Joint Modeling of Topics, Citations, and Topical Authority in Academic Corpora

Jooyeon Kim
KAIST
jooyeon.kim@kaist.ac.kr

Dongwoo Kim
Australian National University
dongwoo.kim@anu.edu.au

Alice Oh
KAIST
alice.oh@kaist.edu

Abstract

Much of scientific progress stems from previously published findings, but searching through the vast sea of scientific publications is difficult. We often rely on metrics of scholarly authority to find the prominent authors but these authority indices do not differentiate authority based on research topics. We present Latent Topical-Authority Indexing (LTAI) for jointly modeling the topics, citations, and topical authority in a corpus of academic papers. Compared to previous models, LTAI differs in two main aspects. First, it explicitly models the generative process of the citations, rather than treating the citations as given. Second, it models each author's influence on citations of a paper based on the topics of the cited papers, as well as the citing papers. We fit LTAI into four academic corpora: CORA, Arxiv Physics, PNAS, and Citeseer. We compare the performance of LTAI against various baselines, starting with the latent Dirichlet allocation, to the more advanced models including author-link topic model and dynamic author citation topic model. The results show that LTAI achieves improved accuracy over other similar models when predicting words, citations and authors of publications.
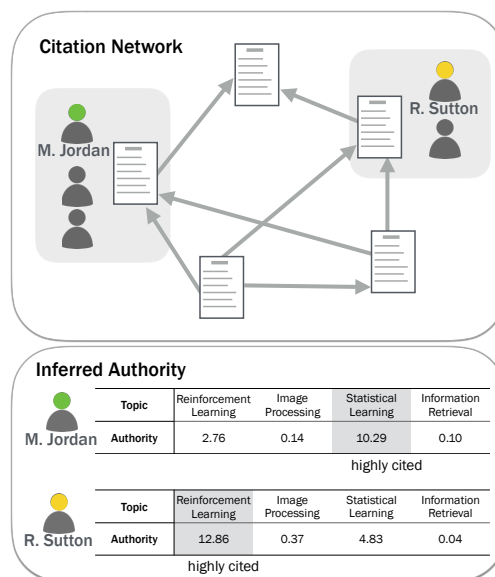
Figure 1: Overview of Latent Topical Authority Indexing (LTAI). Based on content, citation, and authorship information (top), the LTAI discovers the topical authority of authors; it increases when a paper with certain topics gets cited (bottom). Topical authority examples are the results of the LTAI with CORA dataset and 100 topics.

1 Introduction

With a corpus of scientific literature, we can observe the complex and intricate process of scientific progress. We can learn the major topics in journal articles and conference proceedings, follow authors who are prolific and influential, and find papers that are highly cited. The huge number of publications and authors, however, makes it practically impossible to attain any deep or detailed understanding beyond the very broad trends. For example, if we want to identify authors who are particularly influential in a specific research field, it is difficult to do so without the aid of automatic analysis.

Online publication archives, such as Google Scholar, provide near real-time metrics of schol-

arly impact, such as the h-index (Hirsch, 2005), the journal impact factor (Garfield, 2006), and citation count. Those indices, however, are still at a coarse level of granularity. For example, both Michael Jordan and Richard Sutton are researchers with very high citation counts and h-index, but they are authoritative in different topics, Jordan in the more general machine learning topic of statistical learning, and Sutton in the topic of reinforcement learning. It would be much more helpful to know that via topical authority scores, as shown in Figure 1.

Fortunately, various academic publication archives contain the full contents, references, and meta-data including titles, venues, and authors. With such data, we can build and fit a model to partition researchers' scholarly domains into topics at a much finer-grain and discover their academic authority within each topic. To do that, we propose a model named Latent Topical-Authority Indexing (LTAI), based on the LDA, to jointly model the topics, authors' topical authority, and citations among the publications.

We illustrate the modeling power of the LTAI with four corpora encompassing a diverse set of academic fields: CORA, Arxiv Physics, PNAS, and Citeseer. To show the improvements over other related models, we carry out prediction tasks on words, citations and authorship using the LTAI and compare the results with those of Latent Dirichlet Allocation (Blei et al., 2003), relational topic model (Chang and Blei, 2010), author-link topic model, and dynamic author-cite topic model (Kataria et al., 2011), as well as simple baselines of topical h-index. The results show that the LTAI outperforms these other models for all prediction tasks.

The rest of this paper is organized as follows. In section 2, we describe related work, including models that are most similar to the LTAI, and describe how the LTAI fits in and contributes to the field. In section 3, we describe the LTAI model in detail and present the generative process. In section 4, we explain the algorithm for approximate inference, and in section 5, we present a faster algorithm for scalability. In section 6, we describe the experimental setup and in section 7, we present the results to show that the LTAI performs better than other related models for word, citation and authorship prediction.

2 Related Work

In this section, we review related papers, first in the field of NLP and ML-based analysis of scientific corpora, then the approaches based on the Bayesian topic models for academic corpora, and lastly joint models of topics, authors, and citations. In analyzing scientific corpora, previous research presents classifying scientific publications (Caragea et al., 2015), recommending yet unlinked citations (Huang et al., 2015; Neiswanger et al., 2014; Wang et al., 2015; Jiang, 2015), summarizing and extracting key phrases (Cohan and Goharian, 2015; Caragea et al., 2014), triggering a better model fit (He et al., 2015), incorporating authorship information to increase the content and link predictability (Sim et al., 2015), estimating a paper's potential influence on academic community (Dong et al., 2015), and finding and classifying different functionalities of citation practices (Moravcsik and Murugesan, 1975; Teufel et al., 2006; Valenzuela et al., 2015).

Several variants of topic modeling consider the relationship between topics and citations in academic corpora. Topic models that use text and citation networks are divided into two types: (a) models that generate text given citation networks (Dietz et al., 2007; Foulds and Smyth, 2013) and (b) models that generate citation networks given text (Nallapati et al., 2008; Liu et al., 2009; Chang and Blei, 2010). While our model falls into the latter category, we also take into account the influence of the authors on the citation structure.

Most closely related to the LTAI is the citation author topic model (Tu et al., 2010), the author-link topic model, and the dynamic author-cite topic model (Kataria et al., 2011). Similar to the LTAI, they are designed to capture the influence of the authors. However, these models infer authority by referencing only the citing papers' text, while our authority is based on the predictive modeling of comparing both the citing and the cited papers. Furthermore, the LTAI defines a generative model of citations and publications by introducing a latent authority index, whereas the previous models assume the citation structure is given. The LTAI thus explicitly gives a topical authority index, which directly answers the question of which author increases the probability of a paper being cited.

3 Latent Topical-Authority Indexing

The LTAI models the complex relationships among the topics of publications, the topical authority of the authors, and the citations among these publications. The generative process of the LTAI can be divided into two parts: content generation and citation network generation. We make several assumptions in the LTAI to model citation structures of academic corpora. First, we assume a citation is more likely to occur between two papers that are similar in their topic proportions. Second, we assume that an author differs in their authority (i.e., potential to induce citation) for each topic, and an author's topical authority positively correlates with the probability of citations among publications. Also, in the LTAI, when there are multiple authors in a single cited publication, their contribution of forming citations with respect to different citing papers varies according to their topical authority. Lastly, we assign different concentration parameters for a pair of papers with and without citations. In this paper, we use positive and negative links to denote pairs of papers with and without citations respectively.

Figure 2 illustrates the graphical model of the LTAI, and we summarize the generative process of the LTAI, where the variables of the model are explained in the remainder of this section, as follows:

1. For each topic k, draw topic $\beta_k \sim \text{Dir}(\alpha_\beta)$.

2. For each document i:

   (a) Draw topic proportion $\theta_i \sim \text{Dir}(\alpha_\theta)$.
   (b) For each word w_{in}:
      i. Draw topic assignment $z_{in}|\theta_i \sim \text{Mult}(\theta_i)$.
      ii. Draw word $w_{in}|z_{in}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{in}})$.

3. For each author a and topic k:

   (a) Draw authority index $\eta_{ak} \sim \mathcal{N}(0, \alpha_\eta^{-1})$.

4. For each document pair from i to j:

   (a) Draw influence proportion parameter $\pi_{i\leftarrow j} \sim \text{Dir}(\pi_i)$.
   (b) Draw author $a_i|\pi_{i\leftarrow j} \sim \text{Mult}(\pi_{i\leftarrow j})$.
   (c) Draw link $x_{i\leftarrow j} \sim \mathcal{N}(\sum_k \eta_{a_ik}\bar{z}_{ik}\bar{z}_{jk}, c_{i\leftarrow j}^{-1})$.
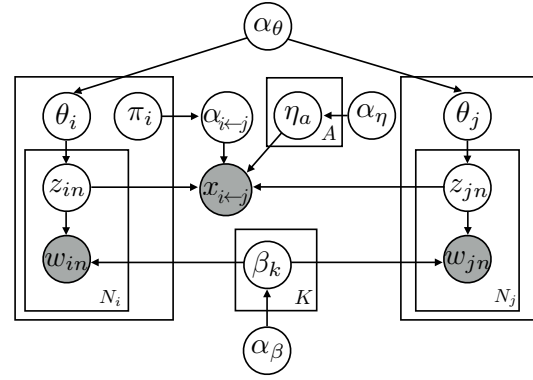

Figure 2: Graphical representation of the LTAI. The LTAI jointly models content-related variables θ, z, w, β, and author and citation related variables η and π.

3.1 Content Generation

To model the content of publications, we follow a standard document generative process of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We inherit notations for variables from LDA; θ is the per-document topic distribution, β is the per-topic word distribution, z is the topic for each word in a document where w is the corresponding word, and $\alpha_\theta, \alpha_\beta$ are the Dirichlet parameters of θ and β.

3.2 Citation Generation

Let $x_{i\leftarrow j}$ be a binary valued variable which indicates that publication j cites publication i. We formulate a continuous variable $r_{i\leftarrow j}$ which is a linear combination of the authority variable and the topic proportion variable to approximate $x_{i\leftarrow j}$ by minimizing the sum of squared errors between the two variables. There is a body of research on using continuous user and item-related variables to approximate binary variables in the field of recommender systems (Rennie and Srebro, 2005; Koren et al., 2009).

Approximating binary variables using a linear combination of continuous variables can be probabilistically generalized (Salakhutdinov and Mnih, 2007). Using probabilistic matrix factorization, we approximate probability mass function $p(x_{i\leftarrow j})$ using probability density function $\mathcal{N}(x_{i\leftarrow j}|r_{i\leftarrow j}, c_{i\leftarrow j}^{-1})$, where the precision parameter $c_{i\leftarrow j}$ can be set differently for each pair of papers as will be discussed below.

Content Similarity Between Publications: In the LTAI, we model relationships between a random pair of documents i and j. The probability of publication j citing publication i is proportional to the similarity of topic proportions of two publications, i.e., $r_{i \leftarrow j}$ positively correlates to $\sum_k \theta_{ik} \theta_{jk}$. Following the relational topic model's approach (Chang and Blei, 2010), we use $\bar{z}_i = \frac{1}{N_i} \sum_n z_{i,n} \approx \theta_i$ instead of a topic proportion parameter θ_i.

Topical Authority of Cited Paper: We introduce a K-dimensional vector η_a for representing the topical authority index of author a. η_{ak} is a real number drawn from the zero-mean normal distribution with variance α_η^{-1}. Given the authority indices η_{a_i} for author a of cited publication i, the probability of a citation is further modeled as $r_{i \leftarrow j} = \sum_k \eta_{a_i k} \bar{z}_{ik} \bar{z}_{jk}$, where the authority indices can promote or demote the probability of a citation.

Different Degree of Contribution among Multiple Authors: Academic publications are often written by more than one author. Thus, we need to distinguish the influence of each author on a citation between two publications. Let $\mathcal{A}_i$ be a set of authors of publication i. To measure the influence proportion of author $a \in \mathcal{A}_i$ on the citation from i to j, we introduce an additional parameter $\pi_{i \leftarrow j}$ which is a one-hot vector drawn from a Dirichlet distribution with $|\mathcal{A}_i|$-dimensional parameter π_i. $\pi_{i \leftarrow ja} \in \{0, 1\}$ is an element of $\pi_{i \leftarrow j}$ which measures the influence of author a on the citation from j to i and sums up to one ($\sum_{a \in \mathcal{A}_i} \pi_{i \leftarrow ja} = 1$) over all authors of publication i. We approximate the probability of citation $x_{i \leftarrow j}$ from publication j to publication i by $p(x_{i \leftarrow j} | z, \pi_{i \leftarrow j}, a_{i \leftarrow j}, \eta_a) \approx \sum_{a \in \mathcal{A}_i} \pi_{i \leftarrow ja} \mathcal{N}(x_{i \leftarrow j} | \sum_k \eta_{ak} \bar{z}_{ik} \bar{z}_{jk}, c_{i \leftarrow j}^{-1})$ which is a mixture of normal distributions with precision parameter $c_{i \leftarrow j}$. Therefore, if topic distributions of paper i and j are similar and if η values of the cited paper's authors are high, the citation formation probability increases. On the other hand, dissimilar or topically irrelevant pair of papers with less authoritative authors on the cited paper will be assigned with low probability of citation formation.

Different Treatment between Positive and Negative links: Citation is a binary problem where $x_{i \leftarrow j}$ is either one or zero. When $x_{i \leftarrow j}$ is zero, this can be interpreted in two ways: 1) the authors of citing publication j are unaware of the publication i, or 2) the

publication j is not relevant to publication i. Identifying which case is true is impossible unless we are the authors of the publication. Therefore the model embraces this uncertainty in the absence of a link between publications. We control the ambiguity by the Gaussian distribution with precision parameter $c_{i \leftarrow j}$ as follows:

$$c_{i \leftarrow j} = \begin{cases} c_+ & \text{if } x_{i \leftarrow j} = 1 \\ c_- & \text{if } x_{i \leftarrow j} = 0 \end{cases} \quad (1)$$

where $c_+ > c_-$ to ensure that we have more confidence on the observed citations. This is an implicit feedback approach that permits using negative examples ($x_{i \leftarrow j} = 0$) of sparse observations by mitigating their importance (Hu et al., 2008; Wang and Blei, 2011; Purushotham et al., 2012). Setting different values to the precision parameter $c_{i \leftarrow j}$ according to $x_{i \leftarrow j}$ induces cyclic dependencies between the two variables. Due to this cycle, the model no longer becomes a Bayesian network, or a directed acyclic graph. However, we note that this setting does lead to better experimental results, and we show the pragmatic benefit of the setting in the Evaluation section.

3.3 Joint Modeling of the LTAI

In the LTAI, the topics and the link structures are simultaneously learned, and thus the content-related variables and the citation-related variables mutually reshape one another during the posterior inference. On the other hand, if content and citation data are modeled separately, the topics would not reflect any information about the document citation structures. Thus, in the LTAI, documents with shared links are more likely to have similar topic distributions which leads to a better model fit. We develop and explain this joint inference in section 4. In section 7, we illustrate the differences in word-level predictive powers of the LTAI and LDA.

4 Posterior Inference

We develop a hybrid inference algorithm in which the posterior of content-related parameters θ, z, and β are approximated by variational inference, and author-related parameters π and η are approximated by an expectation-maximization (EM) algorithm. In algorithm 1, we summarize the full inference procedure of the LTAI.

4.1 Content Parameters: Variational Update

Since computing the posterior distribution of the LTAI is intractable, we use variational inference to optimize variational parameters each of which correspond to original content-related variables. Following the standard mean-field variational approach, we define fully factorized variational distributions over the topic-related latent variables $q(\theta, \beta, z) = \prod_i q(\theta_i|\Psi_{in}) \prod_{N_i} q(z_{in}|\gamma_i) \prod_k q(\beta_k|\lambda_k)$ where for each factorized variational distribution, we place the same family of distributions as the original distribution. Using the variational distributions, we bound the log-likelihood of the model as follows:

$$
\begin{aligned}
\mathcal{L}_{[q]} = \mathbb{E}_q[&\sum_k \log p(\beta_k|\alpha_\beta) + \sum_i \log p(\theta_i|\alpha_\theta) \\
&+ \sum_i \sum_{N_i} \log p(z_{in}|\theta_d) + \log p(w_{in}|\beta_{z_{in}}) \quad (2)\\
&+ \sum_{i,j} \log p(x_{i\leftarrow j}|z_i, z_j, \pi_i)] - \mathcal{H}[q]
\end{aligned}
$$

where $\mathcal{H}[q]$ is the negative entropy of q.

Taking the derivatives of this lower bound with respect to each variational parameter, we can obtain the coordinate ascent updates. The update for the variational Dirichlet parameters γ_i and the λ_k is the same as the standard variational update for LDA (Blei et al., 2003). The update for the variational multinomial ϕ_{in} is:

$$
\begin{aligned}
\phi_{ink} \propto \exp \Bigg\{ &\frac{\sum_j \partial \mathbb{E}_q[\log p(x_{i\leftarrow j}|\bar{z}_i, \bar{z}_j, \pi_i, \eta)]}{\partial \phi_{ink}} \\
&+ \frac{\sum_j \partial \mathbb{E}_q[\log p(x_{j\leftarrow i}|\bar{z}_j, \bar{z}_i, \pi_j, \eta)]}{\partial \phi_{ink}} \quad (3)\\
&+ \mathbb{E}_q[\log \theta_{ik}] + \mathbb{E}_q[\log \beta_{kw_{in}}]\Bigg\}
\end{aligned}
$$

where the gradient of expected log probabilities of both incoming link $x_{i\leftarrow j}$ and outgoing link $x_{j\leftarrow i}$ contribute to the variational parameter. The first expectation can be rewritten as

$$
\begin{aligned}
&\mathbb{E}_q[\log p(x_{i\leftarrow j}|\bar{z}_i, \bar{z}_j, \pi_i, \eta)] \quad (4)\\
&= \mathbb{E}_q[\log \sum_{a\in A_i} p(a_{i\leftarrow j}=a|\pi_i)p(x_{i\leftarrow j}|\bar{z}_i, \bar{z}_j, \eta_a)] \\
&\geq \sum_{a\in A_i} p(a_{i\leftarrow j}=a|\pi_i)\mathbb{E}_q[\log p(x_{i\leftarrow j}|\bar{z}_i, \bar{z}_j, \eta_a)]
\end{aligned}
$$

Algorithm 1 Posterior inference algorithm for the LTAI

Initialize γ, λ, π, and η randomly
Set learning-rate parameter ρ_t that satisfies Robbins-Monro condition
Set subsample sizes S_V, S_E, S_S and S_A
repeat
  Variational update: local publication parameters
  $\mathcal{S}_S \leftarrow S_S$ randomly sampled publications
  for i in $\mathcal{S}_S$ **do**
    for $n = 1$ to N_i **do**
      $S_\leftarrow, S_\rightarrow \leftarrow$ Set of S_V random samples
      Update ϕ_{ink} using Equation 4, 5, 9.
    end for
    $\gamma_i \leftarrow \alpha_\theta + \sum_{N_i} \phi_{in}$
  end for

  EM update: local author parameters
  $\mathcal{S}_A \leftarrow S_A$ randomly sampled authors
  for a in $\mathcal{S}_A$ **do**
    $\mathcal{S}_E \leftarrow S_E$ random publication pairs
    Update η_a using Equation 7, 10
    for i in $\mathcal{D}_a$ and $j = 1$ to D **do**
      $\pi_{i\leftarrow ja} \propto \pi_{ia}\mathcal{N}(\bar{z}_i^\top \eta_a \bar{z}_j, c_{i\leftarrow j}^{-1})$
    end for
  end for

  Stochastic variational update
  for $k = 1$ to K **do**
    $\hat{\lambda}_k \leftarrow \alpha_\beta + \frac{D}{S_S}\sum_{d=1}^{\mathcal{S}_S}\sum_{n=1}^{N_d}\phi_{dn}^k w_{dn}$
  end for
  Set $\lambda^{(t)} \leftarrow (1-\rho^t)\lambda^{(t-1)} + \rho_t\hat{\lambda}$
until satisfying converge criteria

where A_i is the set of authors of i. We take the lower bound of the expectation using Jensen's inequality. The last term is approximated by the first order Taylor expansion $\mathbb{E}_q[\log p(x_{i\leftarrow j}|\bar{z}_i, \bar{z}_j, \eta_a)] = \mathcal{N}(x_{i\leftarrow j}|\bar{\phi}_i^\top \text{diag}(\eta_a)\bar{\phi}_j, c_{i\leftarrow j}^{-1})$. Finally, the approximated gradient of ϕ_{ink} with respect to the incoming directions to document i is

$$
\frac{\sum_j \partial \mathbb{E}_q[\log p(x_{i\leftarrow j}|\bar{z}_i, \bar{z}_j, \pi_i, \eta)]}{\partial \phi_{ink}} \approx \quad (5)
$$

$$
\sum_j \frac{\bar{\phi}_{jk}c_{i\leftarrow j}}{N_i} \sum_{a\in A_i} \eta_{ak}(x_{i\leftarrow j} - \bar{\phi}_i^\top \text{diag}(\eta_a)\bar{\phi}_j)p(a|\pi_i)
$$

where diag is a diagonalization operator and $\bar{\phi}_i$ is $\sum_{n=1}^{N_i}\phi_{in}/N_i$. We can compute the gradient with respect to the outgoing directions in the same way.

4.2 Author Parameters: EM Step

We use the EM algorithm to update author-related parameters π, and η based on the lower bound computed by variational inference. In the E step, we compute the probability of author contribution to the link between document i and j.

$$\pi_{i\leftarrow ja} = \frac{\pi_{ia}\mathcal{N}(\bar{z}_i^\top \eta_a \bar{z}_j, c_{i\leftarrow j}^{-1})}{\sum_{a'\in A_i} \pi_{ia'}\mathcal{N}(\bar{z}_i^\top \eta_{a'} \bar{z}_j, c_{i\leftarrow j}^{-1})} \quad (6)$$

In the M step, we optimize the authority parameter η for each author. Given the other estimated parameters, taking the gradient of $\mathcal{L}$ with respect to η_a and setting it to zero leads to the following update equation:

$$\eta_a = (\Psi_a^\top C_a \Psi_a + \alpha_\eta I)^{-1} \Psi_a^\top C_a X_a \quad (7)$$

Let $\mathcal{D}_a$ be the set of documents written by author a and $\mathcal{D}_a(i)$ be the ith document written by a. Then Ψ_a is a vertical stack of $|\mathcal{D}_a|$ matrices $\Psi_{\mathcal{D}_a(i)}$, whose jth row is $\bar{\phi}_{\mathcal{D}_a(i)} \circ \bar{\phi}_j$, the Hadamard product between $\bar{\phi}_{\mathcal{D}_a(i)}$ and $\bar{\phi}_j$. Similarly, C_a is a vertical stack of $|\mathcal{D}_a|$ matrices $C_{\mathcal{D}_a(i)}$ whose j th diagonal element is $c_{\mathcal{D}_a(i)\leftarrow j}$, and X_a is a vertical stack of $|\mathcal{D}_a|$ vectors $X_{\mathcal{D}_a(i)}$ whose j th element is $\pi_{\mathcal{D}_a(i)\leftarrow ja} \times x_{\mathcal{D}_a(i)\leftarrow j}$. Finally, we update $\pi_{\mathcal{D}_a(i)a} = \sum_j \pi_{\mathcal{D}_a(i)\leftarrow ja}/D$.

5 Faster Inference Using Stochastic Optimization

To model topical authority, the LTAI considers the linkage information. If two papers are linked by citation, the topical authority of the cited paper's authors will increase while the negative link buffers the potential noise of irrelevant topics. This algorithmic design of the LTAI results in high model complexity. To remedy this issue, we adopt the noisy gradient method from the stochastic approximation algorithm (Robbins and Monro, 1951) to subsample negative links for updating per-document topic variational parameter ϕ and authority parameter η. The prior work of using subsampled negative links to reduce computational complexity is introduced in (Raftery et al., 2012). We elucidate how stochastic variational inference (Hoffman et al., 2013) is applied in our model to update global per-topic-word variational parameter λ.

5.1 Updating ϕ and η

Updating $\bar{\phi}_i$ for document i in variational update requires iterating over every other document and computing the gradient of link probability. This leads to the time complexity $\mathcal{O}(DK)$ for every $\bar{\phi}_i$.

To apply the noisy gradient method, we divide the gradient of the expected log probability of link into two parts:

$$\sum_j \frac{\partial \mathbb{E}_q[\log p(x_{i\leftarrow j}|\bar{z}_i, \bar{z}_j, \pi_i)]}{\partial \phi_{ink}} = \quad (8)$$

$$\sum_{j:x_{i\leftarrow j}=1} \frac{\partial \mathbb{E}_q[\log p(x_{i\leftarrow j})]}{\partial \phi_{ink}} + \sum_{j:x_{i\leftarrow j}=0} \frac{\partial \mathbb{E}_q[\log p(x_{i\leftarrow j})]}{\partial \phi_{ink}}$$

where the first and the second terms of the right-hand side are the gradient sums of positive links ($x_{i\leftarrow j} = 1$) and negative links ($x_{i\leftarrow j} = 0$) respectively. Compared to positive links, the order of negative links is close to the total number of documents, and thus computing the second term results in computational inefficiency. However, in our model, we reduced the importance of the negative links by assigning a larger variance $c_{i\leftarrow j}^{-1}$ compared to the positive links, and the empirical mean of $\bar{\phi}_j$ for negative links follows the Dirichlet expectation due to the large number of negative links. Therefore, we approximate the expectation of the gradient for the negative links using the noisy gradient as follows:

$$\sum_{j:x_{i\leftarrow j}=0} \frac{\partial \mathbb{E}_q[\log p(x_{i\leftarrow j})]}{\partial \phi_{ink}} = \frac{D_i^-}{S_V} \sum_{\mathcal{S}_V} \frac{\partial \mathbb{E}_q[\log p(x_{i\leftarrow s})]}{\partial \phi_{ink}} \quad (9)$$

where D_i^- is the number of negative links (i.e. $x_{i\leftarrow j} = 0$) of document i, and S_V is the size of subsamples $\mathcal{S}_V$ for the variational update. We randomly sample S_V documents, compute gradients on the sampled documents, and then scale the average gradient to the size of the negative link D_i^-. This noisy gradient method reduces the updating time complexity from $\mathcal{O}(DK)$ to $\mathcal{O}(S_V K)$.

Now, we discuss how to approximate author's topical authority based on Equation 7. When $K \ll D \times D_a$, the computational bottleneck is $\Psi_a^\top C_a \Psi_a$ which has time complexity $\mathcal{O}(DD_a K^2)$. To alleviate this complexity, we once again approximate the large number of negative links using smaller number of subsamples. Specifically, while keeping the positive link rows Ψ_{a+} intact, we approximate negative link rows in Ψ_a using a smaller matrix Ψ_{a-} that
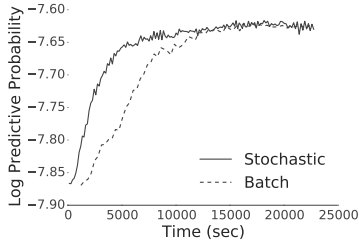
Figure 3: Training time of the LTAI on CORA dataset with stochastic and batch variational inference. Using stochastic variational inference, the per-word predictive log likelihood converges faster than using the batch variational inference.

Dataset	# Tokens	# Documents	# Authors	Avg C/D	Avg C/A
CORA	17,059	13,147	12,111	3.46	12.17
Arxiv-Physics	49,807	27,770	10,950	12.70	67.93
PNAS	39,664	31,054	9,862	1.57	13.18
Citeseer	21,223	4,255	6,384	1.24	4.38

Table 1: Datasets. From left to right, each column shows the number of word tokens, number of documents, number of authors, average citations per document (Avg C/D), and average citations per author (Avg C/A).

has S_E rows, or the size of subsamples for the EM step. Using this approximation, we can represent $\Psi_a^\top C_a \Psi_a$ as

$$\Psi_a^\top C_a \Psi_a = c_+ \Psi_{a+}^\top \Psi_{a+} + \frac{c_- D_a^-}{S_E} \Psi_{a-}^\top \Psi_{a-} \quad (10)$$

with the time complexity of $\mathcal{O}(S_E K^2)$, where D_a^- is the number of rows with negative links in Ψ_a. Although we do not incorporate rigorous analysis on the performance of our model given the size of the subsamples, we confirm that the negative link size greater than 100 does not degrade the model performance in any of our experiment.

5.2 Updating λ

In traditional coordinate ascent based variational inference, the global variational parameter λ is updated infrequently because all the other local parameters ϕ need to be updated beforehand. This problem is more noticeable in the LTAI since updating ϕ using equation 3 is slower than updating ϕ in vanilla LDA; moreover, per-author topical authority variable η is another local variable that algorithm needs

to update a priori. However, using the stochastic variational inference, the global parameters are updated after a small portion of local parameters are updated (Hoffman et al., 2013). Applying stochastic variational inference for the LTAI is straightforward after we calculate the intermediate topic-word variational parameter $\hat\lambda$ by $\alpha_\beta + \frac{D}{S_S} \sum_{d=1}^{S_S} \sum_{n=1}^{N_d} \phi_{dn}^k w_{dn}$ from the noisy estimate of the natural gradient with respect to subsampled local parameters where N_d is the number of words for document d, and S_S is the subsample size for the minibatch stochastic variational inference. The final global parameter for the tth iteration $\lambda^{(t)}$ is updated by $(1 - \rho^t)\lambda^{(t-1)} + \rho_t \hat\lambda$ where ρ_t is the learning-rate. Posterior inference is guaranteed to converge at local optimum when the learning rate satisfies the condition $\sum_{t=1}^\infty \rho_t = \infty, \sum_{t=1}^\infty \rho_t^2 < \infty$ (Robbins and Monro, 1951). In Figure 3, we confirm that stochastic variational inference is applicable for the LTAI and reduces the training time compared to using the batch counterpart, while maintaining similar performance.

6 Experimental Settings

In this section, we introduce the four academic corpora used to fit the LTAI, describe comparison models, and provide information about the evaluation metric and parameter settings for the LTAI[1].

6.1 Datasets

We experiment with four academic corpora: CORA (McCallum et al., 2000), Arxiv-Physics (Gehrke et al., 2003), the Proceedings of the National Academy of Sciences (PNAS), and Citeseer (Lu and Getoor, 2003). CORA, Arxiv-Physics, and PNAS datasets contain abstracts only, and the locations of the citations within each paper are not preserved, whereas the Citeseer dataset contains the citation locations. For CORA, Arxiv-Physics, and PNAS, we lemmatize words, remove stop words, and discard words that occur fewer than four times in the corpus. Table 1 describes the datasets in detail. Note that we obtain citation data from the entire document, not only from the abstract. Also, we consider within-corpus citation only, which leads to less than 13 average citation counts per document for all corpora.

---

[1]Code and datasets are available at http://uilab.kaist.ac.kr/research/TACL2017/

197

(a) CORA   (b) Arxiv-Physics

LTAI   LDA
LTAI-10%   LTAI-20%   LTAI-30%
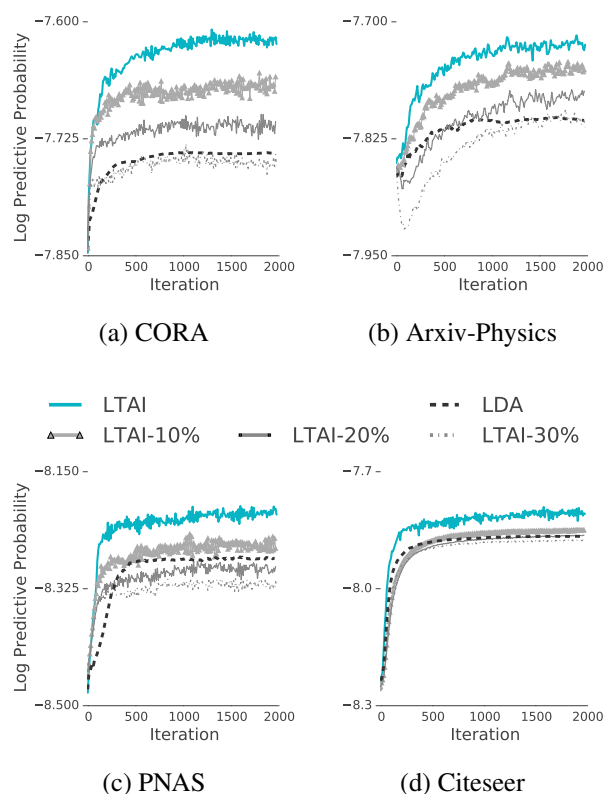
(c) PNAS   (d) Citeseer

Figure 4: Word-level prediction result. We measured per-word log predictive probability on four datasets. As shown in the graphs, our model performs better than LDA.

6.2 Comparison Models

We compare predictive performance of the LTAI with five other models. Different comparison models have different degrees of expressive powers. Each model conducts a certain type of prediction task; while RTM, ALTM, and DACTM predicts citation structures, the topical h-index predicts authorship information. The baseline topic models are implemented based on the inference methods suggested in the corresponding papers; LDA, RTM and the LTAI variants use variational inference, while ALTM and DACTM use collapsed Gibbs sampling. Finally, all the conditions for implementation such as the choice of programming language and modules, except for parts that convey each model's unique assumption, are identically set; thus, the performance differences between models are due to their model assumption and different degrees of data usage, rather than the implementation technicalities.

Latent Dirichlet Allocation: LDA (Blei et al., 2003) discovers topics and represents each publication by a mixture of the topics. Compared to other models, LDA only uses the content information.

LTAI-n%: In LTAI-n%, we remove n% of actual citations and displace them with arbitrarily selected false connections. Note that the link structures are displaced rather than removed. If the citation links are just removed, the LTAI and LTAI-n% cannot be fairly compared as the density of the citation structures will be affected and each model needs different concentration values. Performance difference between the LTAI and this indicates that under identical conditions, using the correct linkage information is indeed beneficial for prediction.

LTAI-C: In LTAI-C the precision parameter $c_{i \leftarrow j}$ has constant value, rather than assigning different values according to $x_{i \leftarrow j}$ as discussed in section 3.

LTAI-SEP: LTAI-SEP has an identical structure as the LTAI, but the topic and the authority variables are separately learned. Once the topic variables are learned using the vanilla LDA, authority and citation variables are then inferred consecutively. Thus, the performance edge of the LTAI over LTAI-SEP highlights the necessity of the LTAI's joint modeling in which both topic and authority related variables reshape one another in an iterative fashion.

Relational Topic Model: RTM (Chang and Blei, 2010) jointly models content and citation, and thus, topic proportions of a pair of publications become similar if the pair is connected by citations. Compared to the LTAI, the author information is not considered, the link structure does not have directionality and the model does not consider negative links.

Author-Link Topic Model: ALTM (Kataria et al., 2011) is a variation of author topic model (ATM) (Rosen-Zvi et al., 2004) that models both topical interests and influence of authors in scientific corpora. The model uses content information of citing papers and names of the cited authors as word tokens. ALTM outputs per-topic author distribution that functions as author influence indices.

Dynamic Author-Citation Topic Model: DACTM (Kataria et al., 2011) is an extension of ALTM that requires publication corpora which preserves sentence structures. To model author influence, DACTM selectively uses words that are close to the point where the citation is presented.

(a) CORA   (b) Arxiv-Physics
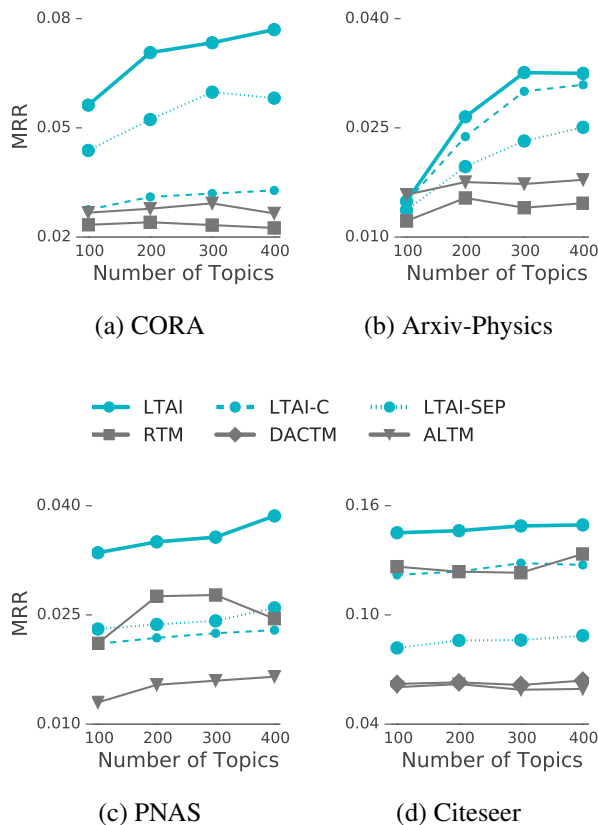
(c) PNAS   (d) Citeseer

Figure 5: Citation prediction results. The task is to find out which paper is originally linked to a cited paper. We measure Mean Reciprocal Rank (MRR) to evaluate model performance. For all cases, the LTAI performs better than the other methods.

In our corpora, only Citeseer dataset preserves the sentence structure.

Topical h-index: To compute topical h-index, we separate the papers into several clusters using LDA and calculate the h-index within each cluster. Topical h-index is used for author prediction in the same manner as we did for our model, except the topic proportions are replaced to the LDA's result and η is replaced to the topical h-index values.

6.3 Evaluation Metric and Parameter Settings

We use MRR (Voorhees, 1999) to measure the predictive performance of the LTAI and the comparison models. MRR is a widely used metric for evaluating link prediction tasks (Balog and de Rijke, 2007; Diehl et al., 2007; Radlinski et al., 2008; Huang et al., 2015). When the models output the correct an-

swers as ranks, MRR is the inverse of the harmonic mean of such ranks.

We report the parameter values used for evaluations. For all datasets, we set c_- to 1. To predict a citation, we set c_+ to 10,000, 100, 1,000, 10, and to predict authorship, we set c_+ to 1,000, 1,000, 10,000, 1,000 for CORA, Arxiv-Physics, PNAS, and Citeseer datasets. These values are obtained through exhaustive parameter analysis. We set α_θ to 1, and α_β to 0.1. We fix the subsample sizes to 500^2. For fair comparison, all the parameters that the LTAI and the baseline models share are set to have the same values, and for other parameters that uniquely belong to the baseline models, the values are exhaustively tuned as done in the LTAI. Finally, we note that all parameters are tuned using the training set, and test dataset is used only for the testing purpose.

7 Evaluation

We conduct the evaluation of the LTAI with three different quantitative tasks, along with one qualitative analysis. In the first task, we check whether using citation and authorship information in the LTAI helps increase the word-level predictive performance. In the second and third tasks, we measure the predictability of the LTAI regarding missing publication-publication linkage and author-publication linkage. With these two tasks, we compare the predictive power of the LTAI with other comparison models and use MRR as evaluation metric. Finally, we observe famous researchers' topical authority scores generated by the LTAI and investigate how these scores capture notable academic characteristics of the researchers.

7.1 Word-level Prediction

In the LTAI, citation and authorship information affect per-document topic proportions, as can be confirmed in equation 3. This joint modeling of content and linkage structure, compared to vanilla LDA that uses content data only, yields better performance in terms of predicting missing words in documents. In this task, we use log-predictive probability, a metric that is widely used in other researches for measuring model fitness (Teh et al., 2006; Asuncion et

---

[2]Although we do not present thorough sensitivity analysis in this paper, we confirm that the performance of our model was robust against adjusting the parameters within a factor of 2.

(a) CORA   (b) Arxiv-Physics

—●— LTAI   ─●─ LTAI-C   ···●··· LTAI-SEP   ─►─ Topical h-index
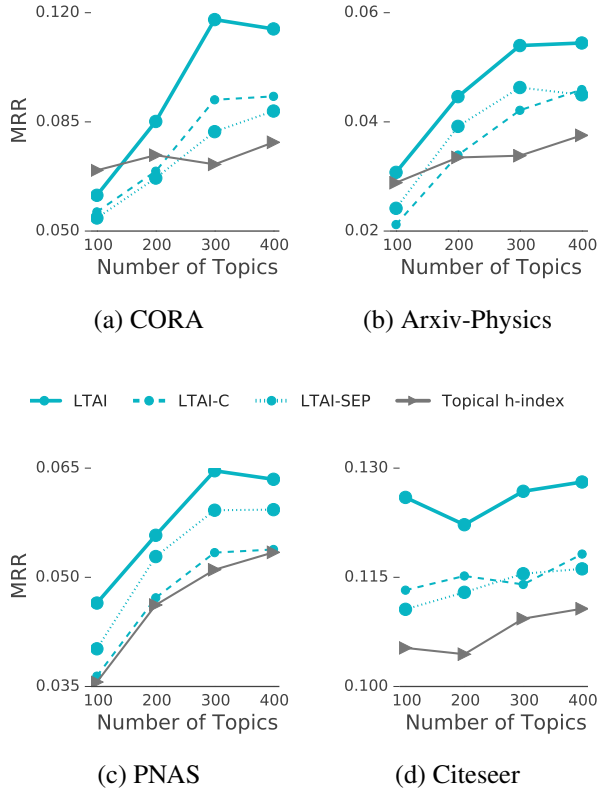
(c) PNAS   (d) Citeseer

Figure 6: Author prediction results. The task is to find out who the author of a cited paper is, given all the citing papers. For all cases, the LTAI performs better than the other methods.

al., 2009; Hoffman et al., 2013). For each corpus, we separate one third of the documents as the test set. For all documents in each test set, we use half of the words for training per-document topic proportion θ and predict the probability of word occurrence regarding the remaining half. Specifically, the predictive probability for a word in a test set w_{new} with respect to the given words w_{obs} and the training document $\mathcal{D}_{train}$ is computed using the equation $p(w_{new}|\mathcal{D}_{train}, w_{obs}) = \sum_{k=1}^{K} \mathbb{E}_q[\theta_k]\mathbb{E}_q[\beta_{k,w_{new}}]$.

Figure 4 illustrates the per-word log-predictive probability in each corpus. We confirm that when using the LTAI, the log predictive probability converges at a higher value compared to the result using LDA. Also, when we corrupt the link structure from 10% to 30% the predictive performances of the LTAI gradually decrease. Thus, the LTAI's superior predictive performance is attributed to its usage of correct citations rather than the algorithmic bias.

7.2 Citation Prediction

We evaluate model predictability regarding which publication is originally citing a certain publication. Specifically, we randomly remove one citation from each of the documents in the test set. To predict the citation link between publications, we first compute the probability that publication j cites i from $p(x_{i \leftarrow j}|z, \mathcal{A}_i, \pi_i) \propto \sum_{a \in \mathcal{A}_i} \pi_{i \leftarrow ja} \mathcal{N}(x_{i \leftarrow j}|\bar{z}_i^\top \text{diag}(\eta_a)\bar{z}_j, c_+^{-1})$. Given the topic proportion of the cited publication θ_i and the topical authorities of the authors η_a, we compute which publication is more likely to cite the publication. Based on our model assumption in subsection 3.2, using topical authority increases the performance of predicting linkage structure.

In Figure 5, the LTAI yields better citation prediction performance than other models for all datasets and with the most number of topics. Since the LTAI incorporates topical authority for predicting citations, it performs better than RTM, which does not discover topical authority. We can attribute the better performance of the LTAI compared to ALTM and DACTM to the LTAI's multiple model assumptions explained in section 3. We note that DACTM requires additional information such as citation location and sentence structure, and thus, is only applicable for limited kinds of datasets.

7.3 Author Prediction

For author prediction, we randomly remove one of the authors from documents in the test set while preserving citation structures. Similar to citation prediction, we predict which author is more likely to write the cited publication based on the topic proportions of cited publication i and a set of citing publications $\mathcal{J}$. We approximate the probability of researcher a being an author of publication i from $p(a|z, \eta_a, x_{i \leftarrow j}) \propto \prod_{j \in \mathcal{J}} \mathcal{N}(x_{i \leftarrow j}|\bar{z}_i^\top \text{diag}(\eta_a)\bar{z}_j, c_+^{-1})$. Because the mixture proportion of an unknown author $\pi_{i \leftarrow ja}$ cannot be obtained during posterior inference, we assume the cited publication is written by a single author to approximate the probability. For author prediction, we choose the author that maximizes the above probability. In Figure 6, the LTAI outperforms the comparison models in most of the settings.

Author	h-index	# cite	# paper	Representative Topic	T Authority
D Padua	12	291	21	parallel, efficient, computation, runtime	10.36
V Lesser	11	303	48	interaction, intelligent, multiagent, autonomous	11.92
M Lam	11	440	20	memory, processor, cache, synchronization	12.74
M Bellare	11	280	43	scheme, security, signature, attack	13.21
L Peterson	10	297	24	operating, mechanism, interface, thread	9.28
D Ferrari	10	377	18	traffic, delay, bandwidth, allocation	14.16
O Goldreich	9	229	49	proof, known, extended, notion	12.57
M Jordan	9	263	27	approximation, intelligence, artificial, correlation	10.15
D Culler	9	565	30	operating, mechanism, interface, thread	12.37
A Pentland	8	207	39	image, motion, visual, estimate	10.82

Table 2: Authors with the highest h-index scores and their statistics from the CORA dataset. We show the authors with their h-index, number of citations (# cite), and number of papers (# paper), representative topic, and their topical authority (T Authority) of the corresponding topic. We show that while the authors have the highest h-indices with lots of papers written and lots of citations earned, the topics that the authors exert authority varies.

7.4   Qualitative Analysis

To stress our model's additional characteristics that are not observed in the quantitative analysis, we look at the assigned topical authority indices as well as other statistics of some researchers in the dataset. In the analyses, we set the number of topics to 100, and use CORA dataset for demonstration.

We first demonstrate famous authors' authoritative topics that can be unveiled using our model. In Table 2, we list top 10 authors with highest h-indices along with their number of citations, number of papers, and their representative topics. Authors' representative topics are the topics with highest authority scores. In the table, we observe that all authors with top h-indices have written at least 18 papers and earned at least 207 citations, which are the top 0.8% and 0.2% values respectively. However, their authoritative topics retrieved by the LTAI do not overlap for any of the authors. This table illustrates that each of the top authors in the table exerts authority on different academic topics that can be captured by the LTAI, while the authors commonly have highest h-index scores as well as other statistics.

We now stress attributes of topical authority index that are different from other topic irrelevant statistics. From Tables 3 to 5, we show four example topics extracted by our model and list notable authors within each topic with their topical authority indices, h-indices, number of citations, and number of papers. In the tables, we first find that all four

authors with highest topical authority values, Monica Lam, Alex Pentland, Michael Jordan, and Mihir Bellare are also listed in the topic-irrelevant authority rankings in Table 2. From this, we confirm that authority score of the LTAI has a certain degree of correlation to other statistics, while it splits the authors by their authoritative topics.

At the same time, the topical authority score correlates less with topic-irrelevant statistics than those statistics correlate with themselves. In Table 5, Oded Goldreich has lower topical authority score for the computer security topic while having higher topic irrelevant scores than the above four researchers, because his main research filed is in the theory of computation and randomness. We can spot authors who exert high authority on multiple academic fields, such as Tomaso Poggio in Table 3 and in Table 4. Similarity, when comparing Federico Girosi and Tomaso Poggio in Table 4, the two researchers have similar authority indices for this topic while Tomaso Poggio has higher values for the other three topic-irrelevant indices. This is a reasonable outcome when we investigate the two researchers' publication history. Federico Girosi has relatively focused academic interest, with his publication history being skewed towards machine-learning-related subjects, while Tomaso Poggio has broader topical interests that include computer vision and statistical learning, while also co-authoring most of the papers that Federico Girosi wrote. Thus, Federico Girosi

Topic: image, motion, visual, estimate, robust, shape, scene, geometric				
Rank	Author	Topical Authority	h-index	# cite	# paper
1	A Pentland	10.82	8	207	39
2	J Fessler	9.09	6	92	26
3	T Poggio	8.22	6	178	27
4	S Sclaroff	7.61	3	69	11
5	K Toyama	6.65	4	41	10

Table 3: Authors who have a high authority score in a computer vision topic.

Topic: approximation, intelligence, artificial, correlation, support, recognition, model, representation				
Rank	Author	Topical Authority	h-index	# cite	# paper
1	M Jordan	10.15	9	263	27
2	M Warmuth	9.57	8	160	17
13	T Poggio	3.48	6	178	27
17	F Girosi	3.22	3	101	9
34	M Jones	2.06	7	151	20

Table 4: Authors who have a high authority score in an artificial intelligence topic.

has a similar authority index for this topic but has lower authority indices for other topics than Tomaso Poggio.

Also, our model is able to capture topic-specific authoritative researchers that have relatively low topic-irrelevant scores. For example, researchers such as Stan Sclaroff and Kentaro Toyama are the top 5 authoritative researchers in a computer vision topic according to the LTAI, but it is difficult to detect these researchers out of many other authoritative authors using the topic-irrelevant scores.

Finally, the LTAI detect researchers' topical authority that is peripheral but not negligible. Mark Jones in Table 4, who has high h-index, a number of citations, and wrote many papers, is a researcher whose academic interest lies in programming language design and application. However, while most of his papers' main topics are about programming language, he often uses inference techniques and algorithms in machine learning in his papers. Our model captures that tendency and assigns a positive authority score for machine learning to him.

Topic: scheme, security, signature, attack, threshold, authentication, cryptographic, encryption				
Rank	Author	Topical Authority	h-index	# cite	# paper
1	M Bellare	13.21	11	280	43
2	P Rogaway	11.98	7	117	13
3	H Krawczyk	7.29	6	75	15
4	R Canetti	7.13	4	40	10
9	O Goldreich	3.70	9	229	49

Table 5: Authors who have a high authority score in a computer security topic.

8 Conclusion and Discussion

We proposed Latent Topical Authority Indexing (LTAI) to model the topical-authority of academic researchers. Based on the hypothesis that authors play an important role in citations, we specifically focus on their authority and develop a Bayesian model to capture the authority. With model assumptions that are necessary for extracting convincing and interpretable topical authority values for authors, we have proposed speed-up methods that are based on stochastic optimization.

While there is prior research in topic modeling that provides topic-specific indices when modeling the link structure, these do not extend to individual indices, and most previous citation-based indices are defined for each individual but without considering topics. On the other hand, our model combines the merits of both topic-specific and individual-specific indices to provide topical authority information for academic researchers.

With four academic datasets, we demonstrated that the joint modeling of publication and author related variables improve topic quality, when compared to vanilla LDA. We quantitatively manifested that including authority variables increases the predictive performance in terms of citation and author predictions. Finally, we qualitatively demonstrated the interpretability by topical-authority outcomes of the LTAI from the CORA corpus.

Finally, there are issues that can be dealt with in future work. We do not consider time information in terms of when papers are published and when pairs of papers are linked; we can use datasets that incorporate timestamps to enhance the model capability to predict future citations and authorships.

References

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *UAI*.

Krisztian Balog and Maarten de Rijke. 2007. Determining expert profiles (with an application to expert finding). In *IJCAI*.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *JMLR*, 3:993–1022.

Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *EMNLP*.

Cornelia Caragea, Florin Bulgarov, and Rada Mihalcea. 2015. Co-training for topic classification of scholarly data. In *EMNLP*.

Jonathan Chang and David Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1):124–150.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *EMNLP*.

Christopher Diehl, Galileo Namata, and Lise Getoor. 2007. Relationship identification for social network discovery. In *AAAI*.

Laura Dietz, Steffen Bickel, and Tobias Scheffer. 2007. Unsupervised prediction of citation influences. In *ICML*.

Yuxiao Dong, Reid Johnson, and Nitesh Chawla. 2015. Will this paper increase your h-index?: Scientific impact prediction. In *WSDM*.

James Foulds and Padhraic Smyth. 2013. Modeling scientific impact with topical influence regression. In *EMNLP*.

Eugene Garfield. 2006. The history and meaning of the journal impact factor. *JAMA*, 295(1):90–93.

Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg. 2003. Overview of the 2003 KDD Cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151.

Yuan He, Cheng Wang, and Changjun Jiang. 2015. Discovering canonical correlations between topical and topological information in document networks. In *CIKM*.

Jorge Hirsch. 2005. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572.

Matthew Hoffman, David Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *JMLR*, 14:1303–1347.

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *ICDM*.

Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and Giles Lee. 2015. A neural probabilistic model for context based citation recommendation. In *AAAI*.

Zhuoren Jiang. 2015. Chronological scientific information recommendation via supervised dynamic topic modeling. In *WSDM*.

Saurabh Kataria, Prasenjit Mitra, Cornelia Caragea, and Giles Lee. 2011. Context sensitive topic models for author influence in document networks. In *IJCAI*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. Topic-link LDA: joint models of topic and author community. In *ICML*.

Qing Lu and Lise Getoor. 2003. Link-based classification. In *ICML*.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163.

Michael Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social studies of science*, 5(1):86–92.

Ramesh Nallapati, Amr Ahmed, Eric Xing, and William Cohen. 2008. Joint latent topic models for text and citations. In *SIGKDD*.

Willie Neiswanger, Chong Wang, Qirong Ho, and Eric P. Xing. 2014. Modeling citation networks using latent random offsets. In *UAI*.

Sanjay Purushotham, Yan Liu, and C.-C. Jay Kuo. 2012. Collaborative topic regression with social matrix factorization for recommendation systems. In *ICML*.

Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. How does clickthrough data reflect retrieval quality? In *CIKM*.

Adrian Raftery, Xiaoyue Niu, Peter Hoff, and Ka Yee Yeung. 2012. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919.

Jasson Rennie and Nathan Srebro. 2005. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics*, 22(3):400–407.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *UAI*.

Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic matrix factorization. In *NIPS*.

Yanchuan Sim, Bryan Routledge, and Noah Smith. 2015. A utility model of authors in the scientific community. In *EMNLP*.

Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *EMNLP*.

Yuancheng Tu, Nikhil Johri, Dan Roth, and Julia Hockenmaier. 2010. Citation author topic model in expert search. In *COLING*.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Ellen Voorhees. 1999. The TREC-8 question answering track report. In *TREC*.

Chong Wang and David Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *SIGKDD*.

Jingang Wang, Dandan Song, Zhiwei Zhang, Lejian Liao, Luo Si, and Chin-Yew Lin. 2015. LDTM: A latent document type model for cumulative citation recommendation. In *EMNLP*.