


# Collective topical PageRank: a model to evaluate the topic-dependent academic impact of scientific papers

Yongjun Zhang<sup>1,2</sup>  · Jialin Ma<sup>2</sup> · Zijian Wang<sup>1</sup> · Bolun Chen<sup>2</sup> ·  
Yongtao Yu<sup>2</sup>

Received: 8 August 2017  
© Akadémiai Kiadó, Budapest, Hungary 2017

**Abstract** With the explosive growth of academic writing, it is difficult for researchers to find significant papers in their area of interest. In this paper, we propose a pipeline model, named collective topical PageRank, to evaluate the topic-dependent impact of scientific papers. First, we fit the model to a correlation topic model based on the textual content of papers to extract scientific topics and correlations. Then, we present a modified PageRank algorithm, which incorporates the venue, the correlations of the scientific topics, and the publication year of each paper into a random walk to evaluate the paper's topic-dependent academic impact. Our experiments showed that the model can effectively identify significant papers as well as venues for each scientific topic, recommend papers for further reading or citing, explore the evolution of scientific topics, and calculate the venues' dynamic topic-dependent academic impact.

**Keywords** Topic model · PageRank · Scientific evaluation

## Introduction

Academic papers play an important role in academic research. By reading significant papers in their area of interest, researchers learn about the problems that have been studied thoroughly, the problems to be solved, and the challenges and potential directions of research in the future. However, the explosive growth of papers makes them difficult to sort to find relevant works. Researchers have to spend a lot of time and effort to locate papers in their area of interest. Some approaches have been developed to evaluate the academic impact of papers, which can help researchers to screen for important papers.

---

✉ Yongjun Zhang  
13511543380@139.com

<sup>1</sup> College of Computer and Information, Hohai University, Nanjing, China

<sup>2</sup> Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an, China

They can be roughly classified into 2 categories: citation count approaches and random walk approaches.

The citation count approaches simply evaluate a paper's academic impact in terms of the citation frequency. The general idea of these approaches is that the more a paper is cited by other papers, the higher its academic impact. Gross and Gross (1927) first proposed the use of the citation count to evaluate the importance of scientific works; then, it was extended to other applications, such as the assessment of journal impact factors (Garfield 2006), national science policies, disciplinary development (MacLean et al. 1998), and the effective improvement of ad-hoc retrieval (Meij and De Rijke 2007; Fujii 2007). There is a major drawback to citation count approaches: they ignore the authority differences of the citing papers. Thus, in a citation count approach, a citation from a paper published in an ordinary venue has the same worth as a citation from a paper published in *Nature*.

Random walk models, such as PageRank (Page et al. 1999), which was originally designed to measure the importance of web pages, have been transferred to estimate a paper's academic impact. They build a directed network in terms of the citation relationships of the papers. The directed network consists of the nodes represented by papers, and the directed edges correspond to the citation of papers, in which a directed edge from the citing paper node to the cited paper node identifies a citation. Then, a random walk is performed to evaluate the importance of the papers. Rather than roughly making use of the paper's citation frequency, these models also take into account the authority of the citing papers to evaluate the paper's academic impact. Therefore, a paper will have a high impact score if it is cited by many other authoritative papers.

Both the citation count and random walk approaches suffer from the following challenges. (1) They show a preference for older papers, since the older papers have more opportunities to be cited, which leads to a serious bias towards older papers, running counter to the notion that researchers want to read the latest papers. (2) Both approaches only make use of the citation relationships of papers to evaluate a paper's academic impact, while the textual content of the papers is ignored, which results in a final impact score that has no connection to the research field. This is a very severe defect, since researchers mostly focus on their areas of interest. Therefore, a field-dependent local paper academic impact evaluation approach would be preferred by researchers.

The state-of-the-art topical PageRank model (Jardine and Teufel 2014) (TPM) tries to overcome the aforementioned drawbacks by incorporating latent-Dirichlet-allocation (LDA) topics into PageRank. The TPM employs the LDA (Blei et al. 2003b) to extract topics from the textual content of papers and approximates the research fields as LDA topics. For each topic, a modified PageRank, which considers the topic proportions of both the cited papers and the citing papers, evaluates the topical impact of each paper. In addition, a linear age-taper strategy is employed to decrease the preference for old papers. The TPM can generate the topic-dependent impact of papers, however, it suffers from the following limits. First, it favors the latest papers due to the linear age-taper strategy. Second, it ignores the interaction of different topics since the LDA model it employs fails to extract the correlations of the topics.

We propose the Collective Topic PageRank Model (CTPM) to address the limits of the TPM. Unlike the TPM, the CTPM employs the correlation topic model (Blei et al. 2007) (CTM) (instead of the LDA) to extract scientific topics as well as their correlations. Then, for each extracted scientific topic, it uses a modified PageRank, which modifies both the bias probability and the transition probability by taking the topic proportions, the prestige of the paper's venue, and the correlations of scientific topics into account to evaluate the

topical academic impact of papers. The CTPM also weights the bias probability with an exponential function of the paper's publication year. The improvements made by the CTPM are summarized as follows:

1. Researchers prefer to read the papers published in prestigious venues to stay current with significant research achievements. The more famous the paper's venue is, the more likely the paper is chosen by researchers. Thus, the CTPM considers the prestige of the venues and incorporates it into the random walk.
2. The CTPM makes use of the correlations of different research fields to reflect their interactions. In TPM, the LDA model is employed to capture topics, and each topic is mapped to a research field. However, the LDA model ignores the correlations of different topics, so that the interactions of different research fields are not utilized by the TPM, whereas the CTPM employs the CTM to explore the correlations of different research fields and then incorporates them into the random walk.
3. As a side effect of the impact evaluation of papers, the CTPM can also evaluate the venue's topic-dependent academic impact. There are several methods used to calculate the journal's impact, e.g., Impact Factor (IF). These approaches can determine the time-varying global academic impact score of a journal. However, the score is topic-independent, so that it cannot indicate how important the venue is in a particular research field. The CTPM overcomes these shortcomings and generates the dynamic topic-dependent impact scores of the venues. A comparison between the CTPM and TPM is shown in Table 1.

We performed some experiments on the AMiner ACM-Citation-network (Tang et al. 2008b) dataset;<sup>1</sup> the experimental results were the following. (1) For each scientific topic, our model identified significant papers and venues. The ranks of the papers and venues were reasonable according to the evaluations. (2) In the evaluation task of the paper's reference items' reintroduction, the CTPM significantly outperformed the state-of-the-art approach, with an 8.76% mean-average precision (MAP) improvement. (3) The CTPM has a wide range of applications, such as topic evolution exploration and the topic-dependent academic impact factor evaluation of venues.

## Related works

Our CTPM is a pipe model combing topic models with PageRank. Thus, prior research works related to the CTPM advanced along 3 lines: using topic models to explore scientific topics, personalizing the PageRank, and combining the two approaches.

### Using topic models to explore scientific topics

Many researchers have observed that the LDA and its extensions are effective tools for exploring scientific literature. The original LDA was proposed by Blei to uncover topics in document collections. As a supervised technique, it can exploit the word co-occurrence patterns in documents to extract semantically meaningful clusters of words (namely, topics) without any labeled data. The most comprehensible way to understand the LDA is to regard it as a mixed component model, in which each topic  $\phi_k$  is a component represented by a multinomial distribution over the words, each document  $d$  is a mixture of

<sup>1</sup> <https://cn.aminer.org/citation>.

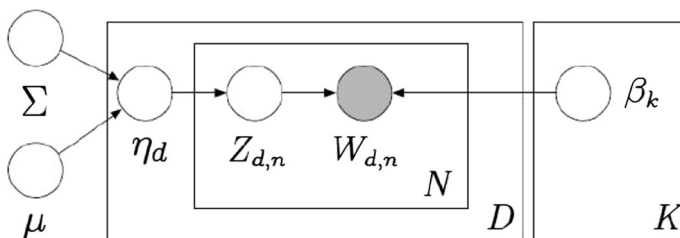
**Table 1** The comparison between the CTPM and TPM

Strategy	TPM	CTPM
Makes use of the papers citations	Yes	Yes
Makes use the paper's textual content	Yes	Yes
Topic model employed	LDA	CTM
Makes use of the paper's venue	No	Yes
Makes use of the correlations of topics	No	Yes
How to decrease the age-bias	Linear age-taper	Gaussian age-taper

components, and  $\theta_{d,k}$  is the proportion of the topic  $k$  in the document  $d$ . In this way, the LDA maps documents to latent topic spaces and allows them to be processed in the lower-dimensional topic space. Since the LDA was proposed, many researchers employed it to explore scientific topics in scientific literature. For instance, Griffiths and Steyvers (2004) found that the topics extracted by the LDA have a close correlation to scientific topics when the LDA is applied to the scientific literature. They created a LDA model for the abstracts of papers published in *PNAS* from 1991 to 2001 and developed a Markov chain Monte Carlo algorithm to extract the scientific topics. Their results show that the extracted scientific topics are consistent with the categories provided by the authors of the papers, which suggests the topics discovered by the LDA should be suitable counterparts of the scientific topics.

However, the LDA uses the *Dirichlet* distribution as the prior distribution of document-topic multinomial distribution  $\theta_d$ , which makes it unable to capture the correlations of the topics, thus hindering the exploration of the interactions between scientific research fields when it is applied to scientific literature. To address this limit, the CTM uses a *Gaussian* distribution  $N(\eta|\mu, \Sigma)$  to model the natural parameterization of  $\theta_d$ , which we denote as  $\eta_d$ . Figure 1 illustrates the graphical model of the CTM. The CTM can capture the correlations of the topics because it uses the covariance matrix  $\Sigma$ , in which each element  $\Sigma_{i,j}$  indicates the correlation between topic pair  $i$  and  $j$ . The experimental results from the articles in *Science* published from 1990 to 1999 show that the CTM fit the data better than the LDA and is an effective tool to extract scientific topics and explore their correlations. Thus, our CTPM employs the CTM rather than the LDA to extract scientific topics and capture their correlations.

Yau confirmed that topic models are powerful tools to explore scientific topics in scientific literature. He employed 3 topic models to cluster scientific papers: the LDA, CTM, and Hierarchical LDA (HLDA) (Blei et al. 2003a). He gathered papers in 7 research fields from the Web of Science manually, and then compared the LDA, CTM, and HLDA



**Fig. 1** Graphical model representation of the CTM

with the baseline  $K$ -means cluster method. His experimental results show the clusters generated by the 3 topic models are close to the research fields, which indicates the superiority of these topic models in exploring scientific topics when they are applied to scientific literature.

Compared to other text documents, scientific literature has some significant non-textual features, such as the authors, venues, publication time, and references. However, the LDA, CTM, and HLDA ignore these non-textual features and only make use of the textual content to train the models, which leaves room for improving their performance. Some other topic models sought to incorporate the non-textual features of the scientific literature into the textual content generative process. For instance, the Author-Topic (AT) model (Steyvers et al. 2004) represents each author by a multinomial distribution over topics; the words in a multi-author paper are assumed to be generated by the mixture of each author's topic. Tang et al. (2008a) proposed the Author-Conference-Topic (ACT) model to simultaneously model the textual content, authors, and publication venues. They implemented 3 variants of the ACT (named ACT1, ACT2, and ACT3). Erosheva et al. (2004) proposed the Mixed-Membership model to link the paper's references with its textual content. Their model represents each topic by both a multinomial distribution over the words  $\phi_k$  and another multinomial distribution over the cited papers  $\psi_k$ . Then, each token  $w_{d,n}$  of each paper is assumed to be generated by its topic  $z_{d,n}$  and the component  $\phi_{z_{d,n}}$ , while each reference item  $c_{d,r}$  is generated by the corresponding topic  $z_{d,r}$  and the component  $\psi_{z_{d,r}}$ . Wang et al. (2013) also used the citation relationships of papers to develop the Citation-LDA model. Their work aimed to explore the evolution of scientific topics. These extensional topic models are probabilistic generative models, which integrate the non-textual data into the generative process of the textual content. This integration improves the performance, but includes a limit: the unseen documents must have the same non-textual features as those in the training documents when the models are used for prediction. This limitation may be fatal for some special tasks, for instance, in recommending papers to read or cite according to an abstract's input, in which the input only includes textual content without any other non-textual features. In contrast, the CTPM makes use of the non-textual data in a different way. It uses a pipelined method to concatenate the topic model to the PageRank-based random walk process. During the first stage, the textual content is modeled by the CTM; during the next state, a modified PageRank algorithm incorporates the non-textual data. The pipelined strategy makes the CTPM more flexible and able to be applied to more tasks, while making full use of the non-textual features in the training data.

## Personalized PageRank algorithms

The CTPM we propose uses a modified PageRank algorithm to generate a topic-dependent score for each paper, which is a revision of the original PageRank. The original PageRank makes use of the link structure of web pages to evaluate their authorities; it was effectively used in Google's search engine. It modeled the authorities of web pages by

$$P_d^t = \alpha B_d + (1 - \alpha) \sum_{d' \in D} T(d' \rightarrow d) P_{d'}^{t-1} \quad (1)$$

$$B_d = \frac{1}{\|D\|} \quad (2)$$

$$T(d' \rightarrow d) = \frac{1}{\|N_{d'}\|} I(d \in L_{d'}) \quad (3)$$

where  $P_d^t$  is the score of page  $d$  in the  $t$ th iteration, and it can be regarded as the probability of choosing page  $d$  from all the web pages.  $B_d$  is called the bias probability, which is used to denote how likely page  $d$  is chosen at random,  $T(d' \rightarrow d)$  denotes the transition probability of page  $d'$  to page  $d$ , and  $\alpha$  weights the probability of the random choice.  $L_{d'}$  is the web pages that  $d'$  links to. From the perspective of probability, the score of page  $d$  can be roughly regarded as the probability of choosing it, and the procedure for choosing it includes the following steps:

1. Select a way to choose page  $d$ . There are 2 ways to it: the random choice way and transition way. The random choice way chooses the page from all the pages at random, while the transition way chooses a page  $d'$  and then transfers to it from  $d'$ . The first step selects the random choosing way with probability  $\alpha$  and the transition way with probability  $1 - \alpha$ ;
2. Choose  $d$  from all pages with a probability  $B_d$  if the way of choosing is the random choice way;
3. If the way of choosing page  $d$  is the transition way, choose any paper  $d'$  with a probability  $P_{d'}$ , then transfer it from page  $d'$  to it with the transition probability of  $T(d' \rightarrow d)$

In the original PageRank, Eq. (2) assumes that each page is equally important, which runs contrary to common sense. Some personalized PageRank models sought to modify the bias probability  $B_d$ . For instance, Gyngyi et al. (2004) proposed the TrustRank algorithm to combat spam pages. They manually selected some seed pages from fewer than 200 sites to identify them as reputable pages and assigned them a high bias probability ( $B_d$ ) to each of them. Their work requires manual selection of seed pages, making it unable to scale to the applications handling huge amount of pages. Based on the TrustRank algorithm, Wu et al. proposed the Topical TrustRank (TTP) algorithm (Wu et al. 2006), in which they partitioned the seed pages into groups in terms of their topical information, then calculated the trust score of each page for each topic separately. They aimed to improve TrustRank by introducing the topics. However, the manual selection and grouping of the seed pages makes their work also subject to the same limit of TrustRank. Gori and Pucci (2007) modified the original PageRank to recommend research papers. Their paper recommendation system accepts a paper  $p$  containing references as an input query; it then recommends papers highly relevant to  $d$ . They modified the bias probability according to the citation relation of papers.

PageRank was also used to evaluate the academic impact of papers. In this case, each citation was regarded as a directed edge from the citing paper to the cited paper. However, the application of original PageRank to scientific papers suffers from a severe drawback: it prefers the older papers because they have more opportunities to be cited. There are some researchers seeking to address this drawback. For instance, Walker et al.'s (2007) CiteRank algorithm used an exponential function of the age to favor more recent papers. The TPM proposed by Jardine and Teufel (2014) decreased the score of older papers with a linear function of their publication age.

There are several researchers who observed that PageRank only works on the citations and does not make use of the textual content of the documents; they sought to integrate the textual content into PageRank's random walk process. For instance, Richardson and Domingos (2002) first implemented PageRank in the information retrieval field. They proposed the query-dependent PageRank (QD-PageRank) to calculate the query-specific scores for web pages. For a specific query  $q$ , their model calculated a query-dependent PageRank score for each page  $d$ , which depended on the textual relevance between  $q$  and  $d$ . However, to obtain the query-dependent PageRank score, the relevance of each page  $d$  to query  $q$  must be calculated at the run time, which is an expensive time cost and limits the model's scalability. Haveliwala (2003) proposed a topic-sensitive PageRank, which modified  $B_d$  to rank pages with respect to 16 manually created topics. Instead of modifying  $B_d$ , Pal and Narayan (2005) proposed a surfer model by changing the transition probability  $T(d' \rightarrow d)$ . Their model split the transition  $T(d' \rightarrow d)$  into within-topic transition and cross-topic transition. These topic-dependent personalized PageRank models require manually predefined topics, which limits their application to massive amounts of data.

### The models combining the topic model with PageRank

We are not the first researchers to combine the topic model with the PageRank to evaluate the topic-dependent scores of scientific literature. To our knowledge, Yang et al. (2009) first combined the LDA with PageRank to calculate the topic-dependent scores of papers. Their work applied the LDA to automatically extract topics from scientific papers and then combine them with PageRank. The model they proposed changed the transition probability to the weighted sum of the intra-topic transition probability and topic-across transition probability. However, their ultimate model obtained the best results when the intra-topic transition probability weight equals 1, thus only the intra-topic transition was considered. Moreover, the calculation of the intra-topic transition was independent of whether the cited paper and the citing paper were relevant to the topic or not, which does not conform to common sense and is not reasonable. Ding (2011) combined the ACT (Tang et al. 2008a) model with a weighted PageRank to rank authors. Her model applied the ACT model to output each author's topic distribution  $P(k|a)$ ; she proposed 2 ways to combine the ACT model with the PageRank: a simple combination ( $I\_PR$ ) or using a topic distribution as a weighted vector for PageRank ( $PR\_t$ ). The aim of her work is different from our work. She sought to rank authors on the author co-citation network, while our work is based on the paper citation network and aims to rank scientific papers.

Other works closest to ours are Yan's work (2014) and the TPM (Jardine and Teufel 2014). The former used the ACT model (Tang et al. 2008a) to extract scientific topics, then modified the bias probability of each paper in terms of its topic proportion, and calculated the topic-based PageRank scores in heterogeneous scholarly networks. The main differences between it and our work are as follows. (1) We use the paper citation network while their model is based on heterogeneous scholarly networks. The building of heterogeneous scholarly networks may be expensive and unavailable in some situations. (2) They used the ACT model, while we employ the CTM. The use of the ACT model may be helpful to improve the exploration of scientific topics, but it caused their model to lose flexibility because the ACT required unseen documents with author and venue data. In contrast, the model in this paper uses the CTM, allowing it to work on unseen documents with only textual data and explore both scientific topics and their correlation. The TPM fits a LDA model on the textual contents of papers to extract scientific topics automatically and

simultaneously represent each paper by a multinomial distribution over the topics, which indicates the paper's main scientific topics. For each extracted scientific topic, the TPM then uses a modified PageRank to rank papers and yield the topic-dependent score of each paper. Their experimental results on the ACL Anthology Network dataset showed its superiority on the tasks of restoring the reference lists and automatically creating reading lists. However, the ACL Anthology Network dataset only consists of the papers about the natural language processing (NLP) research field; the performance of their model may degrade when it is applied to datasets containing the papers from various research fields, which is confirmed by our experiment in “[Recommend papers to read](#)” section. In contrast, our work takes into account the important venue factor of papers and incorporates it into the random walk process, which provides a significant improvement in the model's performance.

## The collective topic PageRank model (CTPM)

The CTPM calculates the topic-dependent academic scores of papers with the following 3 steps. (1) It employs the CTM to determine the topic proportion  $\theta_{d,k}$  of each paper as well as the correlation  $R_{i,j}$  between any 2 topics  $i$  and  $j$ . (2) It incorporates the topic impacts of the venues and the correlations of the topics into PageRank to improve the calculation of both the bias probability and the transition probability. (3) It adjusts the bias probability weight  $\alpha$  dynamically by the age of each paper. The general iterative calculation can be defined as follows:

$$\text{TPR}^{t+1}(d|k) = \alpha_d B(d|k) + (1 - \alpha_d) \sum_{d' \in D} T(d|d', k) \text{TPR}^t(d'|k) \quad (4)$$

## Fit the CTM to papers to extract scientific topics and explore their correlations

The types of scientific expertise needed in different research fields may be interdependent. For example, a paper about genetics may be concerned with health and disease. However, the original LDA ignores the correlations of the topics, which makes it incapable for uncovering the interactions in the scientific fields. Compared with the LDA, the CTM takes the logistic normal distribution as the document-topic distribution priority, which makes it able to explore the correlations of scientific topics when it is applied to scientific papers. Thus, our CTPM employs it to fit scientific papers.

We used the variational inference algorithm to determine each topic  $\phi_k$  as well as the topic multinomial distribution  $\theta_d$  of each paper. Then, we used a lasso regression to calculate the correlation matrix  $R^*$  (Blei et al. 2007). The element  $R_{i,j}^*$  of  $R^*$  identifies the correlation between the topics  $i$  and  $j$ . We normalized the topic correlation  $R^*$  by

$$R_{i,j} = \frac{R_{i,j}^*}{\sum_{j=1}^K R_{i,j}^*} \quad (5)$$

We calculated the topic relevancy of each paper with



$$r_{d,k} = \sum_{i=1}^K \theta_{d,i} R_{i,k} \quad (6)$$

Equation (6) combines the proportion of each topic  $i$  of the paper  $d$  with the correlation between the topics  $i$  and  $k$ . Thus, if the paper  $d$  focuses on a strongly correlated topic  $i$  of topic  $k$  rather than on topic  $k$  directly, it would be relevant to topic  $k$  due to the contribution of both the high  $P(i|d)$  and  $R_{i,k}$ .

## Incorporating the venue and the correlations of topics into PageRank

Consider the following 2 scenarios where researchers choose papers to read. In the first scenario, researchers choose the papers that they are interested in through academic search tools, e.g. Google Scholar. When the retrieved results are presented to them, they prefer to read papers that are published in prominent venues. This decision suggests that the papers published in prominent venues should be chosen with a high probability. In the other scenario, a researcher is reading paper  $d'$  that is highly related to his/her research field; then, the researcher picks some papers from the references of the paper  $d'$  to read further. Generally, the researcher would rather choose papers published in prominent venues than in more common venues, which suggests that the transition probability  $T(d|d', k)$  depends strongly on the prestige of the cited paper's venue.

Our CTPM modified both the bias probability  $B(d|k)$  and the transition probability  $T(d|d', k)$  in Eq. (4) to agree with the above 2 assumptions. In the CTPM, the bias probability is changed to

$$B^{t+1}(d|k) = \frac{\sqrt{V^t(v_d|k)r_{d,k}}}{\sum_{d' \in D} \sqrt{V^t(v_{d'}|k)r_{d',k}}} \quad (7)$$

where  $V^t(v_d|k)$  denotes the impact score of paper  $d$ 's venue in the topic  $k$  in the  $t$ th iteration. We also replaced  $\theta_{d,k}$ , which is the proportion of the topic  $k$  in the document  $d$ , with  $r_{d,k}$ , which is the relevance of the document  $d$  to the topic  $k$ , to take into account the correlations of the topics. Analogous to the bias probability, the transition probability is also changed to

$$T'(d|d', k) = \sqrt{B(d'|k) \frac{r_{d,k}}{\sum_{d'' \in L_{d'}} r_{d'',k}}} \quad (8)$$

$$T''(d|d', k) = \frac{r'(d|d', k)}{\sum_{d'' \in C_{d'}} T'(d|d'', k)} \quad (9)$$

$$T^{t+1}(d|d', k) = \frac{\sqrt{V^t(v_d|k)T''(d|d', k)}}{\sum_{d'' \in C_{d'}} \sqrt{V^t(v_{d''}|k)T''(d|d'', k)}} I(d \in L_{d'}) \quad (10)$$

where  $I(x)$  is the indicator function. Equations (8)–(10) show how the paper's topic-dependent score calculation depends on its venue. Inversely, the topic-dependent impact score of a venue also depends on the topic-dependent impact scores of the papers published in it. The more outstanding the papers of a venue in the topic, the higher its topic-

dependent impact score should be. Our CTPM addresses the calculation of the venue topic-dependent impact score with the following equation:

$$V^t(v_i|k) = \frac{\sum_{d \in v_i} \text{TPR}^t(d|k)/|v_i|}{\sum_j \sum_{d' \in v_j} \text{TPR}^t(d'|k)/|v_j|} \quad (11)$$

where  $|v_i|$  is the number of papers published in venue  $v_i$ . Equation (11) suggests that the topic-dependent impact score of a venue in the  $t$ th iteration depends on the average TPR of the papers published in it.

In order to draw back the preference to old papers, we employ a Gaussian decay function to weight the bias and transition probability. “Appendix 1” describes the use of Gaussian decay function for age-tapering papers in detail (Fig. 8).

## The iterative algorithm

The CTPM calculates the papers topic-dependent score iteratively for each topic. The iterative process is defined as Eq. 4. The calculation will stop when all the topic scores almost do not change. Algorithm 1 describes the iteration process (Fig. 2). To explain our algorithm clearly, we also give a simple example in “Appendix 2”.

## Experiments

### The experimental dataset

We used the AMiner ACM-Citation-network V8 dataset (Tang et al. 2008b) to fit the CTPM. The original dataset is extracted from ACM, it contains 2,381,688 papers and 10,476,564 citation relationships, in which each paper is associated with abstract, authors,

---

#### Algorithm 1 The iteration calculation of $TPR$

---

- 1: Fit the corpus of papers with the  $CTM$  to yield the topic-correlation matrix  $R^*$ , document-topic proportion  $\theta_{d,k}$  and topic-term probability  $\phi_{k,w}$
  - 2: Calculate the topic relevancy of each paper by the Equations (5) and (6)
  - 3: Calculate the weight of random choosing  $\alpha_d$  for each paper  $d$  by the Equation (25)
  - 4: For each topic  $k$  do
    - 5: Initialize  $TPR^0(d|k) = \frac{\gamma_{d,k}}{\sum_{d' \in D} \gamma_{d',k}}$  for each paper  $d$
    - 6: Initialize  $V^0(v|k)$  for each venue  $v$  by the Equation (11)
    - 7: Do until convergence
      - 8: Calculate the bias probability  $B^{t+1}(d|k)$  by the Equation (7)
      - 9: Calculate the transition probability  $T^{t+1}(d|d', t)$  by the Equations (8), (9) and (10)
      - 10: Calculate the  $TPR^{t+1}(d|k)$  according to the Equation (4)
      - 11: Calculate the venue's topic-dependent impact score  $V^{t+1}(v_i|k)$  by the Equation (11)
      - 12: Set  $t \leftarrow t + 1$
- 

**Fig. 2** The iterative calculation algorithm of the CTPM

year, venue, and title. It was preprocessed in 3 steps. The first step guaranteed all papers were unbroken. To accomplish this goal, the papers without any attributes of the title, abstract, venue, publication year, or references were removed. The second step consisted of constructing an effective internal citation network. In this step, we first removed the references not in the dataset, then removed the papers citing fewer than 3 other papers because their references were likely not identified correctly. The third step adjusted the venues of papers. In the original dataset, the venues are chaotic. For example, the venue named *ACM Transactions on Graphics (TOG) - SIGGRAPH 2012 Conference Proceedings* and another venue with the name *ACM Transactions on Graphics (TOG) - SIGGRAPH 2013 Conference Proceedings* should be the same venue. We manually merged the venues of similar names with a new name. For the given example, we merged the 2 venues with the new name *ACM Transactions on Graphics (TOG)*. The final dataset contained 523,720 papers and 2,908,093 citation relationships, as shown in Table 2.

### Fitting the CTM to the data set

We extracted the title and abstract of each paper to constitute the document dataset to train the CTM. A stop-word list<sup>2</sup> was used to prune the stop words. We also removed punctuation, numbers, individual characters, and the words that occurred in fewer than 5 papers or 10 times in the document dataset. Finally, we obtained a vocabulary of 64,859 words, which occurred a total of 48,323,764 times in the document dataset.

One of the key parameters of the CTM is the topic number  $K$ . Many strategies have been developed to choose it. For example, the hold-out perplexity is a common way to determine the best  $K$ . A lower hold-out perplexity of the topic model of  $K$  suggests greater predictive power. However, some research works suggest that topics discovered by the hold-out perplexity measurement may not be comprehensible. Thus, we used a balanced solution between the power predictive and comprehensible topics to find a preferable topic number  $K$  to train the CTM for our CTPM. First, we performed tenfold cross validations with the hold-out perplexity measurement to find reasonable topic number settings. We computed an average hold-out perplexity for each topic number  $K$  among 20 topic number settings, which varied from 10 to 200 with an interval 10 and found the preferable  $K$  is in the range of 80–120, as shown in Fig. 3.

We then used the average topic dissimilarity to find the best  $K$  in the settings of  $K = 80, 90, 100, 110, 120$ . The topic dissimilarity between 2 topics,  $T_i$  and  $T_j$ , was measured by the Jensen–Shannon divergence (JSD)<sup>3</sup> of their topic-word probability distributions, which measures how topic  $T_i$  is distinguished from topic  $T_j$ . For the given topics  $T_i$  and  $T_j$ , the topic dissimilarity between them was calculated as follows:

$$J(T_i, T_j) = \frac{1}{2}(D(T_i||M) + D(T_j||M)) \quad (12)$$

where  $M = \frac{1}{2}(T_i + T_j)$ ,  $D(T_i||M)$  is the Kullback–Leibler divergence<sup>4</sup> from  $M$  to  $T_i$ , which is defined as follows:

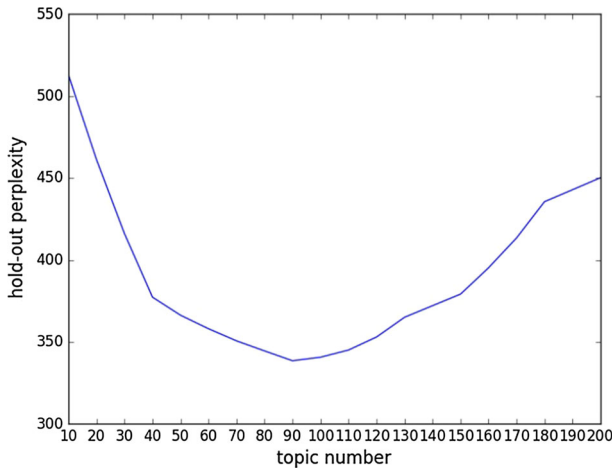
<sup>2</sup> <http://www.nltk.org/>.

<sup>3</sup> [https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon\\_divergence](https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence).

<sup>4</sup> [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_divergence](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence).

**Table 2** Data set statistics

Papers	Citation relationships	Venues	Publication year	Max cited numbers	Mean cited numbers
523,720	2,908,093	7348	1958–2015	3570	5.55



**Fig. 3** The tenfold cross-validated perplexity of different topic numbers for held-out documents. Lower numbers indicate greater predictive power from the CTM

$$D(T_i||M) = \sum_{t=1}^V T_{i,t} \log \frac{T_{i,t}}{M_t} \quad (13)$$

The average topic dissimilarity is an indicator of the topic model quality; a high value suggests that the topics discovered by the topic model are comprehensible. For a topic set  $TS = \{T_1, T_K\}$ , the average topic dissimilarity is:

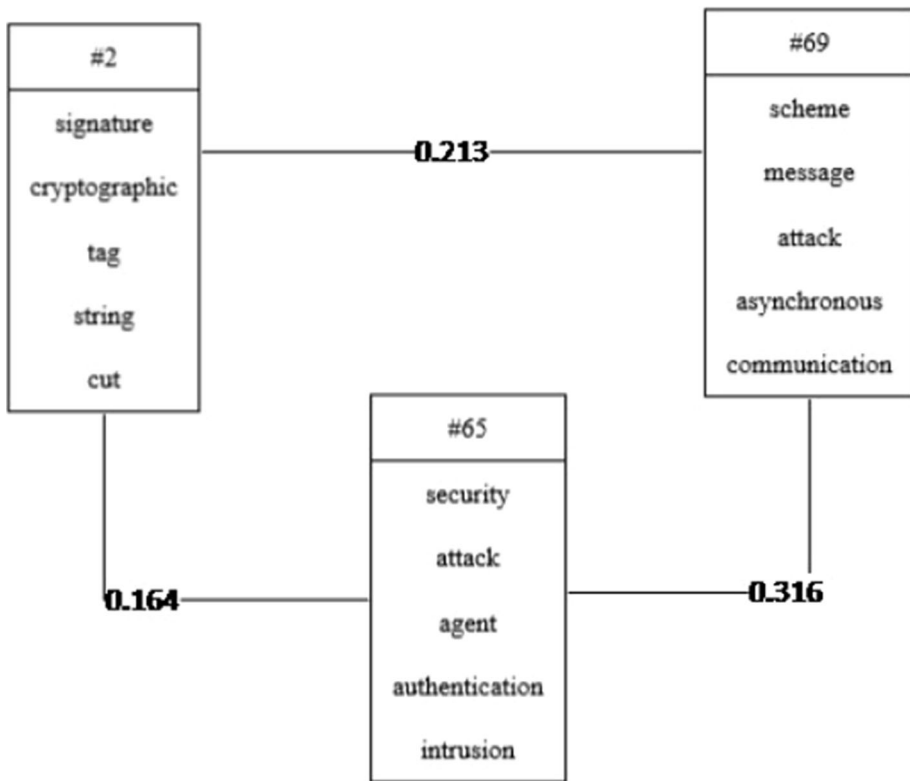
$$\text{avg}(T_1, T_K) = \frac{K(K-1)}{2} \sum_{i=1}^{K-1} \sum_{j=i+1}^K J(T_i, T_j) \quad (14)$$

We calculated the topic dissimilarity for 5 settings of  $K = 80, 90, 100, 110, 120$  and found that  $K = 100$  with the average topic dissimilarity 0.6317 is better than the others, so in our experiments,  $K$  was set to be 100.

Three topics, which are represented by their top 5 words and their correlations, are shown in Fig. 4. Topic #2 is mainly related to encryption and topic #69 is about secure communication; topic #65 is focused on intrusion prevention. Topics #65 and #69 are highly correlated (with a correlation strength of 0.316), which suggests that the research in these scientific fields is strongly interdependent.

### Apply the CTPM to find the top papers in the respective scientific fields

We applied the straightforward PageRank (SP), TPM, and CTPM to the dataset, respectively. Table 3 shows the top 10 papers ranked by SP. These top papers discuss and address issues on different scientific fields. For instance, the paper “A training algorithm for



**Fig. 4** Topics #2, #65, and #69 and their correlations. Each topic is represented by its top 5 words. The link line between 2 topics represents the correlation between them and the correlation strength is illustrated by the number in the middle of the line

optimal margin classifiers” focuses on machine learning algorithms, yet “Communicating sequential processes” discusses communication. The rank list generated by the SP consisted of the papers with the highest impacts, but this may not help researchers find significant papers in their field of interest. For example, although the paper “Chord: A scalable peer-to-peer lookup service for Internet applications” is a seminal work, it may not be of interest to researchers working on data mining technology. Furthermore, the high ranked papers mostly have older publication dates. These results confirmed the old-bias drawback of the SP.

Compared with the SP, the TPM generates topic-dependent scores, which we denote with  $TPR(d|k)$ , where  $k$  is the topic and  $d$  corresponds to the paper. Table 4 shows the results of the top 10 papers of topic #2 yielded by the TPM. As seen from the rank list, it appears that most papers are about cryptography and information security except the 5th (Application of branching cells to QoS aware service orchestrations) and 6th (Certifying Machine Code Safe from Hardware Aliasing: RISC is Not Necessarily Risky) papers. It is notable that there are 9 papers published in 2012–2014 among the top 10 papers, which suggests the TPM favors the latest papers and age-tapers the older papers too much. This result is likely due to the simple age-taper strategy employed by the TPM, which originally was expected to overcome the drawback of the older paper bias of the SP. In addition to

**Table 3** The top 10 papers generated by the SP

Rank	Title	Publication year	Venue
1	A training algorithm for optimal margin classifiers	1992	COLT
2	Mining association rules between sets of items in large databases	1993	SIGMOD
3	Graph-based algorithms for boolean function manipulation	1986	IEEE Transactions on Computer
4	A tutorial on the support vector machines for pattern recognition	1998	Data Mining and Knowledge Discovery
5	Scheduling algorithms for multiprogramming in hard real-time	1973	Journal of the ACM
6	Database mining: a performance perspective	1993	IEEE Transactions on Knowledge and Data Engineering
7	Communicating distinctive image features from scale-invariant keypoints	1978	Communications of the ACM
8	Distinctive image features from scale-invariant keypoints	2004	International Journal of Computer Vision
9	Chord: a scalable peer-to-peer lookup service for internet applications	2001	SIGCOMM
10	A scalable content-addressable network	2001	SIGCOMM

favoring newer papers, we also found the TPM ranked some papers inaccurately. For instance, we found the paper “A new pseudo-random generator from collision-resistant hash functions” published in 2012 was cited only 2 times. However, in topic #2, it was ranked as the second-highest work. It seems that the paper was overstated in the TPM. In contrast, the paper “Dynamic credentials and ciphertext delegation for attribute-based encryption” published in 2012, also with a citation count of 2 in topic #2, obtained a rank of 26. We then resorted to the Google Scholar search tool and found these citation counts have changed to 6 and 147, respectively. These results suggest that the second paper should have a higher rank than the first paper. The reason for this inappropriate ranking by the TPM is that, for more recent papers, the TPM would mainly rank them in terms of their text content, regardless of the difference of their venues, since they have little chance to be cited. For instance, the TPM would generate similar scores for the newer paper published in an ordinary venue with the new paper published in *NIPS* if they have similar content, even if *NIPS* is a renowned venue.

The CTPM also calculated a topic-dependent score for each paper as well as the TPM. The top 10 papers of topic #2 yielded by the CTPM are shown in Table 5. Compared with the results of the TPM, the top 10 papers generated by the CTPM were all highly correlated with the areas of cryptography and information security. Furthermore, the rank list of the CTPM seems more reasonable than that of the TPM. To evaluate the rank results of the CTPM and TPM, we presented the results of Tables 4 and 5 to 7 senior researchers who all have PhD degrees in the Cryptography and Information Security field and asked them which result was better. Their feedback was not surprising: they all indicated that the result of Table 5 was more reasonable.

**Table 4** The top 10 papers on topic #2 generated by the TPM

Rank	Title	Publish year	Venue
1	Practical UC security with a global random oracle	2014	ACM Conference on Computer and Communications Security
2	A new pseudo-random generator from collision-resistant hash functions	2012	Cryptographers' Track at the RSA Conference
3	Batch verification suitable for efficiently verifying a limited number of signatures	2012	International Conference on Information Security and Cryptology
4	Compact round-optimal partially-blind signatures	2012	International Conference on Security and Cryptography for Networks
5	Application of branching cells to QoS aware service orchestrations	2014	Theoretical Computer Science
6	A fully homomorphic crypto-processor design: correctness of a secret computer	2013	International Symposium on Engineering Secure Software and Systems
7	Certifying machine code safe from hardware aliasing: RISC is not necessarily risky	2013	International Conference on Software Engineering and Formal Methods
8	Deterministic polynomial time equivalence between factoring and key-recovery attack on Takagi's RSA	2008	International Conference on Practice and Theory in Public-key Cryptography
9	Fast two-party secure computation with minimal assumptions	2014	ACM Conference on Computer and Communications Security
10	Private collection of traffic statistics for anonymous communication networks	2014	ACM Conference on Computer and Communications Security

We also used some other acknowledged metrics to evaluate the results of Tables 5 and 6 objectively. They were evaluated based on the following ideas. First, we retrieved the cited numbers of the next 1 year to the next 5 years of each paper listed in Tables 5 and 6 using Google Scholar, which are denoted as C1, C2, C3, C4, and C5. We then compared the results of Tables 4 with 5 according to C1 to C5. We also investigated the academic impact of the authors of these papers. The academic impact of the authors is also an important indicator. The papers published by prominent scholars often attract more attention, thus they may have a high academic impact. Third, we evaluated the results in terms of the academic impact of the venues these papers were published in, because papers published in outstanding venues usually have a high academic impact.

The evaluation metrics of C1 to C5 are based on the paper citations. As mentioned in the section introduction. This approach is a well-known and widely used method for evaluating academic impact. However, it suffers from the limitations of having no relation to the research topic and the preference for older papers. Tables 4 and 5 show that the first limitation of the citation evaluation method does not exist, since the papers listed in the 2 tables are almost about cryptography and information security research. The second limit of the citation evaluation method is mostly due to the fact that older papers have a better chance to be cited; this limitation can be avoided because the C1 to C5 metrics we used are the citation numbers of the next 1–5 years after the paper's publication year, rather than the

**Table 5** The top 10 papers on topic #2 generated by the CTPM

Rank	Title	Publish year	Venue
1	Identity-based encryption from Weil pairing	2001	CRYPTO
2	How to break MD5 and other hash functions	2005	EUROCRYPT
3	Finding collisions in the full SHA-1	2005	CRYPTO
4	Hierarchical identity based encryption with constant size ciphertext	2005	EUROCRYPT
5	On ideal lattices and learning with errors over rings	2010	EUROCRYPT
6	Trapdoors for Lattices: simpler, tighter, faster, smaller	2012	EUROCRYPT
7	Efficient identity-based encryption without random oracles	2005	EUROCRYPT
8	Trapdoors for hard lattices and new cryptographic constructions	2008	ACM Symposium on the Theory of Computing
9	Lossy trapdoor functions and their applications	2008	ACM Symposium on the Theory of Computing
10	Simultaneous hardcore bits and cryptography against memory attacks	2009	Theory of Cryptography

**Table 6** The C1 to C5 values of the papers in Tables 4 and 5

Rank	TPM					CTPM				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
1	5	14	21	–	–	64	186	380	741	1150
2	3	4	5	5	7	197	326	455	611	785
3	0	1	2	3	3	213	386	567	762	943
4	1	2	4	8	10	74	138	210	282	372
5	3	3	3	3	3	40	95	163	269	388
6	19	27	34	38	–	54	119	188	288	360
7	5	8	9	11	–	119	203	326	483	658
8	4	6	7	8	9	60	116	194	294	418
9	25	36	45	–	–	47	92	126	177	243
10	6	10	19	–	–	70	108	153	210	277

total citation numbers. We list the values of C1 to C5 of the papers from Tables 4 and 5 in Table 6. The values of C1 to C5 of the top 10 papers generated by the CTPM are much greater than those of the TPM. If we go through the values of C1 to C5 according to the ranked results of the 2 models, we find that the C3 to C5 values in the CTPM are consistent with their rank results in general, which suggests the rank results of the CTPM are reasonable. While the C1 to C5 values in the TPM seem to have no match to their rank results, they are not in a reasonable order. In general, the CTPM is much better than the TPM at choosing the top influential papers.

The second metric we used is the *h*-index of the authors publishing the papers listed in Tables 4 and 5. The *h*-index measures both the productivity and impact of a scholar, which



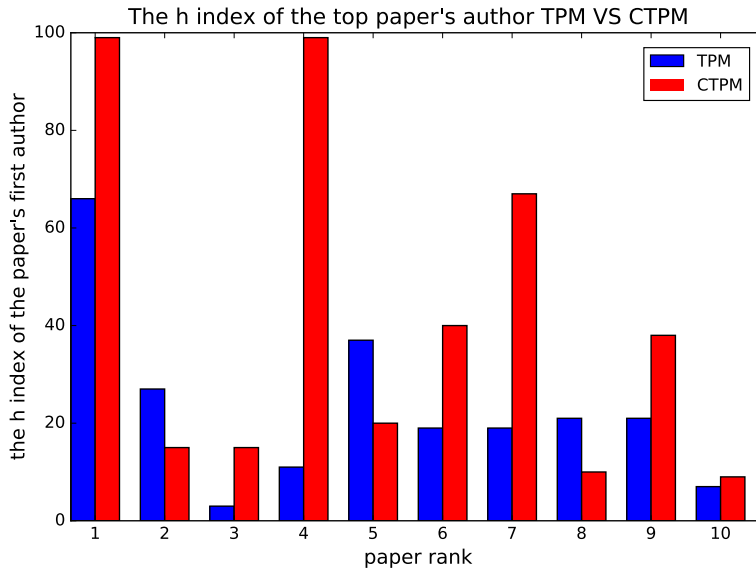
is defined as an author with the index of  $h$  has published  $h$  papers, each of which has been cited by other papers at least  $h$  times. A scholar with a high  $h$  index usually has a high influence in his/her research area, and the papers he/she published often attract extensive attention, thus gaining a high academic impact. Because many papers have multiple authors, we only used the  $h$ -index of the paper's first author to evaluate its impact, since the first author is usually the main contributor to the paper's research. We retrieved the  $h$ -index of the involved authors by Google Scholar and illustrate them in Fig. 5. Overall, the CTPM obtained superior evaluation results to those of the TPM. Specifically, the CTPM had 7 papers with a significantly higher  $h$ -index of first authors. By contrast, the TPM outperformed the CTPM only on the papers with ranks of 2, 5, and 8. Furthermore, the average  $h$ -index of the CTPM (41.2) is close to 2 times of that of the TPM (23.1), which indicates that the CTPM may be superior to the TPM.

The last metric we used was the H5-Index of the venues. The H5-Index of a venue is defined as a venue with an index  $h$  of H5 that published  $h$  papers in the last 5 years, each of which has been cited in other papers at least  $h$  times. We used the H5-index rather than the well-known Impact Factor (IF) because the IFs of journals are available while the IFs of conferences are not. In contrast, the H5-Index of both journals and conferences can be retrieved by some academic search tools, such as AMiner. Making use of the list of the top 1000 venues in Computer Science, which is generated by AMiner according to the H5-Index of the venues, we conducted comparisons on the results obtained by the CTPM and TPM.<sup>5</sup> Table 7 shows that the top 10 papers generated by the CTPM were published in 4 venues and they are all in the list of the top 1000 venues, while in the TPM, nearly a half of the venues (4/10) is beyond the list. Among the top 10 papers, the TPM found 3 papers published in the CCS with a high H5-Index, which makes it look attractive. We made a thorough investigation to explain this fact. The reasons for it are summarized as follows. First, the H5-Index of a venue is a metric that measures its impact in the last 5 years, not the impact over its whole history. The high H5-Index of the CCS implies its outstanding impact is recent. However, in the whole history of the Cryptography and Information Security field, both CRYPTO and EUROCRYPT are as good as or better than the CCS. Second, we found that although topic #2 is about the cryptography and information security topic on the whole, it is more closely related to the cryptography sub-field and slightly farther from the information security sub-field. The top papers generated by the CTPM agree with this fact, and their venues pertain to the cryptography sub-field. However, in the TPM, the top papers and their venues do not conform to the bias feature of topic #2; they are more focused on the information security sub-field. Among them, the CCS is an representative example which may focus more on the information security sub-field than the cryptography sub-field. Overall, the CTPM produces more convincing results than the TPM for the H5-Index metric of the venues of the top papers.

As a side effect of the calculation of the paper topic-dependent impact scores, the CTPM also yields the venue's topic-dependent impact scores. The top 10 venues in topic #2 are shown in Table 8.

We evaluated the results of Table 8 in 2 ways. First, we presented the top venues of Table 8 to the aforementioned researchers and they all approved the ranking results in general. Second, we compared the results with other 3 results. The first was the rank list of computer security and cryptography venues generated by Google Scholar according to the H5-Index of venues, which consisted of the 20 top venues in the area of cryptography and information security. The second was the list of the top journals and conferences

<sup>5</sup> <https://aminer.org/ranks/conf>.



**Fig. 5** *h* Index of the first authors in the TPM and CTPM

**Table 7** The H5-index of venues publishing the top papers of the TPM and CTPM

Rank	TPM		CTPM	
	Venue (short name)	H5-index	Venue (short name)	H5-index
1	CCS	44	CRYPTO	24
2	CT-RSA	19	EUROCRYPT	30
3	ISISC	–	CRYPTO	24
4	ISISCN	–	EUROCRYPT	30
5	TCS	33	EUROCRYPT	30
6	ESSoS	13	EUROCRYPT	30
7	SEFM	–	EUROCRYPT	30
8	IEICE T FUND ELECTR	–	STOC	29
9	CCS	44	STOC	29
10	CCS	44	TCS	33

recommended by the China Computer Federation (CCF).<sup>6</sup> The venues in the CCF list were grouped by the research fields and generated by senior experts in the field through recommendations, reviews, and voting. The third was the list maintained by Guofei Gu,<sup>7</sup> a senior and well-known researcher in the Cryptography and Information Security field. Guofei generated the list of top venues according to the following criteria: acceptance ratio, paper quality and impact, committee member quality, attendee/paper number ratio,

<sup>6</sup> <http://history.ccf.org.cn/sites/ccf/biaodan.jsp?contentId=2903940690850>.

<sup>7</sup> [http://faculty.cs.tamu.edu/guofei/sec\\_conf\\_stat.htm](http://faculty.cs.tamu.edu/guofei/sec_conf_stat.htm).

**Table 8** The top 10 venues on topic #2 of the CTPM

Rank	Venue
1	EUROCRYPT
2	CRYPTO
3	ACM Symposium on Theory of Computing
4	ACM Conference on Computer and Communications Security
5	Theory of Cryptography
6	PKC International Conference on Practice and Theory in Public Key Cryptography
7	RFIDSec
8	Communications of the ACM
9	Journal of Cryptology
10	CT-RSA Conference on Topics in Cryptology

location, history, and industry connections. In Table 9, we show whether the top venues generated by the CTPM are included by other 3 lists.

The results of Table 9 show our CTPM model is a reasonable model. It automatically identified the 6 top venues in the Google list, 7 top venues in the CCF list, and 6 top venues in the Guofei list without any supervised knowledge. Our CTPM also identified 2 other venues, the “ACM Symposium on Theory of Computing” and “Communications of the ACM”, but they are not included in the other 3 lists. Both of these venues are eminent in the area of computer science and do not only focus on cryptography or information security. Moreover, they publish many influential papers about the theory of cryptography and information security. The other 3 lists may leave them out because they are not purely about cryptography and information security, while our CTPM model captured them. These results indicate our model may be more comprehensive and scalable than earlier models.

### Recommend papers to read

We applied the CTPM to the AMiner ACM-Citation-network dataset described in “[The experimental dataset](#)” section to build a paper-recommendation system. The system accepts a query text as input, then outputs 100 papers as the recommendation results. Denoting the input query text as  $q$ , the system first computes its topic distribution  $P(k|q)$  by the CTM fitted to the dataset, then for each paper  $d$  in the dataset, a recommendation score  $RS(d|q)$  is calculated by

$$RS(d|q) = \sum_{k=1}^K P(d|k)P(k|q) \quad (15)$$

where we replaced  $P(d|k)$  with  $TPR(d|k)$ . The top 100 papers with the highest  $RS(d|q)$  are recommended for the reference list of the paper  $d$ . To evaluate the recommendation effectiveness, we compared our model with 8 other models. We chose 1000 papers at random from the paper set as test data. For each paper in the test data, the title and abstract were combined as the input query text, then the top 100 ranked papers were recommended as the citations for it. The ranked papers did not include the papers published after the test

**Table 9** The comparison results of the top 10 venues

Rank	Venue	In Google	In CCF	In Guefei
1	EUROCRYPT	Y	Y	Y
2	CRYPTO	Y	Y	Y
3	ACM Symposium on Theory of Computing	N	N	N
4	ACM Conference on Computer and Communications Security	Y	Y	Y
5	Theory of Cryptography	Y	Y	Y
6	PKC International Conference on Practice and Theory in Public Key Cryptography	Y	Y	Y
7	RFIDSec	N	N	N
8	Communications of the ACM	N	N	N
9	Journal of Cryptology	Y	Y	N
10	CT-RSA Conference on Topics in Cryptology	N	Y	Y

paper, since they have no chance to be cited by the test paper. The recommendation list was compared against the actual citations of the test paper. Like the work performed by Jardine and Teufel (2014) we used MAP as the evaluation metric. Table 10 shows the MAPs of the 9 models.

Our model (MAP = 0.2173) comfortably outperformed the other models. Compared with the other best model (F with MAP 0.2016) (Bethard and Dan 2010), the CTPM has improved the MAP by 7.79%.

We used the popular term-frequency-inverse-document frequency (TFIDF) with cosine similarity (A) as the baseline. For a given query  $q$ , model A calculated the recommendation score for each paper  $d$  by the TFIDF cosine similarity between  $d$  and  $q$ . Then, we employed the LDA with the KL divergence (B) as another metric. For the query  $q$  with a topic distribution  $p_q$ , the KL divergence  $KL(p_q, p_d)$  was calculated for each paper  $d$ . Then, the top 100 lowest KL divergence papers were recommended. As the performance of B (MAP = 0.0609) shows, the use of the LDA led to a remarkable improvement over the TFIDF, which suggests the topic distribution was more effective than that of the TFIDF for paper recommendations. Models C and D used combined strategies to make a trade-off between authorities and the textual content of papers. They first calculated the rank of  $r_1(d|q)$  in terms of the textual similarity between the query  $q$  and the paper  $d$  (Model C

**Table 10** MAP of 9 models

Model	MAP
A: TFIDF cosine	0.0432
B: LDA KL divergence	0.0609
C: TFIDF cosine + citation count	0.0957
D: LDA KL + standard PageRank	0.1361
E: TPR	0.1881
F: B&J; best model	0.2016
G: TPR with Gaussian age-taper	0.1959
H: TPR with CTM	0.1924
I: CTPM	0.2173

used the TFIDF cosine while Model D used the LDA KL). Then, another rank of  $r_2(d)$  was calculated by the citation count (Model C) or standard PageRank (Model D) to take into account the authority of  $d$ , which is independent of  $q$ . The final rank of  $d$  is  $r(d|q) = (r_1(d|q) + r_2(d))/2$ . Models C and D were significant improvements over Models A and B, which indicates that combining the textual similarity with authority is an attractive solution. The TPR used the topic model in conjunction with a mended PageRank to generate the papers topic-dependent scores, and it was a significant improvement over Models C and D. We modified the TPR by replacing the original linear age-taper with a Gaussian age-taper, gaining a 4.16% improvement of the MAP, which suggested that the Gaussian age-taper may be more suitable than the linear age-taper. We also replaced the standard LDA employed by the TPR with the CTM to verify whether the CTM was better than the LDA (Model H). A 2.29% improvement confirmed that the CTM was better than the LDA. Our results of Models E and F were different from that of Jardine and Teufel (2014). In their experiment, Model E obtained a higher MAP than Model F on the ACL Anthology Network (AAN) dataset, while our experimental results were opposite on the AMiner ACM-Citation-network dataset. The difference may be due to the fact that the AAN dataset contains papers published in only the ACL venue, whereas the AMiner ACM-Citation-network dataset consists of papers published in different venues. The difference also suggests that Model E may suffer from a significant performance degradation when it is applied to the papers published in multiple venues. In contrast, Model F uses complex machine learning technologies and may be robust in different datasets.

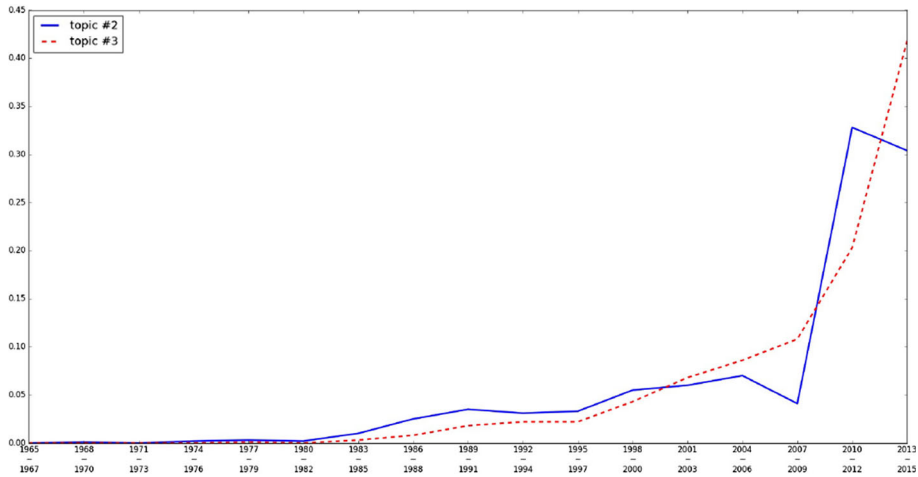
## Explore research topic evolution

Our CTPM can be also applied to outline the evolution of a scientific topic, which may be helpful for researchers to understand questions like “When did the topic arise?”, “When did it become popular?”, and “Is it still a popular topic?”. To answer these questions, we chose the top 1000 papers for each topic to explore the evolution of each topic. We calculated the distributions of the topics over years, which is represented by Eq. (16):

$$P(y|k) = \frac{n_{k,y}}{1000} \quad (16)$$

where  $n_{k,y}$  is the number of papers published in the year  $y$  among the top 1000 papers of topic  $k$ .

We plotted the distribution for topics #2 and #3 in Fig. 6, in which topic #3 corresponds to the area of information retrieval. Topic #3 arose in 1978 and became a research area of increasing interest. From its emergence to 1995, its popularity steadily increased. After 1995, the research works on it began to develop more rapidly. After 2009, there was an explosive increase in this area of study. In contrast to topic #3, the evolution of topic #2 has fluctuated. It arose in 1970, but experienced slow growth for about 10 years. From 1982 to 1991, the first obvious increase in interest emerged. Then, its interest level remained relatively stable. From 2006 to 2009, it dropped slightly, but grew explosively in the next several years. We explored the papers of this period and found the explosive growth could be due to the burst of research interest about security and cryptography in cloud computing, which advanced rapidly in the same period.



**Fig. 6** The evolutions of the topics #2 and #3

### The topic-dependent impact factor calculation of the venues

The IF is an important metric for evaluating the importance of a venue. The IF of a venue  $v$  in the year  $y$  represents the average citation number of papers published in  $v$  and in the 2 years before  $y$ . The IF can be calculated using Eq. (17):

$$\text{IF}(v, y) = \frac{C_{v,y-1} + C_{v,y-2}}{N_{v,y-1} + N_{v,y-2}} \quad (17)$$

where  $C_{v,y}$  denotes the total citation number of papers published in venue  $v$  and in year  $y$ , and  $N_{v,y}$  denotes the total number of papers published in venue  $v$  and in year  $y$ . The IF is an effective metric for evaluating the importance of a venue, but it is subject to some constraints. The most severe one may be that there is no relationship between it and the scientific research fields. Thus, it is a topic-independent metric, while a topic-dependent venue IF may be favored by researchers. Another notable constraint of the IF is that it may be heavily influenced by the popularity of scientific topics. The papers about popular research topics often attract more attention, so they may have a higher citation number, which leads to a higher IF of their venues in general.

Our CTPM can be also applied to calculate the topic-dependent impact factor (TIF) of the venues, which can avoid the aforementioned shortcomings of the IF. Equations (18) and (19) show how to calculate the TIF:

$$\text{TIF}_{v,y,k}^* = \frac{\sum_{d \in D_{v,y-1} \cup D_{v,y-2}} \text{TPR}(d|k)}{N_{v,y-1} + N_{v,y-2}} \quad (18)$$

$$\text{TIF}_{v,y,k} = \frac{\text{TIF}_{v,y,k}^*}{\text{TIF}_{b,y,k}^*} \quad (19)$$

where  $D_{v,y}$  denotes the papers published in the venue  $v$  and in the year  $y$  and  $\text{TPR}(d|k)$  is the score of the paper  $d$  in the topic  $k$ . We first calculated the average TPR of the previous 2 years for each venue using Eq. (18), then ranked  $-\text{TIF}_{v,y,k}^*$  for all venues and chose the

venue with the 10th rank as the base TIF, which is denoted by  $TIF_{b,y,k}^*$  in Eq. (19). The final TIF of a venue can be determined by the ratio of its  $TIF_{v,y,k}^*$  to the base TIF. In our solution of TIF, the 10th rank venue on a topic always has a TIF value 1, regardless of the popularity of the topic. The TIF values of the other venues are also relative to the base TIF. Therefore, the final result may be little affected by the interest in a scientific topic. We show the TIF value of the top 10 venues for topic #2 in 2008 in Table 11.

It seemed that the 10 venues listed in Table 11 can be roughly divided into 4 grade groups by their TIFs: EUROCRYPT and CRYPTO have the closest and highest TIFs, and the next group includes TCC and PKC. CT-RSA and ASIACRYPT also have high TIFs, but these values are a little lower than those of the TCC and PKC group. The remaining 4 venues belong to the last grade group, and their TIFs are close to the base venue, i.e. ACNS' International Conference on Applied Cryptography and Network Security.

We compared the results of Table 11 with the aforementioned Google list in “[Apply the CTPM to find the top papers in the respective scientific fields](#)” section, and the comparison results are listed in Table 12. It can be seen that 8 of the 10 venues in Table 11 are still on the Google list, which indicates the TIF generated by the CTPM is a practical metric to evaluate the topic-dependent dynamic academic impacts of the venues. The TIF rank was different from that in the Google list. We think the reasons for this may be as follows. First, the dataset we used had far fewer records than those available to Google, which may be a major cause of the different ranks. Second, the TIF rank was the result of the venues in 2008, while the Google list reflects the current impact of venues in cryptography and information security. In general, the TIF provides an alternative to evaluate the dynamic impact of the venues, as well as both the IF and H5-Index, while having the advantage of being topic-dependent.

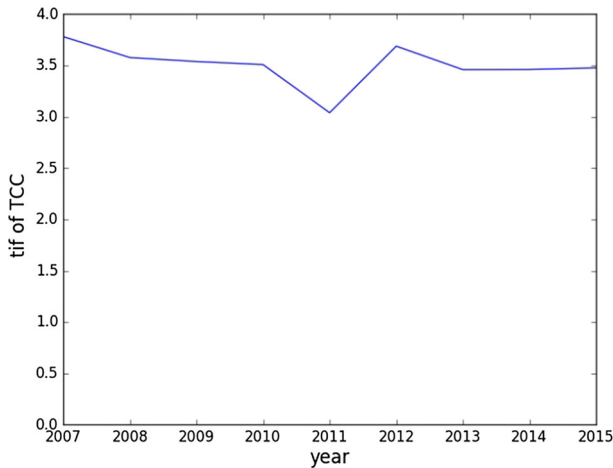
We also illustrated how the TIF of a venue evolved with the years in Fig. 7 in the TCC venue. From 2007 to 2015, the TIF of the TCC on topic #2 varied from 3.04 to 3.78. The maximum TIF occurred in 2007, then it dropped slowly until 2011, in which the TIF was 3.03. In the next few years, it stabilized at the interval around 3.5. The TIFs of the TCC from 2007 to 2015 had a variance of 0.0375 and a mean 3.503, suggesting that this is a prestigious venue with a stable and high academic impact in cryptography, maybe only slightly below EUROCRYPT and CRYPTO. The result also agrees with Table 8 in “[Apply](#)

**Table 11** The TIF values of the top 10 venues for topic #2 in 2008

Venue	TIF
EUROCRYPT	4.802473
CRYPTO	4.351153
TCC	3.577345
PKC	3.255960
CT-RSA	2.702557
ASIACRYPT	2.171382
CHES	1.359957
FSE-International Workshop on Fast Software Encryption	1.132324
ACM Conference on Computer and Communications Security	1.033249
ACNS' International Conference on Applied Cryptography and Network Security	1

**Table 12** The comparison of the TIF rank with the Google list

Venue	TIF	Rank in the Google list	H5
EUROCRYPT	4.802473	7	53
CRYPTO	4.351153	6	53
TCC	3.577345	13	34
PKC	3.255960	16	30
CT-RSA	2.702557	–	–
ASIACRYPT	2.171382	11	36
CHES	1.359957	14	32
FSE- International Workshop on Fast Software Encryption	1.132324	20	28
ACM Conference on Computer and Communications Security	1.033249	1	71
ACNS' International Conference on Applied Cryptography and Network Security	1	–	–


**Fig. 7** The TIF variation of the TCC for topic #2

the CTPM to find the top papers in the respective scientific fields” section, which suggests our calculation of TIF is reasonable and provides an alternative solution to the IF.

## Conclusion

In this paper, we propose a Collective Topic PageRank Model (CTPM) to evaluate the topic-dependent impact of both scientific papers and venues by combining the topic model with the PageRank algorithm. When the CTPM was applied to the scientific literature, we found that it extracted a variety of distinguishable scientific topics, captured the correlations between the scientific topics, and identified the outstanding papers as well as venues with a high impact in each topic. We also have shown that the CTPM can perform a wide range of tasks, such as recommending papers to read or cite, exploring the evolution of



scientific topics, and yielding the topic-dependent impact factors of venues, as demonstrated in “[Experiments](#)” section.

In comparison to previous work, our results show the CTPM has a competitive performance and outperforms not only the state-of-the-art model but also the original PageRank and some modified versions of PageRank (topical and non-topical). We also revealed how the Gaussian age-taper strategy, the CTM, and their combinations with venues improve the performance of PageRank, which may be informative for improving the application of PageRank for scientific literature. Our implementation exhibits some unique advantages. It has no need for predefined scientific topics, as the scientific topics are extracted automatically by the CTM. It relies only on the CTM and PageRank calculation. The trainings of both the CTM and PageRank are not expensive and need to occur only once, which means this tool will be efficient at searching or making predictions. When it is used for recommendations, it achieves competitive performance by only using textual content, publication years, venues of papers, and citation links; there is no need for extensional data, such as the collaborations between authors or the information of search histories, which means it will not suffer from the troublesome cold-start faced by the common recommendation model.

There are some areas of this work that can be improved. One of them is developing a method to extract scientific terms and use them instead of words to represent scientific topics, which may make the approximation of the LDA topic to scientific topic more accurate. Another improvement would be to incorporate the authority of the papers’ authors into the random walk to simultaneously evaluate the academic impact of the papers, venues, and authors at the topic level.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (Grant Nos. 61602202 and 61603146), the Natural Science Foundation of Jiangsu Province, China (Grant Nos. BK20160427 and BK20160428), Top-notch Academic Programs Project of Jiangsu Higher Education Institutions, the Social Key Research and Development Project of Huaian, Jiangsu, China (Grant No. HAS2015020).

## Appendix 1: Decrease the preference for older papers

The original PageRank and most of its modifications set the parameter  $\alpha$  to be 0.1 or 0.15 empirically to weight the contribution of the bias probability, which makes them suffer from bias where older papers are favored. The TPM takes a linear age-taper strategy to address this limit, while our CTPM uses another way to address the bias problem. Taking into account that parameter  $\alpha$  determines the contribution proportions of random choice and transition choosing, the CTPM adjusts  $\alpha$  dynamically according to the papers age. Since new papers have fewer opportunities to be cited than older papers, the CTPM gives newer papers a higher value of  $\alpha$  to indicate that the newer papers should have a high probability to be chosen using random choosing. In addition, we made the following assumptions about the influence of the age of the papers. (1) Researchers prefer up-to-date papers which were published in the last 3 years, so the age-taper of the papers in the last 3 years should be slow. (2) For papers whose ages vary from 4 to 10 years, the age-taper is approximatively linear. (3) If papers are older than 10 years, the age-taper will be slow again. For example, a 20-year old paper has nearly the same timeliness as a 15-year old paper. We applied a Gaussian decay function to make the age-taper agree with the above assumptions. The dynamic is set to

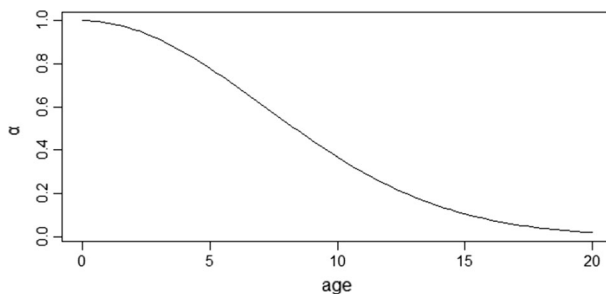
$$\alpha_d = e^{-\frac{g_d^2}{h^2}} \quad (20)$$

where  $g_d$  is the age of the paper  $d$ ;  $h$  is the bandwidth parameter to control the age-taper rate. We experimented  $h$  with different values and found an appropriate setting of 10. The age-taper curve for  $h = 10$  is illustrated in Fig. 8.

According to Eq. (20), a new paper would be chosen mainly by random choice. In particular, the latest paper with an age of 0 years would be chosen completely at random, since it has no chance to be cited.

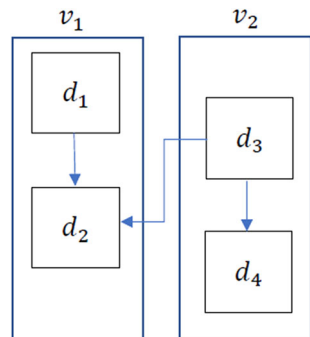
## Appendix 2: A simple example to illustrate our algorithm

To explain our algorithm more clearly, here, we include an example for illustration. Figure 9 shows a citing network with 4 papers,  $d_1$ ,  $d_2$ ,  $d_3$ , and  $d_4$ . Paper  $d_2$  is cited by both  $d_1$  and  $d_3$ , and  $d_3$  cites both  $d_4$  and  $d_2$ . Here, we describe the calculation of  $\text{TRP}^1(d_2|k)$  in detail. The first step is to calculate the Gaussian decay factor of  $d_2$ ; we obtain  $\alpha_2$  easily from Eq. 20. Then we need to calculate the  $\text{TRP}^0(d_2|k)$ . It is also very simple to use the equation  $\frac{r_{d_2,k}}{r_{d_1,k}+r_{d_2,k}+r_{d_3,k}+r_{d_4,k}}$ . After we initialize TPR for all the papers, we calculate the average value of topic-dependent scores of papers for  $v_1$  and  $v_2$ , respectively, which are denoted as  $\text{avg}_{v_1} = \frac{\text{TRP}^0(d_1|k) + \text{TRP}^0(d_2|k)}{2}$  and  $\text{avg}_{v_2} = \frac{\text{TRP}^0(d_3|k) + \text{TRP}^0(d_4|k)}{2}$ . Then the initial topic-dependent scores of  $v_1$  and  $v_2$  are  $V^0(v_1|k) = \frac{2 \text{avg}_{v_1}}{\text{avg}_{v_1} + \text{avg}_{v_2}}$  and



**Fig. 8** The Gaussian age-taper of the CTPM

**Fig. 9** A citing network with 4 papers.  $d_1$  and  $d_2$  are published in the venue  $v_1$ ,  $d_3$  and  $d_4$  are published in the venue  $v_2$



$V^0(v_2|k) = \frac{\text{avg}_{v_2}}{\text{avg}_{v_1} + \text{avg}_{v_2}}$ . So far, we have finished the initialization work and will enter the first iteration. For  $d_2$ , the bias probability is calculated by the following equation:

$$B^1(d_2|k) = \frac{\sqrt{V^0(v_1|k)r_{d_2,k}}}{\sqrt{V^0(v_1|k)r_{d_1,k}} + \sqrt{V^0(v_1|k)r_{d_2,k}} + \sqrt{V^0(v_2|k)r_{d_3,k}} + \sqrt{V^0(v_2|k)r_{d_4,k}}} \quad (21)$$

In our algorithm, the most complicated step is the calculation of the transition probabilities. For  $d_2$ , there are 2 papers  $d_1$  and  $d_3$  citing it, thus the transition probabilities  $T(d_2|d_1, k)$  and  $T(d_2|d_3, k)$  will be greater than 0 and others will be 0. The calculation of  $T^1(d_2|d_3, k)$  requires 3 steps to complete. The first step is obtaining  $T'(d_2|d_3, k)$ , which can be calculated using Eq. 8 and considering  $L_{d_3} = \{d_2, d_4\}$ :

$$T'(d_2|d_3, k) = \sqrt{r_{d_3,k} \frac{r_{d_2,k}}{r_{d_2,k} + r_{d_4,k}}} \quad (22)$$

The second step is obtaining  $T''(d_2|d_3, k)$ . Because  $C_{d_2} = \{d_1, d_3\}$ , it can be calculated using Eq. 9:

$$T''(d_2|d_3, k) = \frac{T'(d_2|d_3, k)}{T'(d_2|d_1, k) + T'(d_2|d_3, k)} \quad (23)$$

In the last step, we calculate  $T^1(d_2|d_3, k)$  as follows according to Eq. 10:

$$T^1(d_2|d_3, k) = \frac{\sqrt{V^0(v_1|k)T''(d_2|d_3, k)}}{\sqrt{V^0(v_1|k)T''(d_2|d_3, k)} + \sqrt{V^0(v_2|k)T''(d_4|d_3, k)}} \quad (24)$$

where the calculation of  $T''(d_4|d_3, k)$  is similar to  $T''(d_2|d_3, k)$ . The calculation of another transition probability  $T^1(d_2|d_1, k)$  follows the same process of  $T^1(d_2|d_3, k)$ . Once we have calculated  $\alpha_2$ ,  $B^1(d_2|k)$ ,  $T^1(d_2|d_3, k)$  and  $T^1(d_2|d_1, k)$ , the  $\text{TRP}^1(d_2|k)$  can be obtained easily with Eq. 4, which is defined as follows:

$$\text{TRP}^1(d_2|k) = \alpha_2 B^1(d_2|k) + (1 - \alpha_2)(T^1(d_2|d_1, k)\text{TRP}^0(d_1|k) + T^1(d_2|d_3, k)\text{TRP}^0(d_3|k)) \quad (25)$$

## References

- Bethard, S., & Dan, J. (2010). Who should i cite: Learning literature search models from citation behavior. In *ACM conference on information and knowledge management, CIKM 2010, Toronto: Ontario, Canada, October* (pp. 609–618).
- Blei, D. M., Jordan, M. I., Griffiths, T. L., & Tenenbaum, J. B. (2003a). Hierarchical topic models and the nested Chinese restaurant process. In *International conference on neural information processing systems* (pp. 17–24).
- Blei, D. M., Lafferty, J. D., Blei, D. M., & Lafferty, J. D. (2007). Correction: A correlated topic model of science. *Annals of Applied Statistics*, 1(2), 634–634.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003b). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Ding, Y. (2011). *Topic-based PageRank on author cocitation networks*. New York: Wiley.
- Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5220–5227.

- Fujii, A. (2007). Enhancing patent retrieval by citation analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, ACM (pp. 793–794).
- Garfield, E. (2006). Citation indexes for science: A new dimension in documentation through association of ideas. *International Journal of Epidemiology*, 35(5), 1123–1127.
- Gori, M., & Pucci, A. (2007). Research paper recommender systems: A random-walk based approach. In *IEEE/WIC/ACM international conference on web intelligence* (pp. 778–781).
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1), 5228.
- Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. *Science*, 66(1713), 385–389.
- Gyngyi, Z., Garcia-Molina, H., & Pedersen, J. (2004). Combating web spam with trustrank. In *Thirtieth international conference on very large data bases* (pp. 576–587).
- Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4), 784–796.
- Jardine, J. G., & Teufel, S. (2014). Topical PageRank: A model of scientific expertise for bibliographic search. In *EACL* (pp. 501–510).
- MacLean, M., Davies, C., Lewison, G., & Anderson, J. (1998). Evaluating the research activity and impact of funding agencies. *Research Evaluation*, 7(1), 7–16.
- Meij, E., & De Rijke, M. (2007). Using prior information derived from citations in literature search. In *Large scale semantic access to content (text, image, video, and sound)* (pp. 665–670). Le centre de Hautes etudes Internationales D’Informatique Documentaire.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*, Technical report. Stanford InfoLab.
- Pal, S. K., & Narayan, B. L. (2005). A web surfer model incorporating topic continuity. *IEEE Transactions on Knowledge and Data Engineering*, 17(5), 726–729.
- Richardson, M., & Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *Nips* (pp. 1441–1448).
- Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 306–315).
- Tang, J., Jin, R., & Zhang, J. (2008a). A topic modeling approach and its integration into the random walk framework for academic search. In *Eighth IEEE International Conference on Data Mining* (pp. 1055–1060).
- Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008b). Arnetminer: Extraction and mining of academic social networks. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 990–998).
- Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06), P06010.
- Wang, X., Zhai, C., & Roth, D. (2013). Understanding evolution of research themes: A probabilistic generative model for citations. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining*, ACM (pp. 1115–1123).
- Wu, B., Goel, V., & Davison, B. D. (2006). Topical trustrank: Using topicality to combat web spam. In *International conference on world wide web* (pp. 63–72).
- Yan, E. (2014). Topic-based pagerank: Toward a topic-level scientific evaluation. *Scientometrics*, 100(2), 407–437.
- Yang, Z., Tang, J., Zhang, J., Li, J., & Gao, B. (2009). Topic-level random walk through probabilistic model. In *Proceedings of joint international conferences on advances in data and web management, APWeb/ WAIM 2009, Suzhou, China, April 2–4* (pp. 162–173).