Academic Network Analysis: A Joint Topic Modeling Approach

Zaihan Yang Liangjie Hong Brian D. Davison
Department of Computer Science and Engineering, Lehigh University
Bethlehem, PA 18015
Email: {zay206, lih307, davison}@cse.lehigh.edu

***Abstract*—We propose a novel probabilistic topic model that jointly models authors, documents, cited authors, and venues simultaneously in one integrated framework, as compared to previous work which embeds fewer components. This model is designed for three typical applications in academic network analysis: the problems of expert ranking, cited author prediction and venue prediction. Experiments based on two real world data sets demonstrate the model to be effective, and it outperforms several state-of-the-art algorithms in all three applications.**

Keywords: Topic Models, Expert Ranking, Prediction, Evaluation

I. Introduction

Social network research has attracted the interests of many researchers, not only in analyzing online social media applications, such as Facebook and Twitter, but also in providing comprehensive services in the domain of scientific research. We define an *academic network* as a kind of social network which integrates scientific factors, such as authors, papers, publishing venues, and their relationships. With the rapid development of online digital libraries, the proliferation of large quantities of scientific literature provides us abundant opportunity to extract the textual content of scientific factors (i.e., publishing papers) as well as their mutual relationships (citation, coauthorship), and therefore stimulates the emergence of many applications that are particularly important in academic domain (in mining and analyzing academic networks), such as expert ranking, citation prediction, cited author prediction, venue prediction, etc.

Generative topic modeling has emerged as a popular unsupervised learning technique for content representation in large document collections. This kind of generative model was first envisioned for pure contextual analysis while ignoring the linkage structure among text data. Representative models of this type of analysis (e.g., [10], [2]) exploit the co-occurrence patterns of words in documents and unearth the semantically meaningful clusters of words (as topics). Researchers have since added extensions to model authors' interests [23], providing a framework for answering questions and making predictions at the level of authors rather than documents, and in a variety of other aspects, such as incorporating link structures and integrating additional context information.

Despite such recent developments (which we review in Section 2), limitations are still present. It is widely acknowledged that one of most prominent advantages of generative topic modeling is that it provides us a flexible and extensible framework to exploit the underlying latent structures over text data as well as their mutual connections. In the academic network, we have multiple kinds of scientific factors and connections; however, most of the previous work considers one aspect of several factors while ignoring some others.

In this paper, we provide a framework that can jointly model authors, papers, cited authors, and venues in one unified model. We name the model as the Author-Citation-Venue topic model (abbreviated as ACVT model), in which we link authors to observed words, cited authors and venues via latent topics. We hypothesize that such a joint modeling has multiple advantages. First of all, this model provides a more comprehensive framework to fully utilize the content words of documents and combines them with other useful contextual information: authors, cited authors and venues. It therefore directly models documents' content relevance, authors' interests, authors' influence, and venues' influence in one model, all of which are important instructive evidence in supporting academic network based applications, such as expert ranking, cited author prediction, and venue prediction. Missing the integration of one sort of contextual information, some certain kind of application would become impossible, for example, if the topic-venue association is not explored, we cannot make valid venue predictions. Our model therefore can be applied in a wider range of applications than previous work. Moreover, incorporating additional contextual and linkage information can help to identify more coherent and complete latent structures over multiple facets. In the ACVT model, we assume that we can achieve better topic-related associations for authors, cited authors and venues when we simultaneously model them together, and such associations with greater coherency are believed to be able to further improve the performance of multiple applications.

In summary, we make the following contributions in this paper:

- We propose a generative model that incorporates multiple facets of academic network: authors, papers, venues and cited authors in an integrated fashion.
- We apply our model, and provided solutions to three tasks in the academic domain: expert ranking, cited author prediction and venue prediction.
- Experiments based on two real world data sets demon-

strate our model to be effective on all three tasks, significantly outperforming several state-of-the-art algorithms.

The rest of the paper is organized as follows. We overview related work in section two. In section three, we introduce the ACVT model as well as the parameter estimation method. We address three applications of this model in section four. In section five, we introduce experimental setup and provide analysis on experimental results. We conclude and outline possible future work in section six.

II. Related Work

A. Author Topic Modeling

Generative topic modeling is a popular unsupervised learning technique for topic-related content representation. Initially, this kind of generative modeling was utilized in pure contextual analysis. Two representative models of this kind, Probabilistic Latent Semantic Analysis (PLSA) [10] and Latent Dirichlet Allocation (LDA) [2], exploit co-occurrence patterns of words in documents and unearth the semantic clusters (topics) of words. In those proposed models, each document would be regarded as a mixture over multiple latent topics.

The original PLSA and LDA idea of document topic modeling has been extended to include modeling of authors' interests. The very first work in this direction is that of Rosen-Zvi et al. [23], which simultaneously models the content of documents and the interests of authors, such that the mixture weights for different topics would be determined by the authors of the documents.

Most recently, a number of models that extend the original idea of LDA and ATM have been proposed, most of which contribute in the direction of incorporating additional contextual information and integrating linkage structures. Link-LDA [4], Pairwise-LDA and Link-PLSA-LDA [20] are three representative topic models that extend PLSA and LDA by integrating citation linkages among papers into topic modeling. However, in these three efforts, no author information has been considered, and the citation prediction is made based upon pairs of papers, which is quite different from the model we propose in this paper that particularly emphasizes the interests and influence of authors.

Several representative works have been proposed to extend ATM. The Author-Conference-Topic (ACT) [25] model adds contextual information, the publishing venues of papers, to represent venues' influence over topics. The Author-Conference-Topic-Connection [28] model extends [25] by introducing an additional latent variable 'subject', from which the confereneces (venues) and topics can be respectively generated. The Citation-Author-Topic (CAT) [27] model directly models the cited authors' information, such that authors' influence over other authors can be considered. As a further extension to the CAT model, the Context Sensitive Topic Models [13] introduces a learning mechanism that can dynamically determine the citation context windows, and to associate terms within citation context windows to cited authors. Our proposed model, the ACVT model, can be regarded as a further extension and combination of the ACT and CAT model, in that we jointly model both the venue and the cited authors information, as compared to ACT which only considers venues, and CAT and the Context Sensitive model that only consider citations.

There are also other topic models which emphasize different aspects of contribution. Liu et al. [15] proposed a model that can jointly model topics, author communities and link information for author community detection. Johri et al. [12] introduced a model that can relax the 'bag-of-words' assumption and can automatically identify multi-word phrases into modeling; Mei et al. [18] conducted temporal author topic analysis, and Song et al. [24] built topic models to disambiguate names. Mei et al. [17] incorporated network regularization technique into an extended version of PLSA. Our ACVT model distinguishes itself from all the work mentioned above by its model design focus and applications.

B. Applications: Expert Ranking, Cited Author and Venue Prediction

Expert ranking has blossomed since the advent of the TREC Enterprise Track initiated in 2005, and the rapid development of online academic search engines, such as ArnetMiner and Microsoft Academic Search. Given a user query, the task of expert ranking basically involves identifying and ranking a list of researchers based on their expertise in that query-specific domain. Two categories of approaches have been the research focus in the past years: the pure content analysis based approach [1], [16], [5], which emphasizes evaluating authors' expertise by measuring the relevance between their associated documents and the query, and the social network based approach [3], [26], [30], [6], [11], which evaluates authors' expertise by exploiting the social interaction of authors and other scientific facets, such as their co-authorships, their citations to other papers/authors, etc. Few prior works directly make use of topic modeling results for expert ranking. The CAT, ACT and ACTC models are the three most representative works we have identified.

Citation prediction has long been a research topic as a specific application in link prediction (e.g., [9], [8]). However, most of them predict citations among papers, and few use topic modeling results. In our paper, we focus on predicting the potential cited authors given a new document, which has seldom been explored by previous work except the work of Tu et al. [27].

In venue recommendation, a ranked list of venues is generated to which a given paper could be submitted. Three prior works [14], [22], [29] particularly address such a problem, however, none of them makes use of a topic modeling approach.

III. Model

Before presenting the model, we first introduce some notation. Suppose W, D, A, V indicate the size of the word vocabulary, the number of papers, the number of authors (cited authors), and the size of venues in the corpus respectively. $\boldsymbol{a_d}$, $\boldsymbol{c_d}$ and N_d denote the set of authors, the set of cited authors, and the number of position-based words in paper d. T denotes

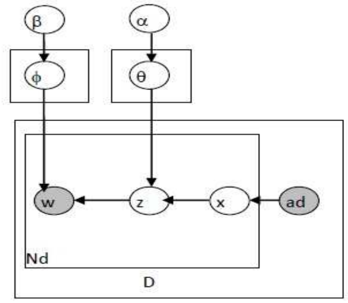

Fig. 1. Graphical Model for the original Author-Topic Model

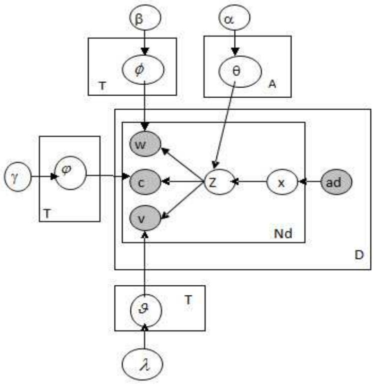

Fig. 2. Graphical Model for the Author-Citation-Venue-Topic Model

the number of latent topics predefined. We further suppose that there exists a $A \times T$ author-topic distribution matrix θ indicating the distribution of authors over topics, a $T \times W$ topic-word distribution matrix ϕ indicating the probability distribution of topics over words, an $T \times A$ distribution matrix φ indicating the distribution of topics over cited authors, and a $T \times V$ distribution matrix ϑ indicating the distribution of topics over venues. $\boldsymbol{z}$, $\boldsymbol{x}$, $\boldsymbol{m}$, $\boldsymbol{s}$ are random variables, representing the topic assignment, author assignment, cited author assignment and venue assignment for each word. α, β, γ, and λ are the Dirichlet prior hyper-parameters that determine θ, ϕ, φ, and ϑ respectively. We list the detailed notation in Table I.

A. Model Description / Generative Process

We depict the graphical model of ACVT in Figure 2 as compared to the original Author-Topic Model shown in Figure 1. As indicated, the graphical model is composed of six plates. Besides the four plates representing Topics, Authors, Documents and words in each document, ACVT introduces two additional plates, representing the topic-cited author association and topic-venue association respectively. As we can see, authors, words, cited authors and venues are all connected via the latent topics. Note that even though the author list and cited author list for any given paper d are assumed to be known, the exact author and cited author assignment for each particular word in paper d are unknown.

TABLE I
NOTATION

Symbol	**Size**	**Description**
W	scalar	size of word vocabulary
D	scalar	number of papers
A	scalar	number authors (cited authors)
V	scalar	number of venues
T	scalar	number of latent topics
N_d	scalar	the number of words in paper d
A_d	scalar	the number of authors of paper d
C_d	scalar	the number of cited authors of paper d
N	scalar	the number of words in corpus
		Observed Data
$\boldsymbol{a_d}$	$\lvert\boldsymbol{a_d}\rvert$	the set of authors of paper d
$\boldsymbol{c_d}$	$\lvert\boldsymbol{c_d}\rvert$	the set of cited authors of paper d
$\boldsymbol{w_d}$	$\lvert\boldsymbol{w_d}\rvert$	the words lists of paper d
v_d	1	the publishing venue of paper d
$\mathcal{A}$	A	the set of authors (cited author) in corpus
$\boldsymbol{w}$	N	the set of word tokens in corpus
$\mathcal{V}$	V	the set of venues in corpus
		Hyper-Parameters
α	$1 \times T$	Dirichlet prior for θ
β	$1 \times T$	Dirichlet prior for ϕ
γ	$1 \times T$	Dirichlet prior for φ
λ	$1 \times T$	Dirichlet prior for ϑ
		Random Variables
θ	$A \times T$	distribution of authors over topics
ϕ	$T \times V$	distribution of topics over words
φ	$T \times A$	distribution of topics over cited authors
ϑ	$T \times C$	distribution of topics over venues
$\boldsymbol{z_{di}}$	$1 \times T$	topic assignments for i^{th} word in paper d
$\boldsymbol{x_{di}}$	$1 \times \lvert a_d\rvert$	author assignments for i^{th} word in paper d
$\boldsymbol{m_{di}}$	$1 \times \lvert c_d\rvert$	cited author assignments for i^{th} word in paper d
$\boldsymbol{s_{di}}$	scalar	venue assignments for ith word in paper d

Within ACVT, each author is associated with a multinomial distribution over topics θ, and each topic is associated with a multinomial distribution over words ϕ, a multinomial distribution over cited authors φ, and a multinomial distribution over venues ϑ. Moreover, θ, ϕ, φ and ϑ follow a Dirichlet distribution with respect to the Dirichlet prior α, β, γ, and λ respectively.

The design of the ACVT model captures the intuition of people writing a paper. Normally, when authors start to write a paper, they should have known what they are going to write about, namely, the topics of their paper. Based upon the chosen topics, they will then choose the exact words to use to represent their intended topics, figure out other related works and their corresponding authors to cite, and determine where to submit this paper. We assume that one paper may address multiple topics, and can be co-authored by more than one author, and that each of the co-authors may have different weights of contributions to a specific topic.

The generative process of the ACVT model can be described as follows. We first sample the author-topic, topic-word, topic-cited author and topic-venue distributions based on the four Dirichlet prior hyper-parameters. Suppose we know the author lists of papers; then for each word in a given paper, we would first draw an author from its author list, then conditioned on this drawn author and his associated author-topic distribution, we sample one topic, based upon which, we further sample the cited author, venue and word according to their topic-related distributions independently.

Under this generative process, the likelihood of the corpus $\boldsymbol{w}$, conditioned on θ, ϕ, φ, and ϑ is:

$$
\begin{aligned}
& p(\boldsymbol{w}|\theta,\phi,\varphi,\vartheta,\mathcal{A},\mathcal{V}) \\
&= \prod_{d=1}^{D} p(\boldsymbol{w_d}|\theta,\phi,\varphi,\vartheta,\boldsymbol{a_d},\boldsymbol{c_d},v_d) \\
&= \prod_{d=1}^{D}\prod_{i=1}^{N_d}\frac{1}{A_d}\sum_{a\in\boldsymbol{a_d}}\sum_{t=1}^{T}\sum_{c=1}^{C_d}\varphi_{tc}\vartheta_{tv_d}\phi_{tw_{di}}\theta_{at}
\end{aligned}
$$

B. Parameter Inference and Estimation

The primary inference goal of our ACVT model is to estimate the posterior distribution of two sets of unknown random variables: (1) the distribution of θ, ϕ, φ and ϑ, and (2) the topic, author, cited author and venue assignments for each word w_{di}: z_{di}, x_{di}, m_{di}, s_{di}.

$$
p(\theta,\phi,\varphi,\vartheta,\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}|D^{train},\alpha,\beta,\gamma,\lambda) \tag{1}
$$

where, $\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}$ indicate the topic, author, cited author and venue assignments for all word tokens in corpus.

Even though calculating these posterior distributions is intractable for exact inference, various approximate inference models have been employed to estimate these posterior distributions in hierarchical Bayesian models, including variational inference [2], expectation propagation[19], and Markov chain Monte Carlo (MCMC) schemes. In this paper, we use Gibbs Sampling [7], a special case of the MCMC approximation scheme, which is not necessarily as computationally efficient as variational inference and expectation propagation, but is unbiased and simple to implement.

The entire inference process involves two steps. Firstly, we obtain an empirical sample-based estimate of $p(\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}|D^{train},\alpha,\beta,\gamma,\lambda)$ using Gibbs Sampling, and then secondly, we infer the posterior distribution of θ, ϕ, φ, and ϑ based upon $\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}$, by exploiting the fact that the Dirichlet distribution is conjugate to the multinomial distribution.

1). Gibbs Sampling for $\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}$

Using Gibbs Sampling, we construct a Markov chain, in which the transition between two successive states results from repeatedly drawing the four-tuple $< z,x,m,s >$, i.e., the assignment of topic, author, cited author, and venue for each word as a block from its distribution, conditioned on all other variables. Such a sampling process would be repeated until it finally converges to the posterior distribution of $\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}$. The corresponding updating equation for this blocked Gibbs Sampler can be defined as:

$$
\begin{aligned}
& p(x_{di}=a,z_{di}=t,m_{di}=c,s_{di}=v|\boldsymbol{U_{known}}) \\
\propto\ & \frac{C^{AT}_{at,-di}+\alpha}{\sum_{t'}C^{AT}_{at',-di}+T\alpha}\frac{C^{TW}_{tw,-di}+\beta}{\sum_{w'}C^{TW}_{tw',-di}+N\beta} \\
\times\ & \frac{C^{TC}_{tc,-di}+\gamma}{\sum_{c'}C^{TC}_{tc',-di}+A\gamma}\frac{C^{TV}_{tv,-di}+\lambda}{\sum_{v'}C^{TV}_{tv',-di}+V\lambda}
\end{aligned}
$$

$$
\begin{aligned}
& \boldsymbol{U_{known}} \\
&= \{w_{di}=w,\boldsymbol{z_{-di}},\boldsymbol{x_{-di}},\boldsymbol{m_{-di}},\boldsymbol{s_{-di}},\boldsymbol{w_{-di}},\boldsymbol{a_d},v_d,\alpha,\beta,\gamma,\lambda\}
\end{aligned}
$$

where C^{AT} represents the author-topic count matrix, and $C^{AT}_{at,-di}$ is the number of words assigned to topic t for author a excluding the topic assignment to word w_{di}. Similarly, C^{TW} represents the topic-word count matrix, and $C^{TW}_{tw,-di}$ is the number of words from the wth entry in word vocabulary assigned to topic t excluding the topic assignment to word w_{di}; C^{TC} represents the topic-cited author count matrix, and $C^{TC}_{tc,-di}$ is the number of cited authors assigned to topic t excluding the topic assignment to word w_{di}, and finally, C^{TV} represents the topic-venue count matrix, and $C^{TV}_{tv,-di}$ is the number of venues assigned to topic t excluding the topic assignment to word w_{di}. Moreover, $\boldsymbol{z_{-di}},\boldsymbol{x_{-di}},\boldsymbol{m_{-di}},\boldsymbol{s_{-di}}$, and $\boldsymbol{w_{-di}}$ stand for the vector of topic, author, cited author and venue assignment and the vector of word observations in the corpus except for the i^{th} word in the d^{th} document respectively.

In implementing this Gibbs Sampling, we simply need to keep track of the four matrices (C^{AT}, C^{TW}, C^{TC}, C^{TV}). By initially assigning words to randomly chosen topic, authors, cited authors and venues, we repeatedly apply this equation to each word in corpus, until finally converged.

2). The Posterior on $\theta,\phi,\varphi,\vartheta$

After we obtain the approximated estimation of $\boldsymbol{z},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}$, the posterior distribution of $\theta,\phi,\varphi,\vartheta$ can be directly computed by exploiting the fact that the Dirichlet distribution is conjugate to the multinomial distribution, and therefore we have:

$$
\begin{aligned}
\theta|\boldsymbol{z},\boldsymbol{x},D^{train},\alpha &\sim Dirichlet(C^{AT}+\alpha) && (2)\\
\phi|\boldsymbol{z},D^{train},\beta &\sim Dirichlet(C^{TW}+\beta) && (3)\\
\varphi|\boldsymbol{z},\boldsymbol{m},D^{train},\gamma &\sim Dirichlet(C^{TC}+\gamma) && (4)\\
\vartheta|\boldsymbol{z},\boldsymbol{s},D^{train},\lambda &\sim Dirichlet(C^{TV}+\lambda) && (5)
\end{aligned}
$$

We can then estimate the posterior mean of $\theta,\phi,\varphi,\vartheta$ by following:

$$
\begin{aligned}
E[\theta_{at}|\boldsymbol{z},\boldsymbol{x},D^{train},\alpha] &= \frac{C^{AT}_{at}+\alpha}{\sum_{t'}C^{AT}_{at'}+T\alpha} && (6)\\
E[\phi_{tw}|\boldsymbol{z},D^{train},\beta] &= \frac{C^{TW}_{tw}+\beta}{\sum_{w'}C^{TW}_{tw'}+W\beta} && (7)\\
E[\varphi_{tc}|\boldsymbol{z},\boldsymbol{m},D^{train},\gamma] &= \frac{C^{TC}_{tc}+\gamma}{\sum_{c'}C^{TC}_{tc'}+C\gamma} && (8)\\
E[\vartheta_{tv}|\boldsymbol{z},\boldsymbol{s},D^{train},\lambda] &= \frac{C^{TV}_{tv}+\lambda}{\sum_{v'}C^{TV}_{tv'}+V\lambda} && (9)
\end{aligned}
$$

IV. Application

We introduce in this section three main applications related with academic network analysis that can be solved by applying our ACVT model.

A. Expert Ranking

The problem of expert ranking is equivalent to the problem of finding experts. The ultimate goal of an expert finding task is to identify people who have relevant expertise to a specific topic of interest. In the academic research environment, estimating a researcher's reputation (contribution) and further ranking academic researchers is of great importance as it can offer support when making decisions about researchers' job promotion, project funding approval, paper review assignment, as well as scientific award assignment.

1) Rank experts by Topic Models: Based on the learning results from the ACVT model, we obtain four distributions: θ, ϕ, φ and ϑ. Suppose we are given a query q, composed of a set of words $\boldsymbol{w}$, then for any given author a in the corpus, the probability of having author a being relevant with the query q, i.e, the expertise of the author a in domain q, can be computed under our ACVT model as:

$$\begin{aligned} p_{TM}(a|q) &\propto p_{TM}(q|a) \quad (10)\\ &= \prod_{w \in q} p(w|a) \\ &= \prod_{w \in q} p(w|a_a)p(w|a_c) \sum_{v \in V(a)} p(w|v) \end{aligned}$$

where $p(w|a_a)$ represents the probability of author a generating word w as an author; $p(w|a_c)$ represents the probability of author a being cited by word w; $p(w|v)$ represents the probability of venue v generating word w. We consider all the publishing venues $V(a)$ of a to evaluate the relevance between author a with respect to word w from the venue aspect of view.

Based upon the learning results from ACVT, we can further have:

$$\begin{aligned} p(w|a_a) &= \sum_t p(w|z)p(z|a_a) = \sum_t \phi_{tw}\theta_{a_a t} \quad (11)\\ p(w|a_c) &= \sum_t p(w|z)p(z|a_c) \quad (12)\\ &\propto \sum_t p(w|z)p(a_c|z) = \sum_t \phi_{tw}\varphi_{ta_c} \\ p(w|v) &= \sum_t p(w|z)p(z|v) \quad (13)\\ &\propto \sum_t p(w|z)p(v|z) = \sum_t \phi_{tw}\vartheta_{tv} \end{aligned}$$

As a result, we can compute $p_{TM}(a,q)$ by:

$$p_{TM}(a|q) \propto \prod_{w \in q} (\sum_t \phi_{tw}\theta_{a_a t})(\sum_t \phi_{tw}\varphi_{ta_c})(\sum_{v \in V(a)} \sum_t \phi_{tw}\vartheta_{tv}) \quad (14)$$

2) Combining with Language Model and Random-walk: We are also interested in examining whether we can achieve better performance when combining the results obtained from Topic Modeling with that of using a language model based approach and a random walk based approach, the two other representative approaches in evaluating researchers' expertise.

To evaluate the relevance between an author a with a query, we can construct a virtual document F_a of author a by concatenating all the publishing papers of author a, and thus the relevance between author a and query q would be equivalent to the relevance of document F_a with query q. Under the standard language model with Jenilek-Mercer smoothing, the probability can be computed by:

$$\begin{aligned} p_{LM}(a|q) &= p_{LM}(F_a|q) \\ &= \prod_{w \in q} \{(1-\lambda)\frac{n(w,F_a)}{n(F_a)} + \\ &\qquad \lambda \frac{\sum_{F_{a'}} n(w,F_{a'})}{\sum_{F_{a'}} n(F_{a'})}\} \quad (15) \end{aligned}$$

A random-walk based algorithm directly models the interaction among network nodes. In this paper, we construct a heterogeneous academic network (as shown in Figure 3, which follows the network design mentioned in paper [25]) which is composed of three kinds of academic factors: authors, papers and venues, and their mutual relationships: $G = (V_a \cup V_d \cup V_v, E_{ad} \cup E_{dd} \cup E_{cd})$. V_a, V_d and V_v represents the collection of authors, papers and venues respectively. Based on our definition, $(d_i, d_j) \in E_{dd}$ if paper d_i cites paper d_j. We further represent each undirected edge into two directed edges in bipartite graphs, and therefore we have both $(a_i, d_j) \in E_{ad}$ and $(d_j, a_i) \in E_{ad}$ if paper d_j is written by author a_i. Similarly, $(v_i, d_j) \in E_{vd}$ and $(d_j, v_i) \in E_{vd}$ if paper d_j is published in venue v_i.

The transition probability between any two nodes in the network is determined by two parameters: the type-based transition parameter $\lambda_{t_1 t_2}$, which determines the probability when the random surfer transfers from node of type t_1 to node of type t_2. The second parameter $p(n_1|n_2)$ determines the transition probability between any two specific nodes, no matter what type of the nodes they are. Under this definition, if the random surfer transfers from node n_1 of type t_1 to node n_2 of type t_2, the transition probability would be $\lambda_{t_1 t_2} p(n_2|n_1)$.

Given this academic network, we apply a PageRank-like [21] propagation algorithm over it to achieve the ranking score for each 'author' node. Suppose the PageRank score of each node n_i is denoted as $r(n_i)$, and then it can be computed by:

$$r(n_j) = \frac{d}{|V|} + (1-d) * \sum_{(n_i,n_j) \in E} \lambda_{t(n_i)t(n_j)} p(n_j|n_i) \quad (16)$$

where $|V|$ is the total number of nodes in the network, and $t(n_i)$ indicates the type of node n_i.

We adopted two methods to combine the ranking performance of topic modeling, language model and random-walk based PageRank. To linearly combine them, the final ranking score of an author a for a given query q can be computed as:

$$p_{Final}(a|q) = \alpha p_{TM}(a,q) + \beta p_{LM}(a,q) + \gamma r(a) \quad (17)$$

where, α, β and γ, satisfying $\alpha+\beta+\gamma = 1$, are the parameters that need to be tuned.

We can also multiply the results obtained from the three methods, which results in the final ranking score presented as:

$$p_{Final}(a|q) = p_{TM}(a|q) \times p_{LM}(a|q) \times r(a) \tag{18}$$

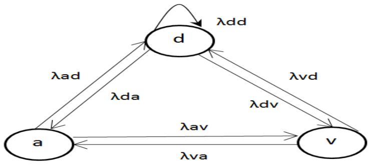

Fig. 3. Heterogeneous Academic Network

B. Cited Author Prediction

We examine in this task the capability of our model in predicting the authors that a given paper might cite in the future. Instead of predicting the cited papers directly, we predict the cited authors. This has applications in real life, since we sometimes follow some authors, especially some authors who are of high reputation in a certain field, and then by going through their publications, an author can locate the most recent and relevant papers to cite.

Suppose we are now given a new document, represented by $\boldsymbol{W_d}$, and suppose we know its author lists $\boldsymbol{a_d}$. In order to predict the potentially cited authors, we need to compute the probability of $p(c|\boldsymbol{w_d})$, the probability of generating c given words $\boldsymbol{W_d}$ and author lists $\boldsymbol{a_d}$. This probability can be computed by making use of the distributions we learned from the training set. We have:

$$\begin{aligned}
p(c|\boldsymbol{W_d}) &= \sum_z \int p(c,z,\theta|\boldsymbol{W_d})d\theta = \sum_z \int \frac{p(c,z,\theta,\boldsymbol{W_d})}{p(\boldsymbol{W_d})}d\theta \\
&\propto \sum_z \int p(c,z,\theta,\boldsymbol{W_d})d\theta \\
&= \sum_z \int p(\boldsymbol{W_d}|z)p(c|z)p(z|\theta)d\theta \\
&= \sum_z p(\boldsymbol{W_d}|z)p(c|z)\int p(z|\theta)d\theta \\
&= \prod_{w\in\boldsymbol{W_d}} [\sum_z \sum_{a\in\boldsymbol{a_d}} p(w|z)p(c|z)\int p(z|\theta)d\theta] \\
&\approx \prod_{w\in\boldsymbol{W_d}} [\frac{1}{|a_d|}\sum_{k=1}^{K}\sum_{a\in\boldsymbol{a_d}} \theta_{ak}\phi_{kw}\varphi_{kc}] \qquad (19)
\end{aligned}$$

where, $a \in \boldsymbol{a_d}$.

C. Venue Prediction

In the task of venue prediction, we aim to predict the potential publishing venue given a paper with both its content and author lists provided. This task is of importance to some researchers, especially researchers that are new to a domain. They may find it difficult to decide where to submit after they finish their work. Similarly, in order to predict the potential venue, we need to compute the probability of $p(v|\boldsymbol{w_d})$. The derivation process is similar to that of the cited author prediction, and therefore we have:

$$\begin{aligned}
p(v|\boldsymbol{W_d}) &= \sum_z \int p(v,z,\theta|\boldsymbol{W_d})d\theta = \sum_z \int \frac{p(v,z,\theta,\boldsymbol{W_d})}{p(\boldsymbol{W_d})}d\theta \\
&\propto \sum_z \int p(v,z,\theta,\boldsymbol{W_d})d\theta \\
&= \sum_z \int p(\boldsymbol{W_d}|z)p(v|z)p(z|\theta)d\theta \\
&= \sum_z p(\boldsymbol{W_d}|z)p(v|z)\int p(z|\theta)d\theta \\
&= \prod_{w\in\boldsymbol{W_d}} [\sum_z \sum_{a\in\boldsymbol{a_d}} p(w|z)p(v|z)\int p(z|\theta)d\theta] \\
&\approx \prod_{w\in\boldsymbol{W_d}} [\frac{1}{|a_d|}\sum_{k=1}^{K}\sum_{a\in\boldsymbol{a_d}} \theta_{ak}\phi_{kw}\vartheta_{kv}] \qquad (20)
\end{aligned}$$

where, $a \in \boldsymbol{a_d}$.

V. Experimental Evaluation

A. Experimental Setup

In order to demonstrate the effectiveness of our model, we carried out a set of experiments on two real world data sets. The first data set is a subset of the **ACM Digital Library**, from which we crawled one descriptive web page for each 172,890 distinct papers having both title and abstract information.

For each published paper, we extracted the information about its publishing venue and references. Due to possible venue name ambiguity, we first converted all upper-case characters into lower-case, and removed all non-alphabetic symbols. We further removed all digits as well as the ordinal numbers, such as the 1st, the 2nd, and applied the Jaccard Similarity match to merge duplicate venue names. We finally obtained 2,197 distinct venues. To remove author names' ambiguity, we represent each candidate author name by a concatenation of the first name and last name, while removing all the middle names. We then use exact match to merge candidate author names. Finally, we obtain 170,897 distinct authors.

The second data set we utilized is the data set 'DBLP-Citation-network V5' provided by Tsinghua University for their ArnetMiner academic search engine [26]. This is data set is the crawling result from the ArnetMiner search engine on Feb 21st, 2011 and further combined with the citation information from DBLP[1]. We name this data set as the **ArnetMiner dataset**. After carrying out the same data processing method as we did for the ACM data set, we find 1,572,277 papers, 795,385 authors and 6,010 publishing venues.

We further carried out a filtering process to remove data noise, and to obtain a smaller subset of both data sets for experiments. We collect for two data sets the papers that have complete information, i.e, title, abstract and venue. Moreover, the papers we collect should have at least one available author and at least one citation. This results in a collection of 92,708

[1] http://www.informatik.uni-trier.de/ ley/db/

TABLE II
STATISTICS OVER ACM AND ARNETMINER DATA SET

Data Set	Paper	Author	Venue	Distinct Word	Word Tokens
ACM	92,708	2,965	1,816	17,341	6,224,821
ArnetMiner	165,330	14,454	2,304	18,151	13,368,826

papers for the ACM data set, and 165,330 papers for the ArnetMiner data set. We further collect authors that have at least one publication and have been cited ten times as minimum, resulting in a set of 2,965 authors and 14,454 authors for ACM and ArnetMiner data sets. We finally filter out the stop words in paper content, and collect sets of 17,341 and 18,151 distinct words for ACM and ArnetMiner respectively that have a word frequency in the entire corpus greater than ten. Table II shows a brief summary of the two data sets we use for experiments.

B. Experimental Methodology and Results

We report in this section results over several groups of experiments. We compare our results with several other state-of-the-art baseline algorithms, and provide analysis for the results.

1) Qualitative Topic Modeling Results: We are interested in examining the modeling results in terms of the four probability distributions we define in the model. In the experiments for both ACM and ArnetMiner data set, we pre-fixed the number of topics to be 50. In this section, we report the top 10 returned words, authors, cited authors, and venues based on their topic-based distributions for one randomly chosen latent topic for ArnetMiner data set as one example.

As shown in Table III, we can observe cohesive and interpretable results. For topic 12, which concerns 'information retrieval'-related research as concluded from the top returned words, we can identify several well-known scientists in this field from both the top 10 author list and cited author list. For example, the top cited author, Prof. Gerard Salton, is regarded as a founding scientist in the field of information retrieval, and the SIGIR Award outstanding contributions in IR research is named after him. The top returned author, Prof. Norbert Fuhr, was presented the Salton Award in 2012 due to "his pioneering, sustained and continuing contributions to the theoretical foundations of information retrieval and database systems."

2) Expert Ranking: **(1). Evaluation Ground Truth**

It has long been acknowledged as one of the problems in expert ranking research that the community lacks both standard query collections and benchmarks for evaluation. Much previous research resorts to human labeling, which is naturally subjective and biased, and is also time-consuming. In this paper, we make use of other evidence and carry out two kinds of evaluations. In the first approach (GT1), we use historical information regarding award winners provided by 16 SIG communities as supporting ground truth. We assume that these award winners are nominated and selected by other researchers in an open and objective way. They are widely acknowledged in their community to have made outstanding contributions in their research fields, and have established world-wide reputations. The corresponding query is generated based on the main research area of that community, for example, the query for SIGIR community is 'information retrieval'. We also check the generated queries with the 23 categories provided by Microsoft Academic engine, and make sure that each query corresponds to one category. We assume that these queries cover the main disciplines of computer science research, and that they represent reasonable topics that users might use for information. These queries are intended to be broad queries.

In the second evaluation approach (GT2), we make use of a benchmark data set with seven queries and expert lists provided by Zhang et al. [31].[2] The expert lists are generated by pooled relevance judgments together with human judgments. Specially, for each query, the top 30 results from three main academic search engines (Libra, Rexa, and ArnetMiner) are collected and merged then further judged by one faculty professor and two graduate students. These queries are more specific queries.

We utilize the traditional IR evaluation metric MAP. We list the query and their corresponding number of experts in Table IV.

TABLE IV
EVALUATION BENCHMARK

Benchmark 1: SIG Community Award Winner	
Query	Expert No.
algorithm theory	7
security privacy	4
hardware architecture	27
software engineering	15
programming language	19
artificial intelligence	14
data mining	7
information retrieval	9
graphics	12
human computer interaction	10
multimedia	2
network communication	18
operating systems	9
database	18
simulation	3
computer education	28
Benchmark 2: ArnetMiner New Expert Lists	
intelligent agents	30
information extraction	20
semantic web	45
support vector machine	31
planning	35
natural language processing	43
machine learning	41

(2). Topic Modeling Results

[2] This data is available online at http://arnetminer.org/lab-datasets/expertfinding/ (the New People Lists).

TABLE III
TOPIC MODELING RESULTS ON ARNETMINER DATA SET

ArnetMiner data set Topic (Information Retrieval)			
Top 10 Words	Top 10 Authors	Top 10 Cited Authors	Top 10 Venues
information	Norbert Fuhr	Gerard Salton	sigir
based	Christopher Manning	W Croft	cikm
web	Jaap Kamps	Hector Molina	world wide web
paper	Kathleen Mckeown	Ricardo Baeza-Yates	acl
search	Gary Lee	Berthier Neto	inf process manage
results	Jian Nie	Justin Zobel	coling
retrieval	Eiichiro Sumita	Fernando Pereira	jcdl
model	Jamie Callan	John Lafferty	jasist
using	Jimmy Lin	Clement Yu	computational linguistics
user	Vimla Patel	Andrew Mccallum	emnlp

We report the experiment results comparing the performance of our ACVT model with the ATM model [23], the CAT model [27], the ACT [25] model, and the ACTC [28] model which is the most recently published work extending ACT [25].

For ACTC [28] model, additional latent variable 'subject' is introduced, and there is no direct author-topic distributions. Instead, each author would be associated with a multinomial distribution over multiple subjects, which have distributions over topics and conferences respectively. There also exists a distribution for topics over words. Under this model, the expertise ranking scheme can be described as:

$$P(a|q) = \prod_{w_i} \sum_{s_j} \sum_{z_t} P(a|s_j) P(s_j|z_t) P(z_t|w_i) \qquad (21)$$

In our experiments, we set the number of latent topics to be 50, and the number of latent subjects to 20 for the ACTC [28] model. For the four hyper-parameters, we set $\alpha = 2$, $\beta = 0.01$, $\gamma = 2$ and $\lambda = 2$. As indicated in the results, our ACVT model works the best in all scenarios and it significantly outperforms the other four models in both ACM and ArnetMiner data sets. Better results can be achieved with the ACVT model using the first benchmark than the second one in both data sets. It can also be observed that under most circumstances, CAT, ACT and ACTC outperform the original ATM, except that working on ArnetMiner data set and using the second benchmark, ACT works slightly worse than ATM. ACTC works better than ACT, and CAT works better than both ACT and ACTC under most circumstances.

Working on ArnetMiner data set, we list in Table VI the Top 10 ranked experts for query 'information retrieval' under five different topic models (ATM, ACT, CAT, ACTC and ACVT) combined with the query. As indicated in the results, we can achieve more valid results using CAT and ACVT than ATM, ACT and ACTC, since several well-known experts in information retrieval can be identified within Top 10, and ranked higher. Furthermore, ACVT can do even better than CAT, since all the returned experts are information retrieval concentrated researchers, while some of the top 10 returned experts by CAT are experts in other fields, for example, Prof.Jeffrey Ullman is famous for his research in compiler, theory of computation and database theory, and Prof.Jennifer Widom is also a well-known database researcher who has won the SIGMOD award in 2007.

TABLE V
COMPARISON OF TOPIC MODELING RESULTS: MAP

ACM data set					
	ATM	CAT	ACT	ACTC	ACVT
GT1	0.0288	0.0688	0.0513	0.0562	**0.1802**
GT2	0.0269	0.0791	0.0780	0.0785	**0.1490**
ArnetMiner data set					
	ATM	CAT	ACT	ACTC	ACVT
GT1	0.0156	0.0919	0.0514	0.0685	**0.1485**
GT2	0.0508	0.0552	0.0673	0.0730	**0.1135**

(3). Combine with Language Model and Random-Walk methods

We examine in this section whether the performance can be improved if we combine topic modeling with a language model-based approach and a random-walk based approach. We report the results for expert ranking in terms of using a language model, a random-walk based method and topic modeling separately, as well as the combined results.

As introduced in section 4.1.2, we adopted two combination methods. For linear combination, we take use of a simple greedy search method to tune the parameters. We gradually change the weight for one particular method from 0 to 1, and let the other two methods evenly share the remaining weights, i.e. ($\alpha \in [0, 1]$, $\beta = \gamma = (1 - \alpha)/2$). Figure 4 and Figure 5 depict the results working on ACM data set using GT1 as the ground truth, and ArnetMiner data set using GT2 as the ground truth respectively. Table VII indicates the results by the multiplication combination method.

Several observations can be made from the results. 1) We can achieve better performance when combining the three

TABLE VII
COMPARISON OF TOPIC MODELING RESULTS: MAP

ACM data set				
	LM	PR	ACVT	LM+PR+ACVT
GT1	0.0752	0.0316	0.1802	**0.1863**
GT2	0.1242	0.0129	0.1490	**0.1529**
ArnetMiner data set				
	LM	PR	ACVT	LM+PR+ACVT
GT1	0.0258	0.0107	0.1485	**0.1750**
GT2	0.1044	0.0104	0.1135	**0.1676**

TABLE VI
EXPERT RANKING RESULTS COMPARISON (ON ARNETMINER DATA SET)

Query: information retrieval				
ATM	ACT	ACTC	CAT	ACVT
Jintao Li	C Giles	Edward Fox	Gerard Salton	W Croft
Ling Duan	Wei-ying Ma	C Giles	Ricardo Baeza-Yates	Gerard Salton
Simone Tini	Ji Wen	Marcos Alves	W Croft	Ricardo Baeza-Yates
Stanley Jr	Maarten Rijke	W Croft	Hector Molina	Hector Molina
Sunil arya	Jian Nie	Berthier Neto	Jiawei Han	Berthier Neto
Karthikeyan Sankaralingam	Irwin King	Maarten Rijke	Rakesh Agrawal	Jiawei Han
Si Wu	Alan Smeaton	Jian Nie	Berthier Neto	Justin Zobel
Cleidson Soua	Chengxiang Zhai	Min Kan	Hans Kriegel	Fernando Pereira
Shi Neo	Rohini Srihari	Mounia Lalmas	Jeffrey Ullman	C Giles
Osman Unsal	W Croft	Mark Sanderson	Jennifer Widom	Wei-ying Ma

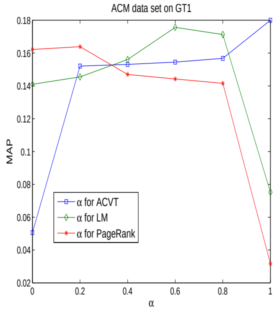

Fig. 4. Combine ranking methods

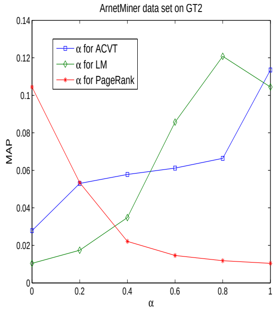

Fig. 5. Combine ranking methods

methods by multiplication than linearly combining them. The best performance under linear combination is always outperformed by multiplication method. This is also true for working on ACM data set with GT2 ground truth, and ArnetMiner data set with GT1 as ground truth. 2) Our ACVT model works better than both the language model and random-walk PageRank-based approach in all experimental scenarios. 2) The language model approach works the second best, and its performance is much better under the first benchmark than the second benchmark. 3) We can achieve improved performance when combing the three approaches together than working on any of them individually. The relative improvement over plain ACVT is 3.45% (ACM under GT1), 2.62% (ACM under GT2), 17.85% (ArnetMiner under GT1) and 47.67% (ArnetMiner under GT2) respectively.

C. Cited Author Prediction

Here we consider the capability of our ACVT model in predicting the authors that any given paper might cite. We take the CAT model as our baseline algorithm, in which cited author information is modeled yet the venue information is missing. In experiments, we select 10,000 papers for the ACM data set, and 18,000 papers for the ArnetMiner data set, as our two testing sets, corresponding to roughly 10% of the total papers in each data set. The criterion for such a selection is that we make sure that the authors of each paper in the testing set has at least one other paper publication in the remaining training set.

TABLE VIII
COMPARISON OF CITED AUTHOR PREDICTION: MAP

Data Set	CAT	ACVT
ACM	0.1029	**0.1154**
ArnerMiner	0.0364	**0.0488**

Predictions are made by following Equation 19. The actual set of cited authors for each test paper serves as our ground truth. We evaluate our performance in terms of MAP, as shown in Table VIII and Precision@K, as shown in Figure 6.

As shown in Table VIII, we can achieve a 12.15% and 34.07% improvement in MAP when using our ACVT model as compared to the CAT model in ACM and ArnetMiner data sets respectively. These demonstrate our model to be more effective in predicting cited authors, and indicate that jointly modeling venue information can provide more cohesive author-topic and topic-cited author associations.

We observed consistent performance in terms of Precision@K across two data sets. Even though the value of Precision@K keeps dropping when K is increased, ACVT outperforms CAT at all different K values. We further notice that there is greater improvement for ACVT over CAT on ArnetMiner data set than ACM data set. For both data sets, the improvement of ACVT over CAT decreases with larger K value.

D. Venue Prediction

We now evaluate the capability of our ACVT model to predict the publishing venue of a given paper. We take the ACT model as our baseline algorithm in which the venue information is modeled yet the cited author information is missing. Similar to the experiments for cited author prediction, we select 10,000 papers and 18,000 papers from ACM and ArnetMiner data sets respectively to work as our testing sets, and make sure that the authors of those chosen papers have at least one other paper in the remaining training sets.

We can perform venue prediction by following Equation 20, and evaluate the results by comparing with the real publishing venue of the given paper.

As demonstrated in Table IX, our ACVT outperforms the ACT model in predicting the publishing venues of any given paper. The improvement of ACVT over ACT is 11.13% for ACM and 71.76% for ArnetMiner. This demonstrates the advantage of jointly modeling multiple facets.

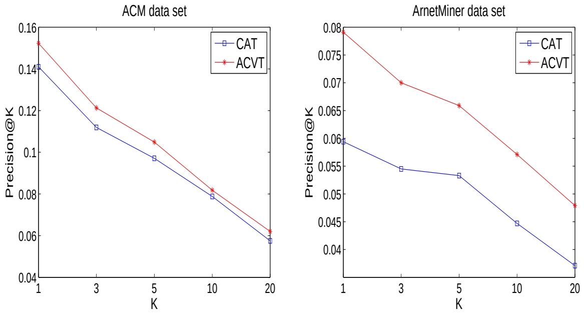

Fig. 6. Cited Autor Prediction: Precision@K

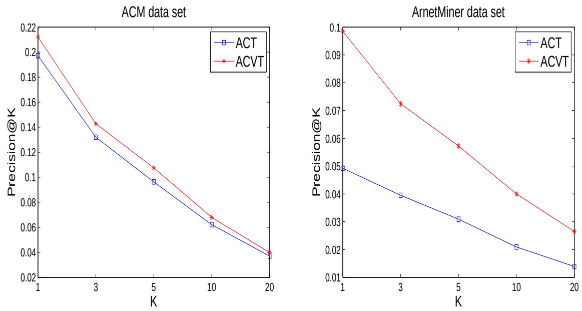

Fig. 7. Venue Prediction: Precision@K

TABLE IX
COMPARISON OF VENUE PREDICTION: MAP

Data Set	ACT	ACVT
ACM	0.3226	**0.3585**
ArnerMiner	0.1151	**0.1977**

Figure 7 shows the performance in terms of Precision@K. We observe similar trend as in the task of cited author prediction. Our ACVT model can outperform the ACT model under all different K values, and we can achieve greater improvement on ArnetMiner data set than on ACM data set.

VI. CONCLUSION AND FUTURE WORK

We proposed in this paper a novel probabilistic topic model (the ACVT model) that can jointly model authors, papers, cited authors and venues in one unified model. As compared to previous work, ACVT can provide a more complete framework to incorporate additional useful contextual information. It is therefore more applicable to multiple applications related with academic network analysis. We have considered performance in three typical applications: expert ranking, cited author prediction and venue prediction. Experiments based on two real world data sets demonstrate that our model can identify more interpretable topic-related associations in terms of authors, cited authors, and venues, and can provide better performance in all three applications as compared with several baseline algorithms.

A number of directions exist for future work, for example, directly modeling author with cited authors or co-authors, including other valuable features, etc. We believe such extra information is likely to provide additional benefit.

REFERENCES

[1] K. Balog, L. Azzopardi, and M. Rijke. Formal Models for expert finding in enterprise corpora. In *SIGIR'06*, 2006.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning*, (3):993–1022, 2003.

[3] H. Deng, J. Han, M. Lyu, and I. King. Modeling and exploiting heterogeneous bibliographics networks for expertise ranking. In *JCDL'11*, 2011.

[4] E. EErosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. In *Proc. of the National Academy Sciences*, pages 5220–5227, 2004.

[5] Y. Fang, L. Si, and A. Mathur. Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search. In *SIGIR'10*, 2010.

[6] S. D. Gollapalli, P. Mitra, and C. L. Giles. Ranking authors in digital libraries. In *JCDL*, pages 251–254. ACM, 2011.

[7] T. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, pages 5228–5235, 2004.

[8] Q. He, D. Kifer, J. Pei, and P. Mitra. Citation Recommendation without author supervision. In *WSDM'11*, 2011.

[9] Q. He, J. Pei, D. Kifer, P. Mitra, and C. Giles. Context-aware citation recommendation. In *WWW'10*, 2010.

[10] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR'99*, pages 50–57, 1999.

[11] X. Jiang, X. Sun, and H. Zhuge. Towards an effective and unbiased ranking of scientific literature through mutual reinforcement. In *CIKM*, pages 714–723, 2012.

[12] N. Johri, D. Roth, and Y. Tu. Experts' retrieval with multiword-enhanced author topic model. In *NAACL-10*, 2010.

[13] S. Kataria, P. Mitra, C. Caragea, and C. Giles. Context Sensitive Topic Models for Author Influence in Document Networks. In *IJCAI'11*, 2011.

[14] N. Lao and W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81:53–67, 2010.

[15] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-Link LDA: Joint Models of Topic and Author Community. In *ICML'09*, 2009.

[16] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM'06*, 2006.

[17] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic Modeling with Network Regularization. In *WWW'08*, 2008.

[18] Q. Mei and C. Zhai. A mixture model for contextual text mining. In *KDD'06*, pages 649–655, 2006.

[19] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *UAI*, 2002.

[20] R. Nallapati, A. Ahmed, E. Xing, and W. Cohen. Joint Latent Topic Models for Text and Citations. In *KDD'08*, 2008.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. In *Stanford InfoLab, Technical Report 1999-66*, 1998.

[22] M. Pham, Y. Cao, R. Klamma, and M. Jarke. A clustering approach for collaborative filtering recommendation using social network analysis. *Journal of Universal Computer Science*, 17(4):583–604, 2011.

[23] M. Rosen-Zvi, T. G. ad M. Steyvers, and P. Smyth. The Author-Topic Model for Authors and Documents. In *UAI'04*, 2004.

[24] Y. Song, J. Huang, and I. Councill. Efficient topic-based unsupervised name disambiguation. In *JCDL'07*, pages 342–351, 2007.

[25] J. Tang, R. Jin, and J. Zhang. A Topic Modeling Approach and its Integration into the Random Walk Framework for Academic Search. In *ICDM'08*, 2008.

[26] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: extraction and mining of academic social network. In *KDD'08*, 2008.

[27] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier. Citation Author Topic Model in Expert Search. In *COLING'10*, 2010.

[28] J. Wang, X. Hu, X. Tu, and T. He. Author-conference topic-connection model for academic network search. In *CIKM*, pages 2179–2183, 2012.

[29] Z. Yang and B. D. Davison. Venue recommendation: Submitting your paper with style. In *ICMLA (1)'12*, pages 681–686, 2012.

[30] Z. Yang, L. Hong, and B. Davison. Topic-driven multi-type citation network analysis. In *RIAO'10*, 2010.

[31] J. Zhang, J. Tang, and J. Li. Expert finding in a social network. In *DASFAA*, pages 1066–1069, 2007.