post p:
4. Choose a topic z_{pi} for t_{pi} and assign that topic to tagger ta_p and resource r_p in post p according to multinomial($\frac{1}{K}$);

/Initial for estimating:

1. Choose $\theta \sim Dir(\alpha), \phi \sim Dir(\beta), \psi \sim Dir(\mu)$
2. For each post p:
3. For each tag tp in p:
4. For tagger tap in p:
5. Select topic z~Multi(θ) for tagger tap for tag tp in post p:
6. Init the parameters: nw[N[tp]$_{p->I}$][z]+=1; na[N[tap]$_{p->I}$][z]+=1; nc[N[rp]$_{p->I}$][z]+=1;
7. nkwsum[z]+=1; nkcsum[z]+=1; naksum[N[tap]$_{p->I}$]+=1;
8. int tagger= N[tap]$_{p->I}$; cid= N[rp]$_{p->I}$; topic=z; tag= N[tp]$_{p->I}$; post=p;
9. end
10. end

/Gibbs Sampling:

1. For each iteration (2000 times):
2. For each post p:
3. For each tag tp in post p:
4. nw[tag][topic]-=1; na[tagger][topic]-=1; nc[cid][topic]-=1; nkwsum[topic]-=1;
5. naksum[tagger]-=1; nkcsum[topic]-=1;
6. For tagger tap in document p:
7. For each topic z:
8. confprob=(nc[N[rp]$_{p->I}$][z]+ μ)/(nkcsum[z]+C* μ);
9. topicauthorprob=(na[N[tap]$_{p->I}$][z]+ α)/(naksum[N[tap]$_{p->I}$]+K* α);
10. wordtopicprob=(nw[N[tp]$_{p->I}$][z]+ β)/(nwsum[z]+V* β);
11. prob[N[tap]$_{p->I}$, z] =wordtopicprob*topicauthorprob*confprob;
12. End for topic z;
13. Random select u~Multi(1/K);
14. For tagger tap:
15. For each topic z:
16. If $\sum_{j=1}^{z} prob\ (j)$ >=u then
17. Break;
18. End for topic z
19. Assign topic=current z;
20. All parameters for z, p, tp, rp should be added 1.
21. Recover the original situation for last instance.
22. End for tag tp
23. End for post p
24. Update θ, φ, ψ
25. End for Iteration

Fig. 1. Algorithm of Gibbs Sampling

This model contains three unassigned parameters: θ, ϕ, ψ (shown in TTR-LDA part in Figure 2). As shown by Tang, Jin, and Zhang (2008), these hyperparametrics can be estimated using Gibbs sampling. The detail process can be seen in Figure 1. In the TTR-LDA model, we consider each resource r as a collection of

topics; each tagger can generate a post to describe the resource. According to the logical process of a tagger's bookmarking activity, we use the following equation to calculate the posterior conditional probability:

$$P(z_{pi} \mid z_{-pi}, ta_p, t_{pi}, r_p, \alpha, \beta, \mu) \propto \theta_{ta_p z_{pi}} \times \phi_{z_{pi} t_{pi}} \times \psi_{z_{pi} r_p}$$

$$\theta_{ta_p z_{pi}} = \frac{na[N[ta_p]_{p->I}][z_{pi}] + \alpha}{\sum_{z'} na[N[ta_p]_{p->I}][z'] + K\alpha}, \phi_{z_{pi} t_{pi}} = \frac{nw[N[t_{pi}]_{p->I}][z_{pi}] + \beta}{\sum_{t'} nw[t'][z_{pi}] + T\beta}, \psi_{z_{pi} r_p} = \frac{nc[N[r_p]_{p->I}][z_{pi}] + \mu}{\sum_{r'} nc[r'][z_{pi}] + R\mu}, \quad (1)$$

The equation above is used to calculate each element in matrix θ, ϕ, ψ during the iterations of Gibbs Sampling and can be found in Step 8-11 in the Gibbs Sampling part in Figure 1. z_{pi} indicates the assignment of topic *z* to post *p* and tag *i* in post *p*; z_{-pi} means the assignment of *z* to other posts and tags before this situation. $\theta_{ta_p z_{pi}}$ means the probability of assigning topic *z* to tagger *ta* in post *p*, at the same time, the *i*th tag in post *p* is also assigned to topic *z*; $\phi_{z_{pi} t_{pi}}$ means the probability of assigning *i*th tag in the post *p* to topic *z*; $\psi_{z_{pi} r_p}$ means the probability of assigning resource *r* in post *p* to topic *z*, at the same time, the *i*th tag in post *p* is also assigned to topic *z*.

For the estimation of hyperparameters α, β, γ, we assign different values for each hyperparameter and run TTR-LDA model to get the results. After several experimentals, we find that different values of hyperparameters have little influence on the performance of the TTR-LDA model, consistent with Lu, Hu, Chen et al.'s (2010) results for Del.icio.us. Using the estimates provided in Tang, Jin, and Zhang (2008), we assign the hyper parameters as: α =50/K (where K is the number of topics), β =0.01 and μ =0.1.

3.2 Community Detection

The Girvan-Neman algorithm is used for community detection (Girvan, et al., 2002). As described in the Related Works section, this algorithm uses betweeness-centrality to detect boundaries between communities. As defined by Girvan and Newman (2002), the betweenness of an edge is the number of shortest paths between pair of vertices. The process for this algorithm is described in Figure 2. A complex network in a social tagging system such as Delicious can be described as *G={U, V}*, where *U* represents the set of nodes/taggers and *V* represents the set of edges. $U=\{tagger_1, tagger_2, tagger_3, ...tagger_{TA}\}$, where *TA* means the total number of nodes/taggers in the network; $V=\{e_1, e_2, e_3, ...e_v\}$, where e_i means two nodes/taggers have a co-bookmark relationship with each other. Then the edge betweenness centrality formula can be seen as below:

$$ebc(e) = \sum_{u \in U, v \in V} \frac{gp_e(u,v)}{gp(u,v)} \quad (2)$$

The formula $gp(u,v)$ represents the number of all paths that pass through nodes *u* and *v*, while $gp_e(u,v)$ represents the number of all paths that pass through nodes u, edge e and node v. In each step, the edge with the highest betweenness is deleted. The loop doesn't end until the clustered communities satisfy the constrained

conditions. In a division of the 50,000 most active taggers, assuming that L different communities are identified, they can be expressed as:

$$Community_{Tagger-Community} = \begin{cases} 1, if\ \ tagger_k \in community_l \\ 0, if\ \ tagger_k \notin community_l \end{cases} \quad (3)$$

3.3 Inference Mechanism

Current models of community detection provide an identification of *how* the communities are structured, but not *why*. Therefore, this paper proposes the TTR-LDA-Community model, which combines community detection with statistical topic mining (i.e., TTR-LDA) to uncover both the structural and semantic features of communities. As described below, this model is an extension of the ACT model (Tang, et al., 2008), with an application to the context of social tagging.

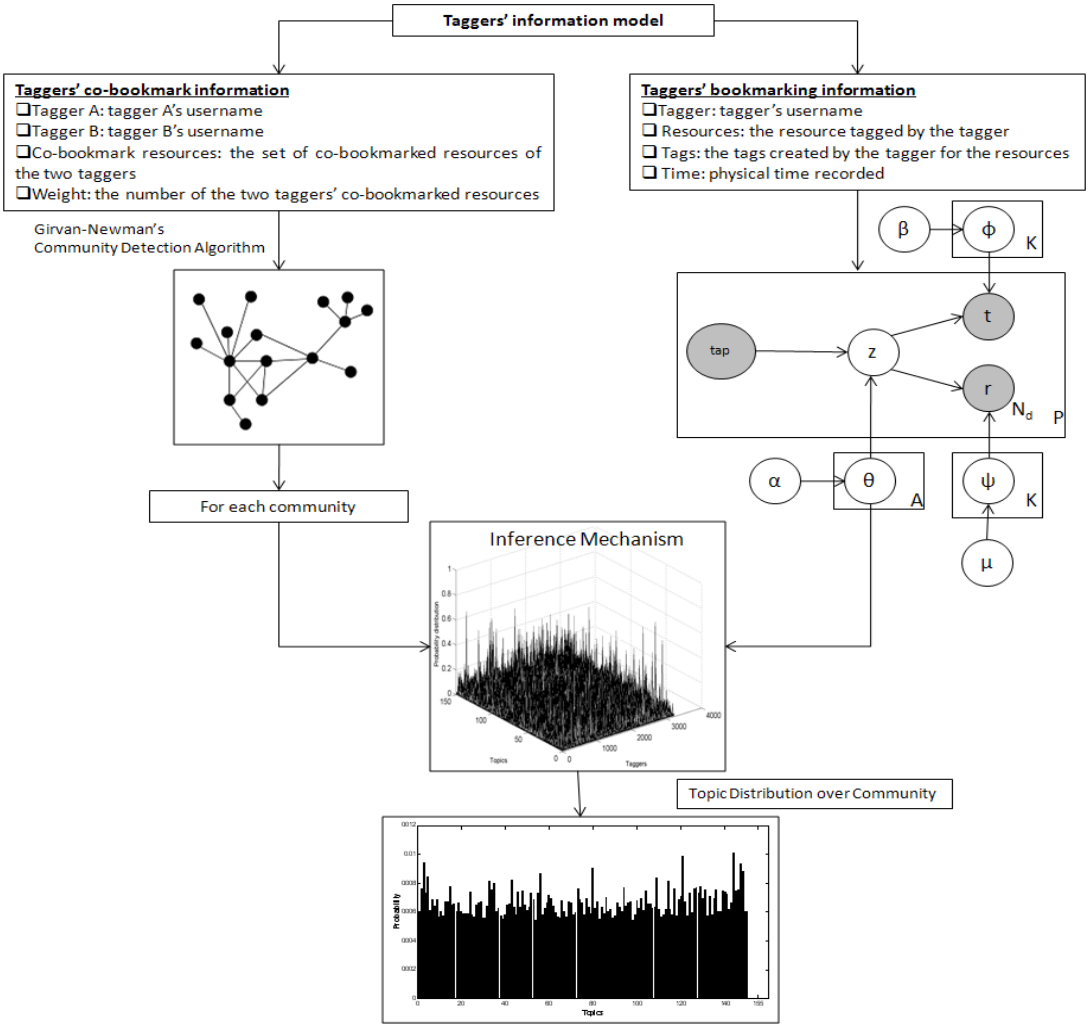

Fig. 2. TTR-LDA-Community Model.

An information profile must first be created for each of the taggers in the sample. This involves identifying the posts created by each unique tagger. For example, if tagger A bookmarked resource B with a set of tags C {*tag1, tag2, tag3… tagn*}; then that tagging instance (post) would be defined as {*tagger=A, resource=B*, *tag*=C {*tag1, tag2, tag3… tagn*}}. In total, 232,212 posts were identified in the sample.

The inference mechanism is used to infer the topic distribution over each community, which is generated by the community detection algorithm; each community includes a set of taggers, where a community is defined by similarities between taggers and tagging behavior. In order to uncover the semantic features of detected communities, the paper uses the inference mechanism to integrate an extended LDA model (i.e., TTR-LDA) and the community detection algorithm.

As shown in Figure 2, the topic-tagger probability distribution is identified through the taggers' information profile. If it is assumed that there are K topics and N taggers, a T*N matrix can be created that records the probability distribution of all topics, for each tagger. This matrix can be expressed as $Matrix_{Topic-Tagger} = \{P(T_i \mid Ta_j) \mid i \in 1:T,\ j \in 1:N\}$.

Second, the taggers' information model is imported into the community model by using the Community Detection algorithm to compute the communities.

Third, an inference mechanism is used to integrate the results from the TTR-LDA model and Community Detection model. $community_l$, which is detected from community detection model, is defined as $community_l = \{CA, E\}$ where *CA* means the set of taggers who belong to $community_l$ and $E=\{e_1, e_2, e_3, ...e_o\}$ means the connection/edges among different taggers in $community_l$, if $community_l$ contains *q* taggers (in this paper, the connection means two taggers have a co-bookmark activity for the same resource, *o* means the total number of edges). Assuming those *q* taggers are $tagger_1, tagger_2, tagger_3, ...tagger_q$; for $tagger_k$ in $community_l$, he/she has a corresponding probability distribution over *K* topics in $Matrix_{Topic-Tagger}$ as $\{P_{k1}, P_{k2}, P_{k3}, ...P_{kK}\}$. Each $community_l$ contains *p* taggers. The *p* taggers can be individually labeled as $tagger_1, tagger_2, tagger_3, ...tagger_p$; for $tagger_k$ in $community_l$, he/she has a corresponding probability distribution over *K* topics in $Matrix_{Topic-Tagger}$ as $\{P_{k1}, P_{k2}, P_{k3}, ...P_{kT}\}$. Therefore, for $community_l$, the probability distribution over K topics can be computed as $\{\frac{\sum_{k=1}^{p} P_{k1}}{p}, \frac{\sum_{k=1}^{p} P_{k2}}{p}, \frac{\sum_{k=1}^{p} P_{k3}}{p}, ... \frac{\sum_{k=1}^{p} P_{kT}}{p}\}$.

To identify the most representative resources for each community, the Pearson similarity correlation is conducted to identify those resources with similar topic distribution within a given community. For example, if a resource has a probability distribution over *K* topics as R=$\{R_1, R_2, R_3, ...R_T\}$ and the community has a probability distribution over *K* topics as C=$\{C_1, C_2, C_3, ...C_T\}$, the similarity between *R* and *C* can be seen as:

$$similarity_{R-C} = \frac{\sum_{i=1}^{T}(R(i)-\bar{R})\times(C(i)-\bar{C})}{\sqrt{\sum_{i=1}^{T}(R(i)-\bar{R})^2}\sqrt{\sum_{i=1}^{T}(C(i)-\bar{C})^2}} \quad (4)$$

$\bar{R}$ denotes the mean value of all R and $\bar{C}$ means the mean value of all *C*.

3.4 Data

There are three major components in social tagging: the tag (the label used by a tagger to describe a resource), the tagger (the individual doing the tagging), and the resource (the item that is tagged). These components, taken together, are called a "triple." The sampling frame for this study was comprised of all triples and the time and date of their creation on Delicious from 2005 to 2008. The sample used for this study was the 50,000 most prolific taggers: those who tagged the largest number of resources. In total, 354,522 unique resources were associated with these 50,000 taggers. These resources were then ranked by the number of taggers (within the sample) associated with the resource (through tagging). Therefore, the top 50,000 taggers were used as the sample for this study. A co-bookmark network was created between taggers and resources, where a connection exists when two taggers tag the same resource. Co-bookmark network reflects the shared interest between taggers. Meanwhile, the set of tags created by those taggers in the co-bookmark network is also collected. The co-bookmark network and the set of tags act as the two inputs of our model, capturing all the three major components of a social tagging website.

In addition, in order to observe the evolution of structure and features of communities, the paper divided the time span (2005-2008) into four slices. Table 2 shows the descriptive statistics of the data in different time slices.

Table 2
Descriptive Statistics of Delicious data in the four time slices

	2005	**2006**	**2007**	**2008**
No. of posts	11,451	49,583	170,165	1,108,782
No. of resources	7,117	25,036	63,273	311,518
No. of tagger	3,616	12,053	28,823	48,688
No. of tag	10,014	31,493	78,661	283,188

As the 2008 time period contained considerably more edges than the previous time periods, it was further divided into four sections, to ensure a similar number of taggers across each time period. The size of the networks created by this division is shown in Table 3 (2-core is used to get the co-bookmarked network with

stronger connection for each time periods: 2-core means that we consider two nodes have a connection if they co-bookmarked more than 2 resources in the system).

Table 3
Information of all taggers and their co-bookmarked network in 2008

	Jan-March	April-June	July-Sep	Oct.-Dec
nodes	6438	8283	15074	15399
edges	14473	21306	47225	81993

4. Results and discussion

In this section, results from applying the proposed model to the experimental data are presented, including an overview of the communities from the community detection module, an overview of the topic distribution from the topic model module, and an integration of structure and topics of communities from the inference mechanism module. Additionally, a dynamic analysis of the interactive impact between the structural and topic features of communities is presented.

4.1 An Overview of Communities

In this analysis, taggers were considered linked if they had co-bookmarked at least two resources for the time period under examination. Each time period was examined for the number of communities in the time period and the sizes of each of the five largest communities within that time period.

Table 4
The summary of community information in each year

	2005	2006	2007	2008
Size of network	3,252	11,811	28,048	38,966
Number of communities	160	224	143	299
Size of largest community	510(15.68%)	2,648(22.42%)	11,149(39.75%)	15,776(34.91%)
Size of second community	499(15.34%)	2,417(20.46%)	8,265 (29.47%)	11,159(24.69%)
Size of third community	389(11.96%)	1,973(16.70%)	6,496 (23.16%)	7,523 (16.65%)
Size of fourth community	310 (9.53%)	1,573(13.32%)	926 (3.30%)	1,318 (2.92%)
Size of fifth community	244 (7.50%)	531 (4.50%)	462 (1.65%)	500 (1.11%)
Ratio of top five communities	60.02%	77.4%	97.29%	80.27%

Table 5
The summary of communities information in each time slice in 2008

	Jan-March	April-June	July-Sep	Oct.-Dec
Size of network	6,438	8,283	15,074	15,399
Number of communities	752	784	767	487
Size of largest community	1,248 (19.38%)	1,829 (22.08%)	4,377 (29.04%)	3,855 (25.03%)
Size of second community	500 (7.77%)	1,146 (13.84%)	2,138 (14.18%)	3,206 (20.82%)
Size of third community	445 (6.91%)	730 (8.81%)	1,670 (11.08%)	2,845 (18.48%)
Size of fourth community	440 (6.83%)	475 (5.73%)	733 (4.86%)	752 (4.88%)
Size of fifth community	247 (3.84%)	457 (5.52%)	674 (4.47%)	521 (3.38%)
Ratio of top five communities	44.73%	55.98%	63.63%	72.60%

As shown in Table 4 and 5, the top five communities in 2007 and 2008 contain more than 90% of all the taggers, suggesting that the main content and topic distributions are dominated by these communities. However, when examining the top five communities within 2008, the proportion is smaller, ranging from 44-73%. One possible explanation is that the birth rate of new posts accelerates over time; for example, there are 232,870 new posts between 2005 and 2007 and 225,951 new posts between April and June of 2008. This may reflect an increase in taggers and resources in 2008. At the beginning of 2008, there are fewer resources bookmarked and, therefore, a lower probability of co-bookmarking. This results in a large number of small communities. However, as the size of the network has grown, the probability of co-bookmarking has also increased. Therefore, as suggested by the value of modularity, the connection strength among taggers in a community decreases, while the size of the community increases. This results in a dominance of a few large communities in the network; rather than an equal distribution among smaller communities.

The communities were then examined in terms of composition, that is, the number of taggers unique to a time period and to a given community. The number of shared taggers between communities was examined diachronically (see Table 6).

Table 6
Matching between communities in different time slices within 2008

		April-June			April-June			July-Sep
Jan-March	**1st**	New taggers 1204 (65.83%)	**July-Sep**	**1st**	New taggers 2970 (67.85%)	**Oct-Dec**	**1st**	New taggers 1660 (43.06%)
	2nd	New taggers 839 (73.21%)		**2nd**	New taggers 1550 (72.50%)		**2nd**	New taggers 1647 (51.37%)
	3rd	New taggers 263 (36.03%)		**3rd**	New taggers 1293 (77.43%)		**3rd**	New taggers 1789 (62.88%)
	4th	New taggers 394 (82.95%)		**4th**	New taggers 577 (78.72%)		**4th**	communities except for five largest ones 451 (59.97%)
	5th	New taggers 373 (81.62%)		**5th**	New taggers 517 (76.71%)		**5th**	New taggers 294 (56.43%)

In Table 6, the largest overlap between communities in the current year and communities in a previous year is shown. For example, the first cell indicates that the largest communities in the second time period in 2008 shared the most taggers with newly registered users in the previous time period. As shown in the table above, a majority of communities in one time period have the largest overlap with newly registered taggers in that time period instead of any communities in the previous time period. This may suggest that the community structure of the Delicious network has been unstable during its evolution. There is also considerable change in the number of new taggers for each of the following time period, suggesting a constant influx of new taggers. However, there are some exceptions: for example, the fourth largest community in Oct-Dec shares a high proportion of taggers from several small communities from previous time periods. This may indicate that these smaller communities merged together during Oct-Dec to form one larger community, with a shared interest profile.

4.2 An Overview of Topic Distribution

To identify the actual topics in each of the communities, the TTR-LDA-Community model was then applied to the top five communities for each time period of 2008. The number of topics to be identified for each time period was determined through perplexity analysis. For this analysis, a small set of training data was used to determine the number of topics that best fits the data. Posts created from October 2008 to December 2008 were used as the sampling frame for the training data. There were 43,453 unique taggers and 350,721 unique posts in this period. 1 out of every 100 posts was chosen until a test set of 3,000 was reached. Perplexity is an index, which describes the performance of a statistical model: the lower the perplexity value, the better a model fits the actual distribution (Rosen-zvi, et al., 2004). The results of the perplexity analysis are shown in Figure 3.

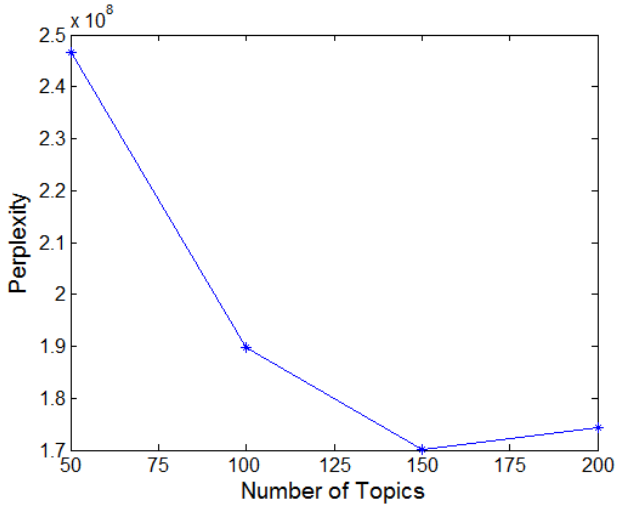

Fig. 3. Perplexity value for increasing number of topics

As shown in Figure 3, the lowest perplexity value (and therefore best fit) is found at 150 topics. Therefore, 150 topics were identified for all posts between October 2008 and December 2008.

Table 7
The overview of 5 top ranked topics in different time period during 2008

	2008 1-3	2008 4-6	2008 7-9	2008 10-12
1st rank topic	Bandslash, bandom fiction	Bandslash, bandom fiction	Web art design	Web 2.0 education and entertainment
2nd rank topic	Fan, humor fiction	Fan, humor fiction	Web 2.0 and on-line education	Web art design
3rd rank topic	Supernatural fiction	Supernatural fiction	Web 2.0 and Social network	Web 2.0 and Social network
4th rank topic	Programming language such as python, java	Web 2.0 and Social network	Web programming language	Web programming language
5th rank topic	Multi-media such as music, videos in youtube	Web 2.0 and on-line education	Operation system such as xp, vista and security problem	Operation system such as xp, vista and security problem

As shown in Table 7, the most popular topics have changed over: in the first 2 months, fiction was very popular in all taggers' groups, after that, Web 2.0 social network and Web art design take up the position of the most popular topics. Operation systems and programming languages remain dominant in all time periods.

4.3 Integration of Structure and Topics of Communities

The top 5 communities for each time period were then extracted and an interest profile for each community was created using an inference mechanism. After using TTR-LDA to examine the training data, topic distributions for all elements were created (e.g., taggers, tags, posts, resources). These topic distributions were used to conduct analysis and make predictions on new posts and resources automatically. The set of communities were identified using the Newman-Girvan algorithm. In order to conduct analysis on the thematic

features of detected communities, one important function of the inference mechanism is to collect tagger-topic, resource-topic, post-topic and tag-topic distributions for each detected community. For example, to identify the topic distribution of all taggers in community *i* which is detected from Newman-Girvan algorithm, the inference mechanism should identify first which taggers belong to community *i* and then search topic distribution (derived from TTR-LDA model) for each selected tagger in community *i*. This process allows for semantic analysis on community structure and improves in the prediction of new taggers or resources. In addition, the Newman-Girvan algorithm can guarantee that one tagger can only appear in one community, which means that one can use the tagger's global topic distribution to replace his/her topic distribution in that community. There are two advantages by making that approximation: 1) compared with applying TTR-LDA to run dataset in each detected community, it can reduce time complexity of the inference mechanism; 2) it can guarantee the consistency of the number of topics in all communities and provide convenience for further analyses. For example, if TTR-LDA is used to run datasets in two different communities separately, TTR-LDA will number the same topic with a random number, which expends time in identifying the same topics from all other communities.

Application of the inference mechanism of the TTR-LDA-Community model provides a topic distribution for this community. By using the inference mechanisms, we can obtain a topic distribution of the community; for example, the most popular topics in the largest community during Oct-Dec 2008 is Web design and management (topic 143), java, jquery, ajax (topic 121), traveling and shopping (topic 3), social networking (topic 80) and politic related topics (topic 47).

4.4 Dynamic Evolution of Structure and Topics of Communities

The topic distribution is then analyzed for each of the top five communities for each of the time periods in 2008. For the analysis, the number of topics was set at 150, for consistency with the training data. This provides an opportunity for examining topics across time. The following procedure was used for this analysis:

1. Identify the top 5 communities from each time period in 2008 and designate the community as *community_i_t* where *t* means the *th* time slice in 2008 and *i* means the *ith* largest community in *th* time slice;
2. Apply TTR-LDA-Community model to each community to compute their topic distribution and denote the result as *model_t_i_Topic(j),* which means the probability of *jth* topic in *ith* largest community in the *tth* time slice;
3. For each topic, obtain the probability distribution of tags belonging to that topic, which denotes the level of representativeness of tags for that topic: the higher the probability is, the more representative the tag is for that topic;
4. For each topic, sort all the tags according to their probability and select 20 top ranked tags to represent the content of the topics; meanwhile, for each community, select the top 5 ranked topics to represent its theme;

5. Perform a similarity analysis for different communities from different time slices. Identify the five topics in each community and compute how many tags are shared by two different communities from different time periods.

Therefore, the analysis provides not only a diachronic examination of topics, but describes the co-occurrence of tags between different communities, which is used to identify similarities of communities. It should be noted that the same tag might belong to different topics within a community.

Each time period was then compared to examine evaluations in topic profiles. For example, *community_i_t* is compared with *community_j_t-1 (j=1,2,…5)*. Topical similarities of the communities are shown in Tables 8, 9, and 10. The number in each cell denotes the occurrences of shared word between the two sets of popular topics of the two corresponding communities.

Table 8
Similarity matrix for communities in July – Sep, 2008 (row) and Oct. – Dec, 2008 (column)

	Largest	**Second**	**Third**	**Fourth**	**Fifth**
Largest	0.350	0.480	0.185	0.005	***0.355***
Second	0.310	0.235	***0.745***	0.005	0.050
Third	0.290	0.525	0.490	0.005	0.130
Fourth	***0.545***	0.380	0.535	***0.010***	0.055
Fifth	0.4550	***0.590***	0.375	0	0.135

Table 9
Similarity matrix for communities in April – June, 2008 (row) and July – Sep, 2008 (column)

	Largest	**Second**	**Third**	**Fourth**	**Fifth**
Largest	***0.315***	0.375	0.275	0.245	***0.530***
Second	0.270	0.150	***0.645***	0.180	0.250
Third	0.010	0	0.010	0	0
Fourth	0.120	0.145	0.180	0.160	0.290
Fifth	0.140	***0.550***	0.255	***0.450***	0.260

Table 10
Similarity matrix for communities in Jan – March, 2008 (row) and April – June, 2008 (column)

	Largest	**Second**	**Third**	**Fourth**	**Fifth**
Largest	***0.405***	0.250	0.010	***0.245***	***0.215***
Second	0	0	0.585	0	0
Third	0.295	0.335	0	0.110	0.095
Fourth	0.270	***0.460***	0	0.145	0.180
Fifth	0	0	***0.730***	0	0

One pattern displayed by the data was the fragmentation of topics over time into more specialized communities. Figure 4 depicts the topic trends consistent with *Community_1_1*.

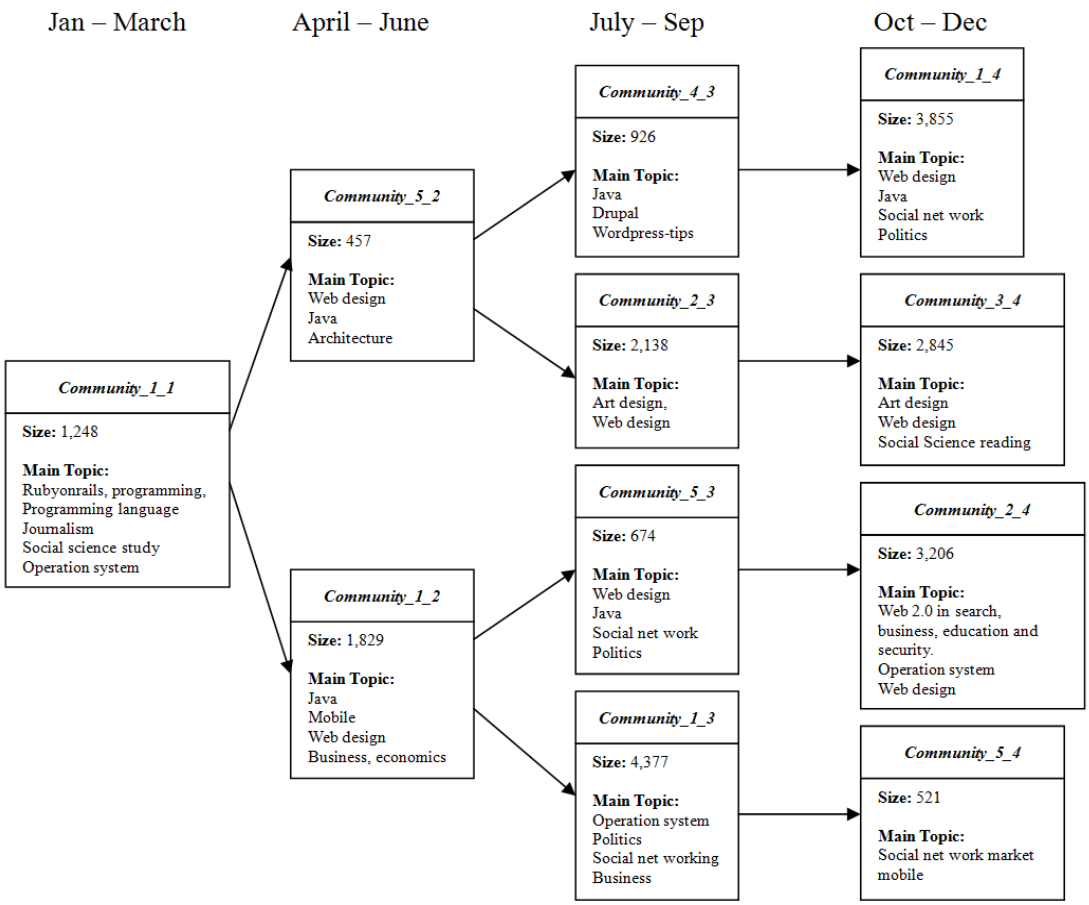

Fig. 4. Parts of the whole evolutionary line of communities over time in time slices within 2008.

As shown, topics affiliated with computer technology fall into a single community for the first time period. In the second time period, this general topic is grouped into two, more specialized communities. This fragmentation pattern continues over time, with new topics emerging within each specialty group. In addition, the size of the communities fluctuates with the influx of new taggers. For example, the size of community about social networks in the 3rd time slice (*Community_1_3*, 4,377) is larger than that in the 4th time slice (*Community_5_4*, 521). The main topics for all communities are listed below.

Table 11
The main topics of all communities in 2008

	Jan—March	April – June	July – Sep	Oct – Dec
Largest	(1) rubyonrails, programming 0.033547 (2) programming language 0.028276 (3) journalism 0.022598 (4) social science study 0.021831 (5) operation system 0.021766	(1) Java 0.0557 (2) mobile, media; marketing 0.032112 (3) flash, web design; 0.030742 (4) Business, economics; 0.025149	(1) Operation system 0.03546 (2) Politics 0.032292 (3) Social net working business 0.032284	(1) Web design 0.0552 (2) Java 0.0382 (3) Social net work 0.0326 (4) Politics 0.03183
Second	(1) super natural fiction 0.080606 (2) fan faction 0.073247	(1) web 2.0, 0.1017 (2) education, 0.03334 (3) social study, 0.031235 (4) music study, 0.024759	Art design, Web design 0.15812	(1)Web 2.0 in search, business, education and security. 0.2057 (2) Operation system 0.04523 (3) Web design 0.03603
Third	(1) mobile, iphone 0.039931 (2) Recherche_documentaire 0.032581 (3) social net working 0.030961	(1) bandslash fiction 0.1401 (2) fan fiction 0.076198 (3) supernatural fiction 0.039619	(1) Web 2.0, 0.066514 (2)E-education,0.037402 (3) photo, music and art et al. 0.036408	(1) Art design 0.11251 (2) Web design 0.02461 (3) Social Science reading 0.022975
Fourth	(1) interactive education 0.038008 (2) math 0.029208 (3) video game 0.029208 (4) enterprise 2.0, social net work, business 0.026649	(1) Imported favorites, toolbar favorites 0.036439 (2) film, movies 0.035648 (3) software, open source 0.03398 (4) electronics 0.029195	(1) java 0.06733 (2) drupal 0.03084 (3)wordpress-tips0.026212	(1) Bandslash fiction Fun fiction Supernatural fiction 0.0964 (2) Cooking 0.026649
Fifth	(1) bandslash fiction 0.1358 (2) doctor. who, drama 0.054392	(1) web design 0.042145 (2) java 0.039373 (3) architecture 0.034056	(1) Web development, java 0.081926 (2) iphone, operation system 0.027655	(1) social net work market 0.12473 (2) mobile 0.033668

5. Evaluation

In this section, community detection, topic model, and inference mechanism are evaluated against various criteria.

5.1 Community Evaluation

Two indices are used to evaluate the quality of the communities detected by the TTR-LDA-Community model: conductance and modularity (Leskovec, et al., 2010). Conductance is used to measure the conductivity for different classes or communities, which is defined as:

$$f(C) = \frac{s_c}{2m_c + s_c}, \quad (5)$$

C denotes the set of nodes in a community, m_c the number of edges in C, and $s_c = |\{(u,v) \mid u \in C \ \& \ v \notin C\}|$ is the number of all (u, v) that satisfy the condition. According to the definition of conductance, a community of high quality should have a small conductance value. Network Community Profile (NCP) is used to compute and display the score of the detected community, based on conductance. Leskovec et al. (2008) define an NCP plot as "the conductance value of the best conductance set of cardinality k in the entire network, as a function of k". Therefore, given graph G, the NCP plot can be expressed as:

$$F(k) = \min_{C \subset G, |C| = k} f(C), \quad (6)$$

where k is equal to the number of nodes in a community. Therefore the NCP plot contains a set of minimum conductance for communities with different sizes in G. The Whiskers tool is also adopted. Whiskers are defined as maximal sub-graphs that can be detached from the rest of the network by removing a single edge (Leskovec, et al., 2008). Rewired networks and rewired whiskers network are used to perform a comparative analysis. The rewired network is a random networks that has the same nodes and the same degree distribution as the original network (Leskovec, et al., 2008). In Figure 5, the 4 small figures in first row is to describe the conductance of original and rewired networks for the taggers' network in 4 time periods in 2008; the 4 small figures in second row is to describe the conductance of whiskers networks for taggers' network in 4 time periods in 2008. The conductance of the original network are drawn in blue line in the first row, rewired random network are drawn in red dashed line in the first row, the original whiskers network are drawn in blue line in the second row, and the rewired whiskers network are drawn in red dashed line in the second row.

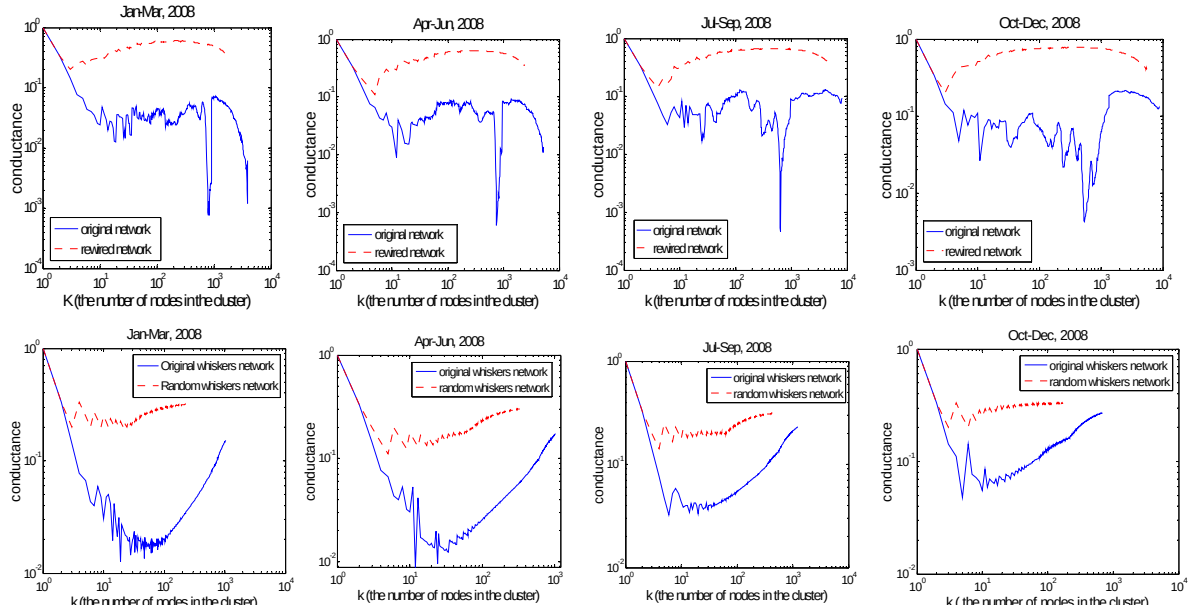

Fig. 5. NCP plot for the taggers' network

Figure 5 shows that, compared with the rewired network and the rewired whiskers, 1) the original network obviously displays a higher quality of communities with a much lower conductance; 2) the value of conductance as the function of community size in the original network and original whiskers show properties of a true large social networks, which tends to be "V" shaped; 3) the original whiskers have the best community granularity (the lowest conductance) between 10-100; and 4) the best community granularity of original networks is around 1000.

Modularity is one of the most widely used methods to evaluate the quality of a division of a network or graph into communities (Leskovec, et al., 2010; Newman, et al., 2004). It can be defined as,

$$M = \frac{1}{4m}(m_C - E(m_C)) \qquad (7)$$

where $E(m_C)$ is the expected number of edges of a random graph which has the same node degree sequence with C. The modularity values of detected communities in different time periods are shown in the Table 12.

Table 12
The modularity values of detected communities in different time periods

Time slice	Modularity	Time period	Modularity
2005	0.320031	2008, Jan-March	0.797043
2006	0.432471	2008, April-June	0.744134
2007	0.502286	2008, July-Sep	0.735286
2008	0.524738	2008, Oct.-Dec.	0.645894

As shown in Table 12, the quality of communities for 2008 is higher than those in earlier time periods. This is likely due to the fact that as the population grows, the community structure becomes more robust. Modularity is also higher in the short-term (for the four time periods of 2008) than in the long-term (all 2008 time periods merged together). This may indicate a higher quality of detected communities for these short time frames. This could be explained by taggers' bookmarking activities: in the short-term, tagging interest may be more focused. However, when these time periods are merged, the tagging interest of a single tagger can be seen across many domains, thereby making the clustering feature of the community weaker.

5.2 Topic Evaluation

To evaluate the inference mechanism used in the TTR-LDA-Community model, symmetrical Kullback–Leibler (sKL) divergence and entropy are calculated based on topic distribution over tags and resources.

5.2.1 sKL Divergence

First, the TTR-LDA model was used to compute the topic distribution over the 1,000 most popular resources in 2008. The most popular topics--the ones with the highest probabilities among those 1,000 popular resources—were about bandslash fiction, fan fiction, and supernatural fiction. However, the number of taggers who have created tags related to bandslash fiction is ranked in the middle of the top 1,000 ranked resources (500-600). One possible explanation is that topics related with web technology have many sub-topics such as operating systems, web design, web 2.0, computer technology, applications, etc., which can also be further divided into smaller topics: for example, programming language is divided into java, rubyonrails, drupal, and C++; application is also varied, for example math models, web application, and hardware. However, for fiction related topics, the theme is relatively concentrated; for example, http://pearl-o.livejournal.com/1000307.html is mainly about bandslash fiction and is very popular among taggers, but this kind of theme does not seem to contain many sub-topics in this context, compared to computer science topics.

Second, the topic distribution of the top 1,000 resources among communities was examined. The results are displayed in Table 13. This can be as interpreted as follows: the *i* in Topic *i(j)* means the *ith* topic in 1,000 most popular resource while *j* means the *ith* topic is ranked as *j* in all the 300 topics. As shown in Table 13, the top 20 ranked topics can be found in the 5 largest communities in different time periods. For each community, there exists at least one topic that is ranked top 10 out of the 1,000 most popular resources.

Table 13
The popular topics distribution over communities

	2008 1-3	2008 4-6	2008 7-9	2008 10-12
Largest	Topic 153(4) Topic 52(5) Topic 171(11) Topic 14(12)	Topic 153(4) Topic 61(7) Topic 76(16)	Topic 39(61) Topic 236(8) Topic 36(82) Topic 220(9) Topic 61(7)	Topic 236(8) Topic 153(4)
Second	Topic 100(3)	Topic 153(4) Topic 76(16)	Topic 194(6)	Topic 236(8) Topic 61(7)
Third	Topic 76(16) Topic 236(8)	Topic 29(1) Topic 100(3)	Topic 194(6) Topic 52(5)	Topic 48(15) Topic 94(6)
Fourth	Topic 38(52) Topic 48(15) Topic 236(8)	Topic 220(9) Topic 66(102)	Topic 153(4) Topic 52(5) Topic 171 (11) Topic 14(12)	Topic 29(1) Topic 100(3)
Fifth	Topic 29(1)	Topic 48(15) Topic 194(5)	Topic 220(9) Topic 14(11)	Topic 236(8) Topic 199(13)

The largest communities for each of the 2008 time periods were additionally analyzed using symmetric KL divergence and entropy. Symmetric KL divergence (*sKL*) was used to analyze the similarity of topic distribution among different resources and entropy was used to compute the features of the topic distribution of an individual resource (Rosen-zvi, et al., 2004). This analysis was done by selecting the largest communities for each time period within 2008 and selecting the set of resources bookmarked. Heat maps of *sKL* divergence are provided in Figure 6.

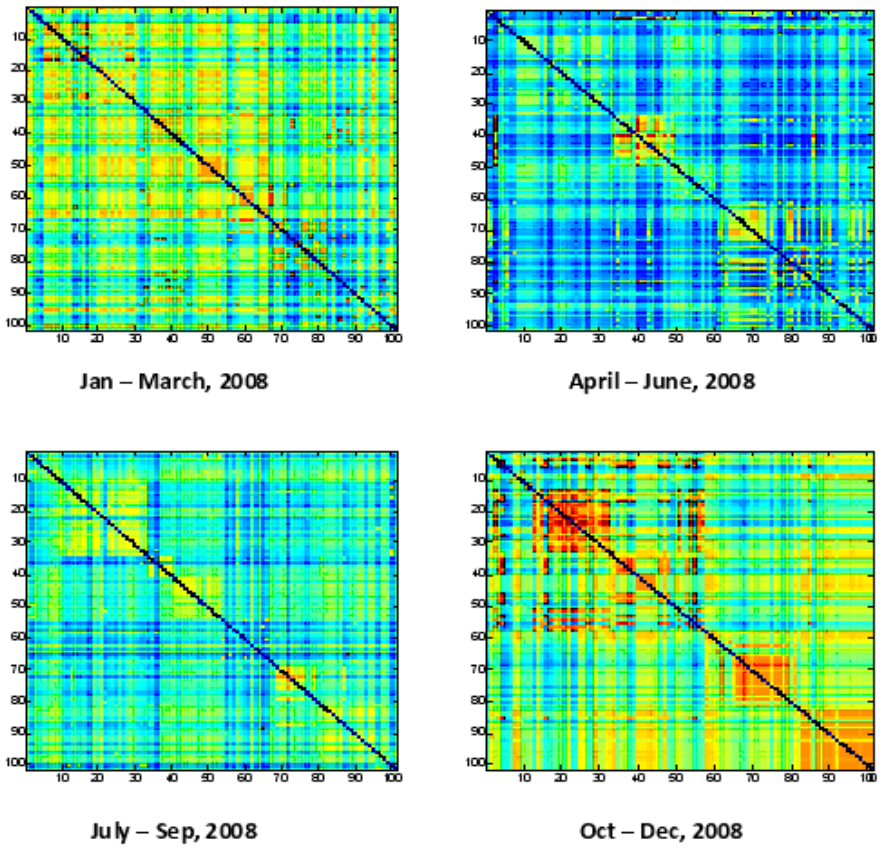

Fig. 6. Heat map of sKL divergence in 4 time periods

The heat map demonstrates which resources have a higher *sKL* divergence with other resources. The darker the color of the cell, the lower *sKL* divergence demonstrated by the corresponding resources—indicating similarity.

The resource pairs with *sKL* equal to zero were deleted. These pairs were predominately links that directed to the same webpage: for example, 'http://…/' and 'http://…/index.html'. Symmetric KL divergence was used to identify and combine these resources. It total, 205,708 resources were identified and recombined.

Interesting results can also be seen for those resource pairs whose *sKL* are relatively low in the largest communities for each time period. The 5 lowest *sKL* and corresponding resource pairs are listed in Tables 14. As demonstrated in Tables 14, resource pairs with low *sKL* have a relatively small number of shared tags. However, there is a high degree of similarity between the topic profiles. For example, the first resource pair in Jan – March, 2008 has 6 shared tags (with high frequencies) out of 110 distinct tags. However, all shared tags cover topics of military and regional crisis. Similarly, the fifth resource pair in Jan -- March, 2008 shares only 6 of 61 distinct tags, but shares a topical affinity for colonialism, imperialism and war. The fourth resource pair in April -- June, 2008 covers issues of art and web design (with only 8 shared out of 116 distinct tags); the third

resource pair in the dataset for July-- Sep, 2008 shares 6 tags (out of 78), predominately about programming languages and technologies.

Traditional similarity comparison methods focus on the proportion of common tags in total between two resources. In contrast, the *sKL* divergence model does not look at the overlap in individual tags, but similarities between the topics of the tags. For example, the fifth resource pair in April -- June, 2008 has 2 shared tags, 'blog' and 'culture'. Though 'blog' has a relative high frequency (it appears 18 times for the first resource and 5 for the second one), the term 'blog' alone is not an adequate representation of the resources. The tag 'culture' appears 8 times for the first resource and 7 times for the second one. Compared with the total number of tags associated with the two resources (173 times for the first resource and 271 times for the second resource), the frequency of "culture" in the two resources is relatively low. Therefore, a traditional similarity measure would find these resources to be dissimilar.

However, the current method identifies a high semantic similarity level. At the macro level, they are both about the same topics with different content: the first resource is an overview of New York City, including culture, blocks, government, restaurants, maps, and social communities; the second resource is about fashion and trends in Berlin and Toronto, including Street-fashion, city views, clothes, art, games, and lifestyle. At the micro level, they cover the same topics with different words: the first resource talks about blocks in the city while the second similarly discusses streets in Berlin; and the first resource is mainly about diet, health, social community, entertainment in New York City while the second resource is mainly about lifestyle, fashion, etc. There are also some representative resources pairs with high number of co-tags, such as "http://www.time.com/time/world/article/0" and "http://www.foxnews.com/story/0", which have 42 common tags and a *sKL* divergence of 3.168791. Though these resources pairs have high number of co-tags, their contents do not have much in common.

Table 14
sKL divergence for resources in 4 time periods in 2008

Jan -- March, 2008	*Number of Co-tag*	*sKL*
http://www.independent.co.uk/news/world/middle-east/our-reign-of-terror-by-the-israeli-army-811769.html http://www.ynetnews.com/articles/0	6	0.000116
http://www.w3.org/html/wg/html5/diff/ http://deseloper.org/read/2008/04/a-simple-modal/	0	0.000325
http://www.slideshare.net/Georgio_1999/how-to-scale-your-web-app http://mongrel.rubyforge.org/wiki/UploadProgress	4	0.000355
http://nubyonrails.com/articles/2006/08/17/memcached-basics-for-rails http://purefiction.net/mongrel_proctitle/	3	0.000355
http://www.independent.co.uk/news/fisk/robert-fisk-how-ireland-exorcised-the-ghost-of-empire-799514.html http://www.nybooks.com/articles/21311	6	0.000355
Mean sKL	-	2.0914
Max sKL	-	12.298439

April -- June, 2008	*Number of Co-tag*	*sKL*
http://funkatron.com/site/comments/google-app-engine-from-a-php-developers-perspective/ http://www.sitepen.com/blog/2008/06/05/easy-repeatable-buildingdeployment-of-pythondojo-projects/	4	0.000104
http://www.bobo.jp/index.html http://www.twist-cube.com/	5	0.000149
https://www.ecotonoha.com/index.html http://www.i-studio.co.jp/	3	0.000296
http://www.i-studio.co.jp/ http://scr.sc/	8	0.000325
http://nyc.everyblock.com/ http://streetclash.blogspot.com/	2	0.000388
Mean sKL	-	2.1645
Max sKL	-	12.095836

July-- Sep, 2008	*Number of Co-tag*	*sKL*
http://www.oreillynet.com/ruby/blog/2008/09/inspect_sql.html http://fuglyatblogging.wordpress.com/2008/10/	3	0.000053
http://www.oreillynet.com/ruby/blog/2008/09/inspect_sql.html http://fuglyatblogging.wordpress.com/2008/10/	5	0.000063
http://guides.rubyonrails.org/2_2_release_notes.html http://halcyon.rubyforge.org/	6	0.000104
http://nettuts.com/web-roundups/10-insanely-useful-django-tips/ http://www.blueskyonmars.com/projects/paver/	4	0.000129
http://jamesdonaghue.com/?p=40 http://websandbox.livelabs.com/	6	0.000163
Mean sKL	-	2.7719
Max sKL	-	12.757012

Oct -- Dec, 2008	*Number of Co-tag*	*sKL*
http://www.uniqlo.com/meets/ http://lucasmotta.com/splash/	5	0.000023
http://www.bencurtis.com/archives/2008/10/drag-and-drop-sorting-with-jquery-and-rails/ http://roman.flucti.com/a-test-server-for-rails-applications	3	0.000116
http://designm.ag/resources/photoshop-space-brushes/ http://thinkdesignblog.com/20-beautiful-free-serif-fonts.htm	3	0.000117
http://www.smashingmagazine.com/2008/07/15/70-beauty-retouching-photoshop-tutorials/ http://psdtuts.com/articles/web/best-of-the-web-october-2/	5	0.000155
http://www.coil-inc.jp/ http://www.uniqlo.com/meets/	6	0.000179
Mean sKL	-	2.5319
Max sKL	-	13.300175

5.2.2 Entropy

Additionally, the topic distributions for different tags can be used to assess the extent to which tags tend to have multiple meanings and are related to more than one topic. In order to assess this, the entropy of each tag's distribution was calculated: Table 15 display the 5 tags with the highest and lowest entropy in each of community. Most of the top-ranked tags are frequently used, formal, single English words, many of which are related to multiple topics. For example, "rest" can refer to a broad range of meanings; adjectives like "useful" can be used to describe different entities, such as open data, images, articles, etc, so they are probably related to many different topics, resulting in high entropies. In addition, "toberead" (others like "totag" and "toread") – tags expressing individual task assignment of the tagger – also rank high in entropy; since different taggers are interested in different topics, it is reasonable that those task-oriented tags are assigned to a broad range of resources and thus scattered among multiple topics. By contrast, tags with lowest entropy are mostly composite words with encoded character (e.g., %3A for space) embedded. Those tags are quite specific; such like "kink%3Aaliensmadethemdoit" and likely belong to a single topic.

Table 15
Entropy of representative resources in 4 time periods in 2008

Tags	High Entropy	Tags	Low Entropy
rest	1.8762	kink%3Aaliensmadethemdoit	0.1571
toberead	3.21999	challenge%3Asweetcharity	0.1571
maps	1.42443	cluetrain	0.22274
games	1.68399	lessons%2Fcourses	0.24026
porn	2.13202	uzumaki-kushina	0.4316
material	-1.52779	behavior%28s%29	-0.07603
useful	-1.4902	r%C3%A1di%C3%B3	-0.11113
slash	-3.86556	wc%3A16-20k	-0.13986
graphic	-1.60578	osg%C3%A9n%C3%A9ral	-0.20611
flash	-2.35009	43folders	-0.24877
home	-1.83082	kost%C3%BCme	-0.03913
download	-3.98765	lancia%3Agod	-0.09007
tools	-1.98148	it101gmu	-0.08511
tutorials	-2.12967	%C3%A9clairage	-0.06275
vc	-1.7841	qi4j	-0.14616
utility	-1.59601	%40webstandards	-0.0473
unix	-2.78639	onlinebusinessschool	-0.05165
books	-1.73143	myinstalleddownloads	-0.06275
art	-2.1005	author%3Aelandrialore	-0.14201
analytics	-1.82205	egovernance	-0.17943

5.3 Algorithm Evaluation

An additional means of evaluation is to evaluate the extent to which the models can be used for recommender systems. In order to assess this, the topic distribution is calculated on the training data (October to December, 2008) using LDA, TTR-LDA and TTR-LDA-Community models. One resource and five tags are recommended per post. The algorithm of resources recommendation is described in Figure 7.

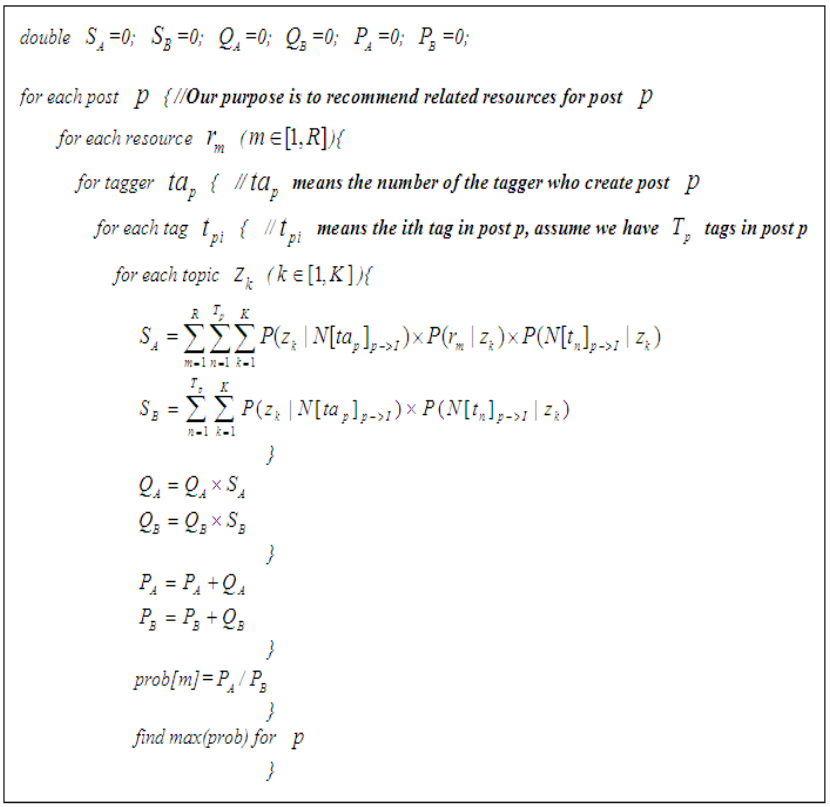

Fig. 7. Recommendation algorithm

The algorithm identifies the resource that has the largest probability of association for a particular post. Precision, Recall, and the F1-measure are used to evaluate the quality of the recommendation. The results for each model are presented in Table 16.

Table 16
Results of Precision, Recall, F1-measure on data-set from 2008 Oct.-Dec.

	Object	Precision	Recall	F1-measure
LDA	Tags for post	0.3502	0.2266	0.2752
TTR-LDA	Tags for post	0.3639	0.2271	0.2797
	Resource for post	0.2690	0.2690	0.2690
TTR-LDA-Community	Tags for post	0.3633	0.2321	0.2809
	Resource for post	0.2873	0.2873	0.2873

As is shown in Table 16, the TTR-LDA-Community model outperforms the other models in terms of precision, recall and F1-measure. This is in large part due to the inherent limitations within each model: LDA can only make predictions on the tags for each post, but cannot predict a resource for each post because LDA only calculates tag probabilities for topics. As shown, TTR-LDA-Community presents a high quality measure for identifying appropriate tags and resources based on a given post.

For further evaluation, a statistical significance test was conducted to compare between related models. This evaluation was done by first selecting 10% of data from each time period (2005, 2006, 2007, January-March 2008, April-Jun 2008, July-September 2008, and October-December 2008) as a test dataset. LDA, and TTR-LDA were applied to the data and used to recommend resources to posts (or tags to posts for LDA) and the performance of each model was compared using a t-test. The *p* values are summarized in Table 17.

Table 17
P-value for model comparison *t* test

	TTR-LDA-Community Vs TTR-LDA	**TTR-LDA-Community Vs LDA**
P-value	<0.05	<0.05
Deviation	+0.012873	+0.007464

As can be seen in Table 17, the p values are smaller than 0.05, indicating that there are significant differences in performance between the TTR-LDA-Community and other models. In addition, average deviation is used to evaluation the performance of the TTR-LDA-Community model in the following manner:

1. Use TTR-LDA-Community model to run all dataset separately and get result set RA{*ra1, ra2, ra3, …ra8*}. RA means the result set of TTR-LDA-Community, *rai* means the result for dataset *i*. Use test dataset to calculate as F1-measure for each result of RA.

2. Repeat Step 1 by using other models: LDA, TTR-LDA, and we got result set RB, RC for each model.

3. Calculate average deviation for the TTR-LDA-Community and each of other competed models using the formula listed below:

$$Average\ Deviation = \frac{\sum_{i=1}^{8}(ra_i - rb_i)}{8} \qquad (8)$$

We found that all deviations are greater than zero, which means that TTR-LDA-Community performs better than other models from a statistical viewpoint. This may be due to the utilization of the inferences mechanism and community detection algorithm as supervised functions, which can narrow the scope of the recommended items.

6. Conclusion

This paper proposes the TTR-LDA-Community model, an integrated model which combines TTR-LDA and Community detection using an inference mechanism. By applying this model to Delicious data, the paper observed the clustering of active taggers into communities; the topic distributions within communities; and the ranking of taggers, tags, and resources within these communities.

Using community detection, the paper observed the changes in community structure diachronically. Social tagging communities seem to experience a large intake of newcomers, significantly altering the participant base over time. In addition, quality of communities detected in short term is higher than that in long term, showing temporary stability followed by a sharp change in the topology of user network on Delicious over time. There also is evidence of a dominance of large communities: the largest of the communities incorporate the majority of participants, although many smaller communities exist. In addition, the changes of conductance as the function of community size display features of large true social network: communities of different sizes and qualities existed within the network. Changes in conductance over the whisker networks also reveal several core nodes in each community.

By examining topical features of communities, the paper finds large differences between communities. Some communities have a core group of topics, while the topic profiles for other communities are varied. Topics may also appear in a few communities simultaneously, but then split into sub-topics and scatter through many communities. In summary, topics seem to be a dynamic feature of communities: emerging, blending, and disappearing over time. The TTR-LDA-Community model provides a high quality algorithm for identifying both communities and the topics connecting these communities. This algorithm can be applied for other domains for community detection and profiling and for the provision of recommendation systems.

Acknowledgments

This work is supported by NIH-funded VIVO project (NIH grant U24RR029822).

Daifeng Li is funded by China National Natural Science Foundation (70971083), the Graduate Innovation Fund of Shanghai University of Finance and Economics (cxjj-2008-330), the 2009 Doctoral Education Fund of Ministry of Education in China (20090078110001) and the NIH VIVO project (uf09179).

Jie Tang is supported by the Natural Science Foundation of China (No. 60703059), Chinese National Key Foundation Research (No. 60933013), and National High-tech R\&D Program (No. 2009AA01Z138).

References

Andersen, R., Chung, F., & Lang, K (2008) Local partitioning for directed graphs using pagerank. Internet Mathematics, 51, 3-22.

Blei, D. M., NG, A.Y., & Jordan, M., I (2003) Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Clauset, A., Newman, M., E., J., & Moore, C (2004) Finding community structure in very large network. *Physical Review E, 70*, 066111.

Chang, J., & Blei, D (2009) Relational topic models for document networks. Proceedings of Conference on AI and Statistics AISTATS'09.

Cohn, D., & Hofmant, T (2001) The missing link – a probabilistic model of document content and hypertext connectivity. In Advances in Neural Information Processing Systems, 13, 430-436.

Dietz, L., Bickel, S., & Scheffer, T (2007) Unsupervised prediction of citation influences. In Proceedings of the 24th international conference on Machine learning, 233–240.

Erdos, P., & Renyi, A (1959) On random graphs. I, *Publicationes Mathematicae, 6*, 290-291.
Flake, G., Tarjan, R., & Tsioutsiouliklis, K (2003) Graph clustering and minimum cut trees. *Internet Mathematics*, *14*, 385–408.
Erosheva, E., Fienberg, S., & Lafferty, J (2004) Mixed membership models of scientific publications. Proceedings of the National Academy of Sciences, 101, 5220–5227.
Girvan, M., & Newman, M. E. J (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*12, 7821–7826.
Leskovec, J., Lang, K., Mahoney, M. (2010) Empirical comparison of algorithms for network community detection. In *proceedings of the nineteenth International World Wide Web Conference.* North Caroline, USA.
Leskovec, J., Lang, K., Dasgupta, A., & Mahoney, M (2008) Statistical properties of community Structure in Large Social and Information Networks. In WWW '08: Proceedings of the 17th International Conference on World Wide Web, pages 695–704.
Leskovec, J., Lang, K., Dasgupta, A., & Mahoney, M (2008) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. arXiv:0810.1355v1.
Li, D, He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., & Li, J (2010) Community-based topic modeling for social tagging. The 19th ACM International Conference on Information and Knowledge Management (CIKM2010), Oct 26-30, Toronto, Canada.
Liu, Y., Niculescu-Mizil, A., & Gryc, W (2009) Topic-link LDA: joint models of topic and author community. Paper presented at Proceedings of the 26th Annual International Conference on Machine Learning.
Lu, C., Hu, X., Chen, Y., Park, J., He, T., Li, Z. (2010). The topic-perspective model for social tagging systems. The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 7.
Mccallum, A., Corrada-emmanuel, A., & Wang, X (2004) The author-recipient-topic model for topic and role discovery in social networks: experiments with enron and academic email. *Technical Report UM-CS-2004-096.*
Milgram, S (1967) The small world problem, *Psychology Today, 2,* 60–67.
Nallapati, R., & Cohen, W (2008) Link-PLSA-LDA: A new unsupervised model for topics and influence in blogs. Proceedings of International Conference on Weblogs and Social Media ICWSM'08 pp. 84–92.
Newman, M. E. J (2001) The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America, 98*2, 404-409.
Newman, M. E. J (2003) The structure and function of complex networks. *Siam Review* 45, 167–256 .
Newman, M. E. J, & Girvan, M (2004). Finding and evaluating community structure in networks. physical review E, 69:026113.
Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D (2004) Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*9, 2658-2663.
Rosen-zvi, M., Griffiths, T., Steyvers, M., & Smyth, P (2004) The author-topic model for authors and documents. *In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence pp. 487-494. Virginia: AUAI Press.*
Si, X., & Sun, M (2009) Tag-LDA for scalable real-time tag recommendation. Journal of Computational Information Systems. 1, 23-30.
Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008) ArnetMiner: Extraction and mining of academic social networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining SIGKDD'2008. pp.990-998
Tang, J., Jin, R., & Zhang, J (2008) A topic modeling approach and its Integration into the random walk framework for academic search. In Proceedings of 2008 IEEE International Conference on Data Mining ICDM'2008 pp. 1055-1060. Washington, DC: IEEE Computer Society.